UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

GERAÇÃO DE REGRAS DE IDENTIFICAÇÃO DE PRODUTOS EM DESCRIÇÕES TEXTUAIS DE COMPRAS APRESENTADAS EM PORTAIS DE TRANSPARÊNCIA PÚBLICA

Eduardo Soares de Paiva

Orientadora
Prof. Dra. Kate Cerqueira Revoredo

RIO DE JANEIRO, RJ – BRASIL
Fevereiro de 2017

GERAÇÃO DE REGRAS DE IDENTIFICAÇÃO DE PRODUTOS EM DESCRIÇÕES TEXTUAIS DE COMPRAS APRESENTADAS EM PORTAIS DE TRANSPARÊNCIA PÚBLICA

Eduardo Soares de Paiva

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

_____

Prof. Kate Cerqueira Revoredo, D.Sc. – UNIRIO

_____

Claudia Cappelli Aló, D.Sc. – UNIRIO

_____

Flávia Cristina Bernardini, D.Sc. – UFF

RIO DE JANEIRO, RJ – BRASIL
Fevereiro de 2017

Paiva, Eduardo Soares de.
P142 Geração de regras de identificação de produtos em descrições textuais de compras apresentadas em portais de transparência pública / Eduardo Soares de Paiva, 2017.
111 f.

Orientadora: Kate Revoredo.
Dissertação (Mestrado em Informática) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2017.

1. mineração de texto. 2. transparência pública. 3. Tratamento de dados. 4. Processamento internsivo de dados. I. Revoredo, Kate, oriente. II. Título.

PAIVA, Eduardo Soares de. **GERAÇÃO DE REGRAS DE IDENTIFICAÇÃO DE PRODUTOS EM DESCRIÇÕES TEXTUAIS DE COMPRAS APRESENTADAS EM PORTAIS DE TRANSPARÊNCIA PÚBLICA**. UNIRIO, 2017. 109 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Os portais de transparência pública vêm se constituindo em importantes canais de comunicação entre o governo e a sociedade. No entanto, nem sempre o formato das informações apresentadas nesses portais é o mais apropriado. Por exemplo, as descrições de compras em formato de texto dificultam a análise dessas compras, pois para se saber os produtos que estão sendo adquiridos é necessária uma leitura e interpretação de cada descrição de compra, o que é humanamente impossível, devido ao grande volume de dados apresentados. Dessa forma, o objetivo desse trabalho é fazer a identificação automática dos produtos que são especificados de forma textual nas descrições de compras. Logo, a questão de pesquisa dessa dissertação é: como identificar de forma automatizada os produtos a partir das especificações textuais que são usadas para caracterizá-los nas descrições dos gastos que são apresentados nos portais de transparência pública? Para isso, é proposto um processo de descoberta de conhecimento em dados textuais capaz de gerar regras que possibilitam a identificação de produtos a partir das descrições textuais de compras. A pesquisa foi realizada utilizando a metodologia CRISP-DM (Cross Industry Standard Process for Data Mining) e sua avaliação foi dividida em duas partes: a primeira avalia as regras geradas, enquanto que a segunda verifica a qualidade dos resultados obtidos no processo de identificação de compras propriamente dito. Os estudos concluíram que o processo proposto apresentou resultados satisfatórios, porém ainda existem muitas outras possibilidades de melhorias que podem ser exploradas em trabalhos futuros.

Palavras-chaves: transparência pública, mineração de texto, tratamento de dados, processamento intensivo de dados.

ABSTRACT

The public transparency portals are becoming important communication channels between government and society. However, not always the portals present the information in the most appropriate format. For example, the description of purchases in text format hinders analysis of purchases, as to know the products that are being acquired, it is necessary reading and interpreting of each purchase description, what is humanly impossible due to large data volume presented. Thus, this work goal is automatically identifying the products that are textually specified in the purchase descriptions. So this dissertation research question is: How to automatically identify products by textual specifications, used to characterize them in expenditure, descriptions presented in the public transparency portals? For this, a knowledge discovery process is proposed in textual data capable of generating rules that allow products identification from purchases textual descriptions. This research was performed using the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology and its evaluation was divided into two parts: the first evaluates the rules generated, while the second checks the results quality obtained in identifying process of the purchases. The studies concluded that the proposed process presented satisfactory results, but there are still many other possibilities for improvement that can be explored in future work.

Keywords: *Public transparency, text mining, data treatment, data intensive processing.*

Sumário

Capítulo 1 – Introdução 12
1.1 Motivação 12
1.2 Definição do Problema 14
1.3 Objetivos 15
1.4 Enfoque de Solução e Hipótese 16
1.5 Metodologia 17
1.6 Organização da Dissertação 19
Capítulo 2 – Fundamentação Teórica 20
2.1 Transparência Pública 20
2.1.1 Publicidade 21
2.1.2 Dados Governamentais Abertos 22
2.1.3 Portais de Transparência Pública 24
2.2 Mineração de Texto 27
2.2.1 Pré-processamento de Texto 28
2.2.2 Representação Textual 30
2.2.3 Descoberta de Conhecimento 33
2.3 Processamento Intensivo de Dados 34
2.4 Trabalhos Relacionados 38
Capítulo 3 – Proposta 42
3.1 Entendimento do Negócio 42
3.2 Entendimento dos Dados 44
3.3 Preparação dos Dados e Modelagem 46
3.3.1 Pré-processamento 47
3.3.2 Geração de Frases Candidatas 49
3.3.3 Filtragem de Frases Frequentes 50
3.3.4 Poda de Sub frases 51
3.3.5 Geração de Regras 53
3.3.6 Refinamento de Regras 54
3.3.7 Aplicação das Regras 55
Capítulo 4 – Avaliação 57
4.1 Infraestrutura 57
4.2 Projeto de Avaliação 58
4.2.1 Avaliação das Regras 58
4.2.2 Avaliação da Identificação 60
4.2.3 Estudo de Caso 60
4.3 Base de Dados 61
4.4 Experimentos 62
4.4.1 Análise das Regras Geradas 63
4.4.2 Análise dos Resultados 70
Capítulo 5 – Aplicações 74
5.1 Cálculo de Preços de Referência dos Produtos Comprados Pela Administração Pública 74
5.2 Identificação de Compras com Preços Muito Acima do Esperado 75
5.3 Possibilidade de Cancelar uma Compra Superfaturada Antes de sua Concretização 77

5.4 Comparação Entre Valores Pagos em Compras Licitadas e Não Licitadas..... 78
5.5 Análise das Marcas Mais Compradas 79
5.6 Identificação de Fornecedores Vendendo o Mesmo Produto com Preços Diferentes..... 80
5.7 Acompanhar a Tendência dos Preços de Produtos 82
5.8 Comparar a Relação Entre Preço Pago e Quantidade Comprada 83
5.9 Outras Aplicações 85
Capítulo 6 – Conclusão 87
6.1 Contribuições 87
6.2 Análise da Solução Proposta..... 90
6.3 Limitações..... 92
6.4 Trabalhos Futuros 92
6.5 Considerações Finais 93
Referências 95
Apêndice I – Análise de Outliers..... 100

Lista de Figuras

Figura 1 - Descrição Textual de Compras 15
Figura 2 - Fases do CRISP-DM (adaptado de SHEARER, 2000) 18
Figura 3 - Transparência Ativa e Passiva (CGU, 2013) 25
Figura 4 - Degraus da Transparência (CAPPELLI; LEITE, 2008) 26
Figura 5 - Um framework tradicional para análise de texto, adaptado de (HU e LIU, 2012) 28
Figura 6 - Representações da Bag of Words 32
Figura 7 - Modelo de execução do MapReduce, adaptado de (DEAN; GHEMAWAT, 2001) 36
Figura 8 - Estágios da Execução da Despesa 44
Figura 9 - Tela de Empenho e Modelo de Dados de Empenho 45
Figura 10 - Processo de geração de regras de identificação de produtos 46
Figura 11 - Resultado do Pré-processamento 48
Figura 12 - Algoritmo de geração de Frases Candidatas 49
Figura 13 – Exemplo do processo de Geração de Frases 50
Figura 14 – Exemplo do processo de filtragem de frases 51
Figura 15 - Algoritmos de Poda de Sub Frases 52
Figura 16 - Exemplo do processo de poda de sub frases 52
Figura 17 – Aplicação das regras 55
Figura 18 - Exemplos de configurações de clusters 59
Figura 19 - Qualidade das regras geradas por experimento 69
Figura 20 - Recortes das telas do Portal das Transparência para registros considerados outliers da regra R_3 73
Figura 21 - Tela do Portal da Transparência do Governo Federal com empenho de produtos com preços muito acima do esperado 76
Figura 22 - Tela de pagamento do Portal da Transparência do Governo Federal 78
Figura 23 - Telas do Portal da Transparência com o mesmo fornecedor vendendo o mesmo produto com preços diferentes 81
Figura 24 - Recorte de telas do Portal da Transparência de compras com valores unitários discrepantes 85

Lista de Tabelas

Tabela 1 - Exemplos de operações suportadas pelos RDDs no Spark 37
Tabela 2 - Parâmetros dos experimentos realizados..... 62
Tabela 3 - Tempo de processamento e quantidade de regras geradas por experimento. 62
Tabela 4 - Clusters das regras gerados no experimento 6 65
Tabela 5 - Clusters das regras gerados no experimento 5 66
Tabela 6 - Clusters das regras gerados no experimento 4 67
Tabela 7 - Clusters das regras gerados no experimento 3 67
Tabela 8 - Clusters das regras gerados no experimento 2 68
Tabela 9 - Clusters das regras gerados no experimento 1 68
Tabela 10 - Outliers identificados por regra..... 71
Tabela 11 - Amostra de preços de referência calculados 74
Tabela 12 - Amostra de preços muito acima do esperado..... 75
Tabela 13 - Amostra de preços praticados em compras com e sem licitação 79
Tabela 14 - Amostra de marcas mais compradas por produtos..... 80
Tabela 15 - Amostra dos preços do litro da gasolina durante o ano de 2015..... 82
Tabela 16 - Amostra de produtos com valores unitários muito altos 84
Tabela 17 - Amostra de produtos com valores unitários muito baixos 84

Lista de Notações

$\mathcal{D}$ Conjunto de Descrições de Compras

d_i Descrição de Compra pertencente ao conjunto $\mathcal{D}$

$df(i)$ Número de documentos em um corpus de texto contendo o termo *i*

G Frase composta por uma sequência de palavras *w*

M Número total de caracteres correspondentes presentes em duas strings a serem comparadas

N Número de documentos em um corpus de texto

n Tamanho de uma frase G, formada por uma sequência de palavras *w*

n_i Número de Ocorrências do termo *i*

$\mathcal{P}$ Conjunto de Produtos

p_j Produto pertencente ao conjunto $\mathcal{P}$

q Números de pontos vizinhos a serem considerados em um processo de detecção de outliers

r Números outliers a serem encontrados em um processo de detecção de outliers

$\boldsymbol{tf_i}$	Frequência absoluta do termo *i*
$\boldsymbol{tfidf(i)}$	Frequência inversa de um termo *i*
$\boldsymbol{T}$	Número total de caracteres presentes em duas strings a serem comparadas
$\boldsymbol{w}$	Palavra pertencente ao texto a ser analisado

Capítulo 1 – Introdução

Este capítulo fornece uma visão geral da pesquisa, apresentando sua motivação, o problema abordado, os objetivos a serem alcançados, o enfoque de solução, a definição da hipótese, a metodologia empregada e por fim, a organização do restante da dissertação.

1.1 Motivação

Com o avanço da Internet e das tecnologias digitais, questões de transparência eletrônica, democracia digital, governo aberto, ciberdemocracia e outros termos que associam a atuação governamental às ferramentas apoiadas pelo uso de Tecnologia da Informação vêm ganhando cada vez mais importância para a sociedade. Os ecossistemas digitais são ambientes que estimulam a participação cidadã e consequentemente aumentam o grau de democratização dos governos.

Acompanhando essa tendência, a área de transparência tem encontrado um terreno fértil para promover o estimulo ao controle social dos gastos públicos. Os portais de transparência pública vêm se transformando em importantes canais de comunicação entre o governo e a sociedade. Por meio desses portais, o cidadão tem acesso a uma série de informações, que facilitam o acompanhamento e o controle das atividades governamentais.

Visando atender à crescente demanda por informações públicas, o governo brasileiro tem se empenhado para disponibilizar seus dados. Nesse contexto, duas leis podem ser consideradas como marcos no processo de abertura dos dados governamentais. A Lei Complementar 131(BRASIL, 2009) determina a disponibilização, em tempo real, de informações pormenorizadas sobre a execução orçamentária e financeira da União, dos Estados, e dos Municípios; e a Lei 12.527 (BRASIL, 2011) permite que qualquer pessoa possa solicitar qualquer informação produzida pelo governo, desde que tal

informação não esteja expressamente classificada como não pública (ex: reservada, secreta ou ultrassecreta).

Além dessas novas leis, que permitem implantar uma nova cultura na Administração Pública, outras iniciativas foram adotadas para aumentar a disponibilidade dos dados governamentais, como a criação da INDA - Infraestrutura Nacional de Dados Abertos (BATISTA; SILVA; MIRANDA, 2013), que é a política do governo brasileiro para dados abertos, e que tem como finalidade garantir e facilitar o acesso pelos cidadãos, pela sociedade e em especial pelas diversas instâncias do setor público aos dados e informações produzidas ou custodiadas pelo Poder Executivo Federal.

Adicionalmente, o Brasil também é signatário da OGP (OPEN GOVERNMENT PARTNERSHIP, 2011), uma iniciativa internacional que pretende difundir e divulgar, globalmente, práticas governamentais relacionadas à transparência dos governos, ao acesso a informação pública e à participação social.

Outro fator que tem viabilizado a institucionalização desses instrumentos de transparência foi o alto grau de informatização dos governos. Atualmente, praticamente todas as atividades realizadas pelo poder público são controladas por algum sistema de informação (JARDIM, 2004). A automatização dos processos governamentais gera grandes volumes e variedades de dados, que são utilizados como fontes de informação de transparência pública.

Porém, essas iniciativas por si só não garantem um aumento efetivo do grau de transparência da informação que o cidadão tem das atividades governamentais. Isso acontece porque a maioria dos dados disponibilizados para o cidadão não foram concebidos com esse propósito. Em geral as informações são oriundas de sistemas coorporativos cujo objetivo é propiciar o controle administrativo das contas públicas e por isso nem sempre o seu formato é o mais apropriado para o cidadão entender o que realmente elas representam.

Dentre essas informações não tratadas, estão as descrições de compras feitas pela Administração Pública. Os produtos comprados são descritos em formato textual de livre preenchimento, o que inviabiliza análises de compras de produtos iguais e prejudica o acompanhamento sistemático dos gastos.

Outro agravante nesse contexto é o elevado volume de dados disponibilizados diariamente por esses sites. Apesar da grande quantidade de informações apresentadas permitir uma maior abrangência e mais insumo para que o cidadão possa acompanhar a atuação governamental, a falta de mecanismos de classificação e organização dessas

informações (com relação a importância desses gastos) acaba fazendo com que dados relevantes fiquem escondidos no grande volume de informações disponibilizadas, dificultando o entendimento e reuso dessas informações.

Alguns trabalhos já se propuseram a melhorar a qualidade das informações dos portais de transparência. PAIVA; REVOREDO; BAIÃO (2016) propuseram um integrador para os dados do portal da transparência do Governo Federal, que são oriundos de diversas fontes distintas. Com relação ao tratamento dos dados textuais, (CARVALHO et al., 2013; CARVALHO et al., 2014a; PAIVA; REVOREDO, 2016b) apresentam soluções que visam extrair informações dos dados textuais apresentados em portais de transparência. No entanto, ainda não existe uma solução abrangente, capaz de aliar a interpretação dos dados textuais ao grande volume de dados. Dessa forma, faz-se necessário o desenvolvimento de pesquisas a fim de estudar e propor novas formas de tratamento de informações que possam melhorar a qualidade dos dados que estão sendo disponibilizados nos portais de transparência pública. Isso pode propiciar melhores condições de acompanhamento das atividades governamentais, por parte do cidadão, e consequentemente, ajudar a consolidar a transparência.

1.2 Definição do Problema

Diariamente, os diversos Portais de Transparência pública disponibilizam centenas de milhares de registros. Só o Portal da Transparência do Governo Federal (CGU, 2004) apresenta cerca de 46 mil novos registros a cada dia.

No entanto, essas informações não estão todas estruturadas e fáceis de serem tratadas. Algumas das informações mais relevantes na identificação de uma determinada compra vêm descritas em formato de texto, o que dificulta análises sistematizadas. Essas informações não são estruturadas para permitir que os usuários possam especificar as características do material que ele está adquirindo, de acordo com as suas necessidades. Porém, tal flexibilidade acarreta uma maior dificuldade em se identificar as compras que estão sendo feitas.

Na Figura 1 é apresentada uma descrição de compra presente em um portal de transparência, sendo que, para se identificar o que realmente está sendo especificado nessa descrição, é necessária a leitura e interpretação do texto descritivo. Dessa forma, a partir da descrição textual apresentada na Figura 1, deve-se abstrair que tal compra se refere a blanquete de peru. No entanto, para se extrair esse tipo de informação, das

compras que são apresentadas nos portais, seriam necessárias análises de várias descrições de compras como essa, o que possibilitaria a obtenção de conhecimentos como por exemplo o preço médio praticado por produto, quantidade média comprada e maiores fornecedores. Porém, o grande volume de dados dos portais de transparência torna esse tipo de análise impossível de ser feita de forma manual.

185,00000 KG COMPOSTO ALIMENTAR SALAME, TIPO BLANQUETE DE PERU, INGREDIENTES CARNE DE PERU, APRESENTAÇÃO PEÇA INTEIRA, TEMPERATURA CONSERVAÇÃO 0 A 10 ¨C. SIMILAR À MARCA SADIA, RESENDE, PERDIGÃO. MARCA: Seara ITEM DO PROCESSO: 00006 ITEM DE MATERIAL: 000137081

Figura 1 - Descrição Textual de Compras

Sendo assim, faz-se necessária a criação de um mecanismo capaz de analisar as descrições textuais das compras, que são apresentadas nos portais de transparência, e identificar o produto a que cada uma delas se refere.

Logo, o problema a ser tratado pode ser considerado como uma questão de classificação textual e é definido da seguinte forma:

Dada uma coleção de descrições de compras $\mathcal{D} = \{d_1, d_2, d_3, \dots, d_k\}$, deseja-se definir um conjunto $\mathcal{P}$ de produtos $\mathcal{P} = \{p_1, p_2, p_3, \dots, p_n\}$ e uma função $\mathcal{F}: \mathcal{D}X\mathcal{P} \rightarrow \{0, 1\}$, ou seja, uma função $\mathcal{F}$ que atribui um valor 0 ou 1 para cada par (d_i, p_j), tal que $d_i \in \mathcal{D}\ e\ p_j \in \mathcal{P}$. Se o valor atribuído for 1, diz se que a descrição de compra d_i se refere ao produto p_j. Caso contrário, diz-se que a descrição de compra d_i não se refere ao produto p_j. Cabe ressaltar que é possível que uma mesma descrição esteja relacionada a mais de um produto e que o conjunto $\mathcal{P}$ de produtos pode receber novos elementos a qualquer momento.

Sendo assim, o problema em questão pode ser caracterizado da seguinte maneira: Como identificar de forma automatizada os produtos a partir das especificações textuais que são usadas para caracterizá-los nas descrições dos gastos que são apresentados nos portais de transparência pública?

1.3 Objetivos

Os objetivos dessa pesquisa são descritos da seguinte forma

Objetivo Principal:

- Identificar os produtos mais comprados pela Administração Pública, por meio da análise das descrições textuais de compras, apresentadas nos portais de transparência.

Objetivos intermediários:

- Determinar quais são as descrições de compras que descrevem os produtos mais comprados pela Administração Pública.
- Gerar regras de identificação para esses produtos mais comprados.

Cabe ressaltar que, devido à natureza dos dados dos portais de transparência ou seja, grandes volumes de dados que são atualizados diariamente de forma incremental, a solução a ser apresentada está condicionada à satisfação de dois requisitos: ser escalável[1] e ter condições de processar grandes volumes de dados[2].

1.4 Enfoque de Solução e Hipótese

Conforme apresentado na Seção 1.2, o problema a ser tratado consiste em identificar de forma automatizada os produtos a partir das especificações textuais que são usadas para caracterizá-los nas descrições dos gastos que são apresentados nos portais de transparência pública.

Dessa forma, considerando-se uma frase como sendo uma sequência de tokens contínuos e partindo-se da premissa de que descrições de produtos similares apresentam alguma sequência de tokens iguais, considerou-se a seguinte hipótese:

Se forem identificadas as sequências de tokens que mais se repetem em um determinado conjunto de descrição de compras, então, essas sequências de tokens caracterizarão os produtos mais comprados desse conjunto de descrições.

Optou-se pela caracterização apenas dos produtos mais comprados devido à grande quantidade de diferentes produtos que a Administração Pública pode adquirir (centenas de milhares), o que torna inviável uma abordagem capaz de identificar todos os

[1] Escalável: aquilo que tem condições para crescer de forma uniforme ou para suportar um aumento de carga (DICIONÁRIO DA LÍNGUA PORTUGUESA, 2016).
[2] Nesse contexto, entende-se por grande volume de dados, bases de dados de qualquer tamanho (tão grande quanto o necessário), sendo que o hardware a ser utilizado deve ser definido de acordo com esse volume de dados.

produtos possíveis. Logo, esse trabalho pretende criar regras de identificação para os produtos mais comprados, otimizando assim o esforço de identificação de tais produtos.

A solução proposta faz uso de mineração de texto e está inspirada na abordagem apresentada em (EL-KISHKY et al., 2014; LIU et al., 2015). Esses trabalhos utilizam técnicas de mineração de frases para a definição dos principais tópicos abordados em corpus de texto.

Sendo assim, o trabalho ora proposto utiliza a mineração de frases no contexto dos dados de portais de transparência, a fim de permitir a identificação dos produtos mais relevantes que são apresentados nas descrições textuais de compras.

Visando atender aos requisitos de escalabilidade e de capacidade de processamento de grandes volumes de dados, todo o processo de identificação de produtos foi projetado para utilizar o Apache Spark (ZAHARIA et al. 2010; ZAHARIA et al. 2012), um framework para processamento de big data que roda de forma paralela em cluster de computadores. Dessa forma, aumentos expressivos no volume de dados a ser processado podem ser compensados com a inclusão de novas máquinas ao cluster, sem o comprometimento da performance.

1.5 Metodologia

Para atingir o objetivo desse trabalho é proposto uma metodologia que segue as etapas enunciados pelo CRISP-DM (Cross Industry Standard Process for Data Mining) (SHEARER, 2000). O CRISP-DM tem como objetivo fornecer orientações básicas para a condução do processo de KDD (GOLDSCHMIDT; BEZERRA, 2015). Esse modelo está dividido em 6 fases e tem como finalidade definir os passos a serem seguidos em um projeto de mineração de dados. No entanto, ele também pode ser aplicado à mineração de textos, desde que as fases de entendimento dos dados e de preparação de dados sejam adaptadas para dados textuais (MAIA, 2015). Na Figura 2 são indicadas as fases do modelo, sendo que, as setas internas representam as dependências mais importantes e frequentes entre essas fases e a seta externa indica que a metodologia é iterativa, ou seja, pode ser necessário se passar mais de uma vez por essas fazes até se atingir o objetivo final.

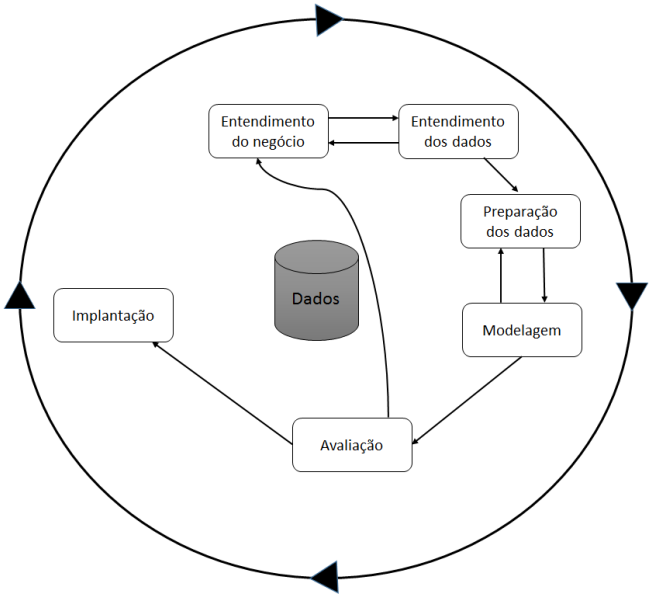


Figura 2 - Fases do CRISP-DM (adaptado de SHEARER, 2000)

O entendimento do negócio é a fase inicial do processo e se concentra em entender os objetivos e requisitos do projeto a partir da perspectiva do negócio. Essa fase é responsável por converter esse conhecimento em uma definição do problema de mineração e em um plano preliminar projetado para atingir os objetivos (AZEVEDO; SANTOS, 2008).

A fase de entendimento dos dados geralmente ocorre ao mesmo tempo da compreensão do contexto (GOLDSCHMIDT; BEZERRA, 2015). Essa etapa é responsável por realizar um estudo minucioso dos dados disponíveis, a fim de propiciar maior familiaridade com estes dados. Ela envolve atividades como: identificação de problemas de qualidade de dados, descrição do formato dos dados, quantidade de registros, atributos e relacionamento entre atributos.

A atividade de preparação dos dados é responsável pelo pré-processamento dos dados envolvidos na mineração. Nessa fase, os dados são selecionados, limpos, integrados, formatados e adequados para as atividades subsequentes. O objetivo principal dessa fase é deixar os dados prontos para serem utilizados na etapa de modelagem. Na fase de modelagem, são definidos e aplicados os algoritmos de mineração de dados. Nessa fase, também é feita a escolha dos parâmetros, utilizados pelos algoritmos de mineração, que melhor se adaptam ao conjunto de dados em questão.

Durante a avaliação, os resultados são analisados, a fim de se verificar a qualidade dos modelos gerados. Normalmente essa análise é baseada em indicadores e métricas comparativas, cujo objetivo é confrontar os resultados obtidos com os esperados.

Finalmente, a implantação consiste no planejamento e acompanhamento das ações a serem realizadas pelo modelo de conhecimento gerado. Nessa fase, os modelos gerados são colocados à disposição do usuário para serem usados.

1.6 Organização da Dissertação

O restante do texto dessa dissertação está organizado da seguinte forma: no Capítulo 2 é apresentada a fundamentação teórica com os principais conceitos relacionados aos assuntos abordados nesse trabalho. No capítulo 3 é apresentada a proposta de solução para o problema de identificação de produtos em descrições textuais de Compras. No Capítulo 4 é feita a avaliação da solução proposta, enquanto que, no Capítulo 5 são apresentadas algumas aplicações para solução desenvolvida. Finalmente no Capítulo 6 é apresentada a conclusão da dissertação.

Capítulo 2 – Fundamentação Teórica

Essa dissertação aborda aspectos de três áreas de conhecimento: transparência pública, mineração de texto e processamento de grandes volumes de dados. Dessa forma, esse capítulo faz uma revisão dos principais conceitos ligados a essas áreas, bem como de alguns trabalhos relacionados à extração de conhecimento de dados governamentais textuais.

2.1 Transparência Pública

Transparência pública está relacionada ao acesso à informação pública, sendo que, a Lei 12.527 (BRASIL, 2011), em seu artigo 4º, define informação pública da seguinte forma: "dados, processados ou não, que podem ser utilizados para produção e transmissão de conhecimentos, contidos em qualquer meio, suporte ou formato".

A transparência e a participação popular são princípios basilares da gestão democrática, que permitem aos cidadãos se informarem sobre a agenda proposta pelo governo e participarem das decisões sobre assuntos públicos e relacionados a seus legítimos interesses particulares (MARQUES, 2010).

A publicidade é outro conceito que tenta tornar possível a participação popular na gestão governamental. A publicidade é um dos princípios da Administração Pública apresentado no artigo 37 da Constituição federal (BRASIL, 1988).

Nesse contexto, também estão inseridos os dados abertos governamentais, que são dados livremente disponíveis para toda sociedade utilizar como desejar, sem restrições de licenças, patentes ou mecanismos de controle.

Logo, esses três conceitos; transparência, publicidade e dados abertos; são complementares e foram criados basicamente para atingir os mesmos objetivos: promover

a participação popular, permitir o controle social, prevenir a corrupção e estimular um melhor gerenciamento na Administração Pública. Porém, apesar das semelhanças, tais termos não são sinônimos e cada um tem funções específicas dentro da difícil tarefa de cooperar com a operacionalização de uma gestão democrática. Esses conceitos são descritos a seguir.

2.1.1 Publicidade

O artigo 37 da Constituição Federal enumera 5 princípios[3] que devem ser seguidos pela administração pública: legalidade, impessoalidade, moralidade, publicidade e eficiência. Logo, a questão da publicidade dos atos governamentais não é uma opção a ser escolhida pelo gestor público, mas sim um mandamento constitucional.

Segundo ALEXANDRINO e PAULO (2008), o princípio da publicidade apresenta uma dupla acepção: a primeira diz respeito à necessidade de publicação oficial dos atos administrativos, a fim de que eles possam produzir efeitos externos. Já a segunda está relacionada à necessidade de transparência da atividade administrativa como um todo.

A exigência de publicação oficial dos atos externos da Administração não é um requisito de validade dos atos administrativos, mas sim um pressuposto de sua eficácia. Assim, enquanto não verificada a publicação do ato, este não estará apto à produção dos efeitos perante seus destinatários externos ou terceiros (ALEXANDRINO; PAULO, 2008).

Entende-se como oficial a publicação do ato no Diário Oficial da União, no Diário Oficial dos Estados, Diário Oficial do DF e dos Municípios em que haja impressa oficial. Nos demais municípios (onde não existam diários oficiais), admite-se a fixação do ato na sede da prefeitura ou da Câmara, como forma de conferir publicidade ao ato. No entanto, esses meios de comunicação não alcançam grande parte da sociedade. Por essa razão, o segundo aspecto do princípio da publicidade diz respeito a transparência da atividade administrativa como um todo.

Com relação a esse segundo aspecto do princípio da publicidade, a Constituição ainda determina, no parágrafo primeiro do artigo 37, que "a publicidade dos atos, programas, obras, serviços e campanhas dos órgãos públicos deverá ter caráter educativo,

[3] Os princípios da Administração Pública são mandamentos que norteiam a atuação de todos os agentes públicos na execução de seus atos.

informativo ou de orientação social, dela não podendo constar nomes, símbolos ou imagens que caracterizem promoção pessoal de autoridades ou servidores públicos".

Logo, o princípio da publicidade está diretamente ligado a necessidade de se dar transparência aos atos públicos, a fim de propiciar aos cidadãos condições de acompanhar a atuação governamental.

2.1.2 Dados Governamentais Abertos

AGUNE, GREGORIO FILHO e BOLLIGER (2010) definem dados governamentais abertos (ou simplesmente governo aberto) como sendo a disponibilização, através da Internet, de informações e dados governamentais de domínio público para a livre utilização pela sociedade. Esses autores ressaltam ainda que de acordo com esse conceito, deve-se garantir o acesso a dados primários, de forma que o interessado possa combiná-los e cruzá-los para produzir novas informações e aplicações, colaborando com o governo na geração de conhecimento social a partir das bases governamentais.

Segundo DIETRICH et al. (2009), dado aberto é aquele que pode ser livremente utilizado, reutilizado e redistribuído por qualquer pessoa, e deve, no máximo, atribuir à fonte original e/ou fazer o compartilhamento desses dados utilizando a mesma licença em que as informações foram apresentadas.

Porém, independente da definição adotada, o fato é que para que um determinado conjunto de dados seja considerado como dado aberto, algumas condições devem ser respeitadas. Os dados governamentais são considerados abertos quando publicados de acordo com os 8 princípios[4] elencados por ativistas do governo aberto. Segundo esses princípios, para os dados serem considerados abertos, eles devem ser:

- Completos: Todos os dados públicos devem ser disponibilizados. Dado público é aquele que não está sujeito a restrições de privacidade, segurança ou outros privilégios.
- Primários: São apresentados tal como colhidos da fonte, com o maior nível possível de granularidade, sem agregação ou modificação. Por exemplo, um

[4] Esses princípios foram definidos em um encontro (que ocorreu em dezembro de 2007 na Califórnia, nos EUA) com representantes de empresas da área de tecnologia e membros de universidades e do governo.

gráfico não é considerado como dado aberto, mas os dados utilizados para construir a planilha que deu origem a ele podem ser abertos.

- Atuais: Devem ser publicados o mais rápido possível para preservar seu valor. Em geral, têm periodicidade: quanto mais recentes e atuais, mais úteis para seus usuários.
- Acessíveis: São disponibilizados para a maior quantidade possível de pessoas, atendendo, assim, aos mais diferentes propósitos.
- Compreensíveis por máquina: Devem estar estruturados de modo razoável, possibilitando que sejam processados automaticamente. Por exemplo, uma tabela em PDF é muito bem compreendida por pessoas, mas para um computador é apenas uma imagem, já uma tabela em formato estruturado, como CSV ou XML, é processada mais facilmente por softwares e sistemas.
- Não discriminatórios: Devem estar disponíveis para qualquer pessoa, sem necessidade de cadastro ou qualquer outro procedimento que impeça o acesso.
- Não proprietários: Nenhuma entidade ou organização deve ter controle exclusivo sobre os dados disponibilizados.
- Licenças livres: Não devem estar submetidos a copyrights, patentes, marcas registradas ou regulações de segredo industrial.

Eaves (2009) acrescenta ainda três leis para os dados governamentais abetos:

- Se o dado não for encontrado e indexado na web, ele não existe;
- Se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser aproveitado;
- Se algum dispositivo legal não permitir sua replicação, ele é inútil.

No entanto, se o governo não se empenhar em disponibilizar os seus dados, a cultura dos dados abertos nunca será difundida, e nesse ponto o Brasil tem conseguindo importantes avanços, tanto em nível nacional como em governos estaduais e municipais.

No contexto nacional, a principal iniciativa nesse sentido foi a criação da Infraestrutura Nacional de Dados Abertos (INDA). A INDA pretende, em um primeiro momento, reunir organizações comprometidas com a publicação de seus dados para padronizar as melhores práticas de governo aberto, e em uma segunda fase, pretende convencionar não só os padrões e tecnologias utilizadas, mas também os conceitos e a

forma como as diferentes áreas da administração pública visualizam e modelam seus negócios e seus sistemas (BATISTA; SILVA; MIRANDA, 2013).

Porém, os dados abertos do governo federal não são publicados em portais de transparência pública. Essas informações são disponibilizadas em sites específicos, e o maior repositório de dados abertos governamentais em nível nacional é o portal brasileiro de dados abertos[5], que é mantido pelo Ministério do Planejamento, e tem como objetivo centralizar a busca e o acesso a dados governamentais abertos.

O público que utiliza dados abertos não é o mesmo atendido pelos portais de transparência pública, pois, a utilização de dados abertos pressupõe uma certa capacidade para manipular dados em formatos originais. Já os portais de transparência têm o objetivo de apresentar os dados de uma forma mais fácil de ser entendida, sem a necessidade de se fazer cruzamentos de informações ou outras análises mais especializadas.

2.1.3 Portais de Transparência Pública

CAPPELLI e LEITE (2008) definem transparência organizacional como sendo a existência de políticas, padrões e procedimentos que visam fornecer aos interessados informações sobre a organização segundo características gerais de acesso, uso, qualidade de conteúdo, entendimento e auditabilidade. Esse conceito pode ser instanciado no contexto de transparência pública, sendo que, nesse caso, as informações prestadas dizem respeito a dados gerados, ou mantidos, pela administração pública.

O Brasil possui várias normas que enfatizam a necessidade da transparência pública, porém, duas leis específicas se destacam como marco legal referentes a esse assunto: a Lei 12.527 (BRASIL, 2011), Lei de Acesso à Informação, no que tange à transparência passiva, e a Lei Complementar 131 (BRASIL, 2009) referente à transparência ativa.

A distinção entre transparência ativa e passiva se dá pela forma como o governo dá acesso à informação. Na transparência ativa, a divulgação da informação se dá por iniciativa do próprio setor público, que torna públicas as informações de interesse geral ou coletivo, ainda que não tenha sido expressamente solicitado (CGU, 2013), principalmente pela Internet (ex. portais de transparência). Já na transparência passiva, a divulgação se dá quando algum órgão ou ente é demandado pela sociedade a prestar

[5] http://www.dados.gov.br/

informações, desde que tais informações não sejam resguardadas por sigilo (CGU, 2013). A obrigatoriedade de prestar as informações solicitadas está prevista especificamente no artigo 10 da Lei 12.527, como por exemplo em situações em que alguém solicita a algum órgão informações sobre um assunto específico. Na Figura 3 é ilustrada a diferença entre essas duas formas de fornecimento de acesso à informação.

Figura 3 - Transparência Ativa e Passiva (CGU, 2013)

Uma vez definido o que é transparência, o passo seguinte é se definir como se tornar transparente. Pois a simples disponibilização da informação não garante a transparência. Para se fazer transparente, é necessário que a informação disponibilizada seja efetiva, ou seja, que ela seja útil para a necessidade de conhecimento do usuário. De acordo com CAPPELLI; LEITE e ARAÚJO (2010) a transparência não é uma qualidade binária, mas algo que possui estágios evolutivos. CAPPELLI; LEITE e ARAÚJO (2010) propõem um modelo de estágios para se medir o grau de transparência. Segundo CAPPELLI (2009), a transparência pode ser dividida em degraus, e cada um desses degraus pode ser estabelecido através da institucionalização de algumas características. Dessa forma, CAPPELLI e LEITE (2008) definem 5 degraus de transparência, conforme apresentado na Figura 4.

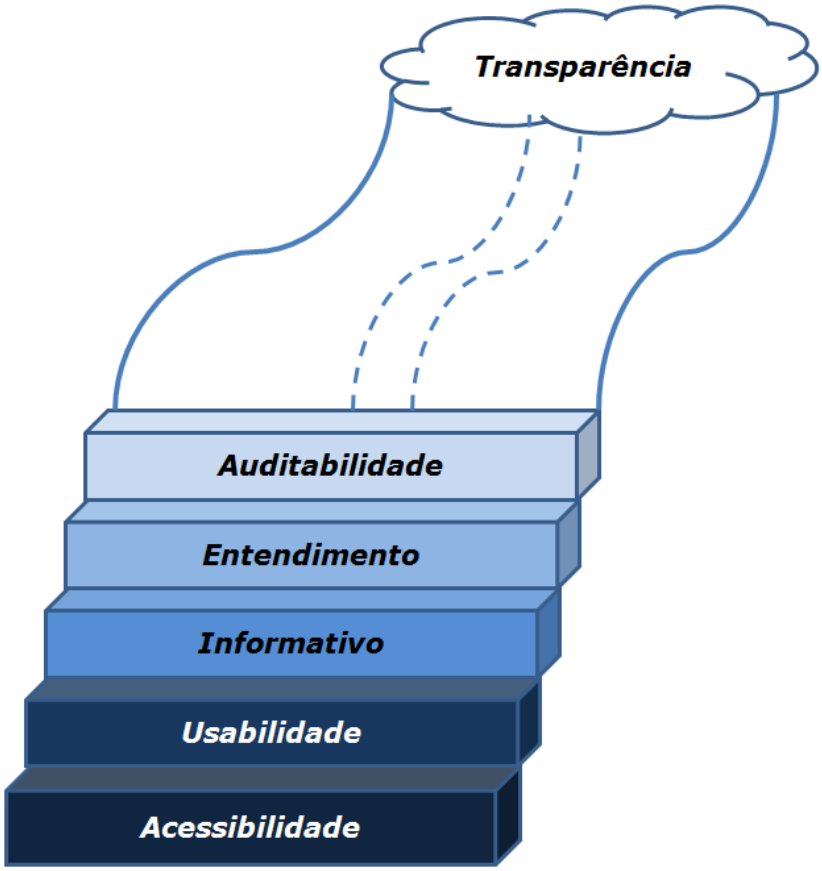


Figura 4 - Degraus da Transparência (CAPPELLI; LEITE, 2008)

CAPPELLI e LEITE (2008) define os 5 degraus de transparência da seguinte forma:

- Degrau 1 – Acessibilidade: A transparência é realizada através da capacidade de acesso. Esta capacidade é identificada através da aferição de práticas que efetivam características de portabilidade, disponibilidade e publicidade na organização.
- Degrau 2 – Usabilidade: A transparência é realizada através das facilidades de uso. Esta capacidade é identificada através da aferição de práticas que efetivam características de uniformidade, simplicidade, operabilidade, intuitividade, desempenho, adaptabilidade e amigabilidade na organização.
- Degrau 3 – Informativo: A transparência é realizada através da qualidade da informação. Esta capacidade é identificada através da aferição de práticas que efetivam características de clareza, completeza, corretude, atualidade, comparabilidade, consistência, integridade e acurácia na organização.
- Degrau 4 – Entendimento: A transparência é realizada através do entendimento. Esta capacidade é identificada através da aferição de práticas que efetivam

características de concisão, compositividade, divisibilidade, detalhamento e dependência na organização.

- Degrau 5 - Auditabilidade: A transparência é realizada através da auditabilidade. Esta capacidade é identificada através da aferição de práticas que efetivam características de validade, controlabilidade, verificabilidade, rastreabilidade e explicação das informações na organização.

Logo, de acordo com a classificação proposta por CAPPELLI; LEITE e ARAÚJO (2010) a simples publicação ou disponibilização dos dados só garante o atingimento do primeiro degrau de transparência. Dessa forma, os governos devem também se empenhar em implementar praticas que propiciem o atingimento dos demais degraus de transparência.

2.2 Mineração de Texto

Há muito tempo tem-se buscado técnicas para se extrair conhecimento de textos. Os primeiros trabalhos nesse sentido ocorreram ainda na década de 50. LUHN (1958) utilizou informações estatísticas, obtidas da distribuição de frequência das palavras, para calcular uma medida relativa de importância das palavras, a fim de sumarizar o conteúdo dos textos.

Com o passar do tempo e com o aumento da produção de conteúdo digital, também se aumentou o interesse pela extração de conteúdo de documentos no formato textual, assim como o número de trabalhos voltados para essa área. Várias subdivisões desse grande assunto sugiram, como por exemplo análise de sentimentos (YU; WANG, 2015), reconhecimento de entidades nomeadas (DERCZYNSKI et al., 2015), classificação de textos (ALTINEL; GANIZ; DIRI, 2015) e clusterização para identificar grupos de textos semelhantes (LIU et al., 2015).

No entanto, independente da finalidade do processo de mineração de texto, o fato é que a possibilidade de se obter vantagens pela identificação do conteúdo que está sendo expresso nos textos tem motivado as pesquisas nessa área. HEARST (1999) afirma que o processo de Mineração de Textos também pode ser descrito como um processo de identificação de informações desconhecidas de uma coleção de textos.

HU e LIU (2012) propõem um framework para análise textual composto por três passos consecutivos: pré-processamento de texto, representação do texto e descoberta de conhecimento, conforme mostrados na Figura 5.

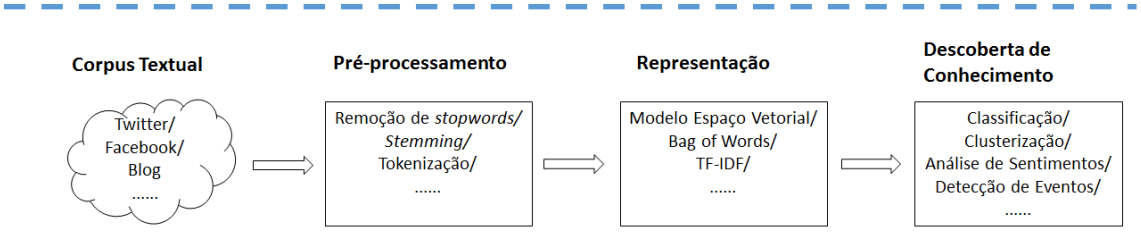


Figura 5 - Um framework tradicional para análise de texto, adaptado de (HU e LIU, 2012)

2.2.1 Pré-processamento de Texto

O pré-processamento do texto objetiva fazer o documento de entrada mais consistente para facilitar a representação textual (WEISS et al., 2010). A seguir são apresentadas as principais atividades que são executadas durante o pré-processamento dos textos.

- ***Case Folding***

Case Folding é o processo de converter todos os caracteres de um documento no mesmo tipo de letra – ou todas maiúsculas ou minúsculas. Isso tem a vantagem de acelerar comparações no processo de indexação.

- **Tokenização**

A tarefa de tokenização consiste na identificação dos termos dentro do texto. Tal atividade tem o objetivo de dividir um documento textual em unidades mínimas, denominadas tokens. Para o caso do segmento de texto: "CARNE BOVINA IN NATURA", a tokenização apresentaria como saída os seguintes tokens: "CARNE", "BOVINA", "IN" e "NATURA".

- **Retirada de *"stopwords"***

Quando uma palavra é muito comum, ou seja, ela aparece em uma frequência muito maior do que as demais palavras, essa palavra acaba se tornando irrelevante[6] para o processo de extração de conhecimento, pois tal palavra não irá trazer nenhum conhecimento relevante a respeito daquele texto. Normalmente essas palavras são preposições, artigos, pronomes, advérbios e outras palavras que não acrescentam muita informação ao contexto do domínio. Essas palavras são chamadas de *stopwords* e normalmente são retiradas do texto durante o pré-processamento.

A lista de palavras consideradas *stopwords* varia de idioma para idioma, podendo inclusive variar de uma região para outra, mesmo em localidades em que se fala o mesmo idioma. Uma palavra também pode ser considerada *stopwords* ou não de acordo com o contexto em que está sendo analisado. Por exemplo, quando se está analisando o texto de abstracts de artigos científicos, o termo "article" pode ser considerado uma *stopwords*, pois provavelmente deve aparecer em praticamente todos os abstracts ("this article ..."), no entanto, na análise de outros contextos essa palavra pode não ser tão comum, e por isso não deve ser considerada uma *stopword*.

- **Uso de dicionários ou Thesauros**

Um dicionário pode ser definido como um vocabulário controlado que representa sinônimos, hierarquias e relacionamentos associativos entre termos para ajudar os usuários a encontrar a informação de que eles precisam (LOPES, 2004). O thesaurus pode ser utilizado para mapear termos variantes (sinônimos, abreviações, acrônimos, e ortografias alternativas) em um único termo preferido para cada conceito, possibilitando que palavras grafadas de forma diferentes, mas com o mesmo significado, possam ser consideradas com o mesmo conteúdo informativo.

- ***Stemming* ou lematização**

Na escrita natural, muitas vezes a mesma palavra pode aparecer grafadas de diferentes formas, isso pode ocorrer devido a variações causadas pela inclusão de diferentes terminações ao radical original da palavra, como por exemplo, sufixos, desinências verbais ou desinência de gêneros. Essas variações trazem dificuldades ao

[6] Cabe ressaltar que essas palavras são consideradas irrelevantes apenas para os casos de extração de conhecimento automático de informações, por não trazerem diferenças semânticas quando se compara os textos. No entanto, dentro da estrutura gramatical de uma determinada língua elas são tão importantes quanto as demais, sendo essenciais para darem o sentido completo de um determinado texto.

processo de análise textual automatizada. A fim de minimizar essa variabilidade, utiliza-se um procedimento denominado *stemming*, cujo principal objetivo é transformar essas palavras em um formato normalizado.

O processo de *stemming* é realizado considerando cada palavra isoladamente e tentando reduzi-la a sua provável palavra raiz. Isto tem a vantagem de eliminar sufixos, indicativos de formas verbais e/ou plurais. Os algoritmos de stemming empregam artifícios da linguística e são dependentes do idioma.

Alguns exemplos de métodos de *stemmer* são: Stemmer S (HARMAN, 1991), que remove as terminações "ies", "es" e "s" das palavras de língua inglesa; stemmer de Porter (PORTER, 1980), que identifica as diferentes inflexões referentes a uma mesma palavra e faz a substituição pelo radical dessa palavra, e o método de Lovins (LOVINS, 1968), um método para a remoção de sufixo de palavras.

2.2.2 Representação Textual

Para minerar texto, é preciso processá-los em um formato em que os algoritmos de mineração de dados possam utilizá-lo (WEISS et al., 2010). A forma mais comum para modelar documentos é transformá-los em vetores numéricos e então trabalhar com operações de álgebra linear (HU e LIU, 2012). Essa representação é chamada de *"Bag of Words"* (BOW). Na BOW, uma palavra é representada como uma variável separada com um peso numérico de importância. No entanto, existem várias formas de se calcular esse peso da palavra, abaixo são apresentados os principais modelos de ponderação utilizados:

- Modelo Binário: esse método representa apenas a noção de existência ou inexistência do termo no documento, ou seja, caso o termo esteja presente no documento, ele recebe o valor 1, e caso ele não esteja presente nesse documento ele recebe o valor 0.
- Modelo baseado em Frequência Absoluta - mais conhecida como "term frequency" (tf), indica o número de ocorrências de uma palavra em um documento.
- Modelo baseado em Frequência Relativa - consiste na frequência absoluta (tf) normalizada pelo tamanho do documento (número de palavras nele contidas). Na Equação 1, tf_i indica a frequência absoluta de um determinado termo i, enquanto que n_i indica o número de ocorrências desse termo i, e $\sum_k n_k$ o número de todos os termos presentes no texto analisado.

$$tf_i = \frac{n_i}{\sum_k n_k} \qquad (1)$$

- Modelo baseado na Frequência Inversa de Documentos - Uma das maneiras mais utilizadas para se identificar o peso do termo é a aplicação da frequência inversa de documentos. Essa medida indica a importância de um termo em um documento, considerando todo o corpus a ser analisado. Dessa forma, o peso de um termo dentro de um documento aumentará a medida que o número de ocorrências desse termo nesse mesmo documento aumenta. Porém, esse valor irá diminuir se outros documentos também possuírem esse mesmo termo. Na Equação 2, $tfidf(i)$ indica a frequência inversa de um termo i, tf_i indica o número de ocorrências desse termo i no documento (frequência absoluta), $df(i)$ o número de documentos contendo o termo i, e N o número de todos os documentos

$$tfidf(i) = tf_i . \log(\frac{N}{df(i)}) \qquad (2)$$

A seguir são apresentados três seguimentos de texto distintos e a Figura 6 mostra a representação da BOW, utilizando-se os modelos de representação previamente apresentados.

- Segmento 1: CARNE BOVINA IN NATURA CARNE BOVINA
- Segmento 2: PEIXE IN NATURA FILÉ DE PEIXE
- Segmento 3: CARNE IN NATURA PEITO DE FRANGO

	CARNE	*BOVINA*	*IN*	*NATURA*	*PEIXE*	*FILÉ*	*DE*	*PEITO*	*FRANGO*
Segmento 1	1	1	1	1	0	0	0	0	0
Segmento 2	0	0	1	1	1	1	1	0	0
Segmento 3	1	0	1	1	0	0	1	1	1

(a) Representação Binária

	CARNE	*BOVINA*	*IN*	*NATURA*	*PEIXE*	*FILÉ*	*DE*	*PEITO*	*FRANGO*
Segmento 1	2	2	1	1	0	0	0	0	0
Segmento 2	0	0	1	1	2	1	1	0	0
Segmento 3	1	0	1	1	0	0	1	1	1

(b) Representação por Frequência Absoluta

	CARNE	*BOVINA*	*IN*	*NATURA*	*PEIXE*	*FILÉ*	*DE*	*PEITO*	*FRANGO*
Segmento 1	0,333	0,333	0,167	0,167	0	0	0	0	0
Segmento 2	0	0	0,167	0,167	0,333	0,167	0,167	0	0
Segmento 3	0,167	0	0,167	0,167	0	0	0,167	0,167	0,167

(c) Representação por Frequência Relativa

	CARNE	*BOVINA*	*IN*	*NATURA*	*PEIXE*	*FILÉ*	*DE*	*PEITO*	*FRANGO*
Segmento 1	0,352	0,954	0	0	0	0	0	0	0
Segmento 2	0	0	0	0	0,954	0,477	0,176	0	0
Segmento 3	0,176	0	0	0	0	0	0,176	0,477	0,477

(d) Representação por Frequência Inversa – TF-IDF

Figura 6 - Representações da Bag of Words

Nos modelos de representação baseados em BOW, cada um dos termos é considerado uma variável, e os valores associados a cada um desses termos é o valor dessa variável para o documento em questão, segundo o modelo utilizado. Logo, cada documento é representado por um conjunto de variáveis (cada termo será uma variável). Dessa forma, utilizando-se a BOW e tomando como exemplo o modelo baseado na frequência absoluta, apresentado na Figura 6(b), a variável CARNE possui os valores 2, 0 e 1 para os segmentos de texto 1, 2 e 3, respectivamente.

A utilização de um modelo de representação baseado em BOW tem a vantagem de vetorizar os segmentos ou documentos, no processo de mineração de textos, permitindo que técnicas baseadas em vetores possam ser utilizadas para definir a similaridade entre documentos, dentre outras atividades.

2.2.3 Descoberta de Conhecimento

A partir do momento em que os corpus de textos são transformados em vetores, pode-se aplicar métodos de aprendizado de máquina a esses dados, como por exemplo algoritmos de classificação e clusterização (HU e LIU, 2012).

- **Clusterização**

Clusterização é uma abordagem de aprendizado não supervisionado que tem o objetivo de agrupar um conjunto de objetos, a fim de maximizar a sua similaridade dentro do mesmo grupo (denominado cluster) e minimizar a semelhança que eles têm para objetos em outros grupos (clusters) (JAIN; MURTY; FLYNN, 1999). Em Mineração de texto, os algoritmos de clusterização são utilizados para agrupar documentos similares. Dessa forma, a medida de similaridade é o elemento chave, sendo que, em clusterização de documentos, a medida de similaridade de cosseno é a mais comumente utilizada.

Uma vez definidos os clusters de documentos, outro desafio é a definição de rótulos capazes de representar o conteúdo principal do grupo de documentos pertencentes a cada um dos clusters.

- **Categorização**

A categorização visa identificar os tópicos principais em um documento e associar este documento a uma ou mais categorias pré-definidas (YANG; PEDERSEN, 1997). Segundo (WEISS et al., 2010), a categorização é a tarefa de aprender uma função alvo H que mapeia cada conjunto de atributos x para um rótulo de classes y pré-definido. Dessa forma, dado um conjunto de documentos rotulados com as suas respectivas categorias, denominado conjunto de treinamento, o objetivo é construir um modelo capaz de atribuir o rótulo correto de um novo conjunto de documentos, não categorizados, a partir do conhecimento obtido pelo conjunto de dados de treinamento.

- **Sumarização**

A sumarização é um processo de redução da quantidade de texto em um documento, sem perder o seu significado chave (HAHN; MANI, 2000). LOPES (2004) distingue 2 tipos de sumarização: sumarização por abstração e sumarização por extração.

A sumarização por abstração faz um resumo de forma análoga ao processo utilizado pelos humanos, ou seja, primeiro abstrai o conhecimento presente no texto, e

depois apresenta as principais ideias desse texto através da escrita de um resumo. Conforme citado por (LOPES, 2004), apesar de já terem sido feitas algumas pesquisas nessas áreas, a criação de resumos por abstração não é uma alternativa viável.

Já na sumarização por extração, sentenças inteiras ou parágrafos são copiados do documento original numa tentativa de construir um texto menor que ainda conserve as principais ideias do documento original (LOPES, 2004). Esse tipo de sumarização é baseado na medida de importância relativa dos documentos.

Essa dissertação não utiliza nenhum desses métodos de descoberta de conhecimento citados, pois é proposto um método específico para o problema em questão. No entanto, o resultado obtido pelo método de descoberta de conhecimento em dados textuais proposto se assemelha aos resultados obtidos com a utilização do método de clusterização, pois as descrições de compras de produtos iguais são grupadas em um mesmo conjunto, assim como textos semelhantes são grupados em um mesmo cluster no método de clusterização.

2.3 Processamento Intensivo de Dados

Atualmente, a escalada do volume e variedade de dados vem trazendo dificuldades para o processamento dessas informações, inclusive para o caso de informações governamentais. LIN; DYER (2010) apontam que a única abordagem viável para esse tipo de problema é a dividir e conquistar. A estratégia básica dessa abordagem é particionar um problema grande em subproblemas menores. Dessa forma, os subproblemas independentes podem ser tratados de forma paralela por diferentes *threads*, núcleos de processadores ou máquinas em clusters (LIN; DYER, 2010).

Porém, em ambientes tradicionais de programação paralela ou distribuída, o programador precisa tratar explicitamente questões complexas, como por exemplo: forma de paralelização, tolerância a falhas, distribuição de dados entre os nós de processamento, balanceamento de carga, acesso à memória compartilhada, dentre outros.

O MapReduce, introduzido por (DEAN; GHEMAWAT, 2001), é um modelo de programação paralela para grandes volumes de dados. Ele é inspirado na estratégia dividir e conquistar, mas abstrai do programador a complexidade dos problemas típicos do gerenciamento de aplicações distribuídas, permitindo que o desenvolvedor possa se dedicar apenas na solução do problema a ser tratado, deixando que a aplicação execute a distribuição e o paralelismo.

Neste modelo, que roda em clusters[7] de computadores, as tarefas de processamento são distribuídas entre os nós (máquinas do cluster), implementando uma arquitetura mestre-escravo, na qual, um nó mestre é responsável pelo gerenciamento, e os nós escravos são responsáveis pelo processamento propriamente dito.

No modelo mapreduce, o programador só precisa desenvolver duas funções principais: Map e Reduce (DEAN; GHEMAWAT, 2008). Essas funções trabalham basicamente com pares de chaves e valores, sendo que a função Map gera um conjunto de pares (chave-valor) intermediários, que são passados para a função Reduce, que é responsável por processar e juntar todos esses pares (chave-valor) intermediários, associando os pares com a mesma chave em uma espécie de sumarização dos resultados. Entre as operações de Map e Reduce, esses pares de chave e valor passam pela operação de *"Shuffle and Sort"*. Nessa fase, que ocorre de forma automatizada, os pares chave-valor, gerados pelos nós que executam funções *map*, são distribuídos para os nós responsáveis pelas funções de *reduce*. No entanto, antes que a função *reduce* comece, esses pares são ordenados pela chave, de forma que a função *reduce* já receba os pares de maneira ordenada pela chave.

Existem várias ferramentas que implementam o modelo de programação MapReduce, dentre as quais, o Hadoop [8]e o Apache Spark[9] são as mais conhecidas. Na Figura 7 é apresentado o sistema de execução de programas MapReduce no Hadoop (WHITE, 2012).

DEAN; GHEMAWAT (2001) descrevem a execução de um programa MapReduce, apresentada na Figura 7, pelos seguintes passos: (i) a base de dados de entrada, contendo os arquivos a serem analisados, é armazenada em um sistema de arquivo distribuído composto por diversos registros divididos em blocos. Antes do início da execução propriamente dita, cópias do programa desenvolvido pelo usuário (funções de Map e Reduce) são distribuídas por todos os nós (máquinas) do cluster (fluxo 1 da Figura 7); (ii) o modelo implementa uma arquitetura mestre-escravo, e as tarefas de Map e Reduce (definidas pelo usuário) são atribuídas aos nós escravos pelo nó mestre (fluxos 2 da Figura 7); (iii) um nó escravo recebe a tarefa Map a ser executada tendo como entrada um bloco de dados designado pelo nó mestre (fluxo 3 da Figura 7), sendo que, após o

[7] Clusters de computadores são máquinas interligadas que trabalham de forma conjunta a fim de executar uma mesma tarefa de processamento.
[8] http://hadoop.apache.org/
[9] http://spark.apache.org

processamento da função Map, o nó escravo armazena os pares (chave-valor) intermediários em uma memória temporária, para posterior envio à fase seguinte; (iv) periodicamente, os pares produzidos pelos nós responsáveis pela execução da função Map são escritos em discos locais (fluxo 4 da Figura 7), sendo que a localização desses discos locais são passadas para o nó mestre, que é responsável por passar a localização desse armazenamento temporário aos escravos responsáveis pela execução da tarefa de Reduce; (v) quando um nó escravo, responsável pela tarefa de Reduce, é notificado pelo mestre sobre essa localização, usa chamadas remotas para ler esses dados no disco local do escravo que executou a tarefa de Map (fluxo 5 da Figura 7); (vi) o nó escravo responsável pela função de reduce itera esses dados intermediários ordenados e para cada chave ele agrega os valores correspondentes, de acordo com a função reduce definida pelo usuário; (vii) após a conclusão de todas as tarefas de map e reduce, o resultado da execução é disponibilizado em arquivos de saída (fluxo 6 da Figura 7).

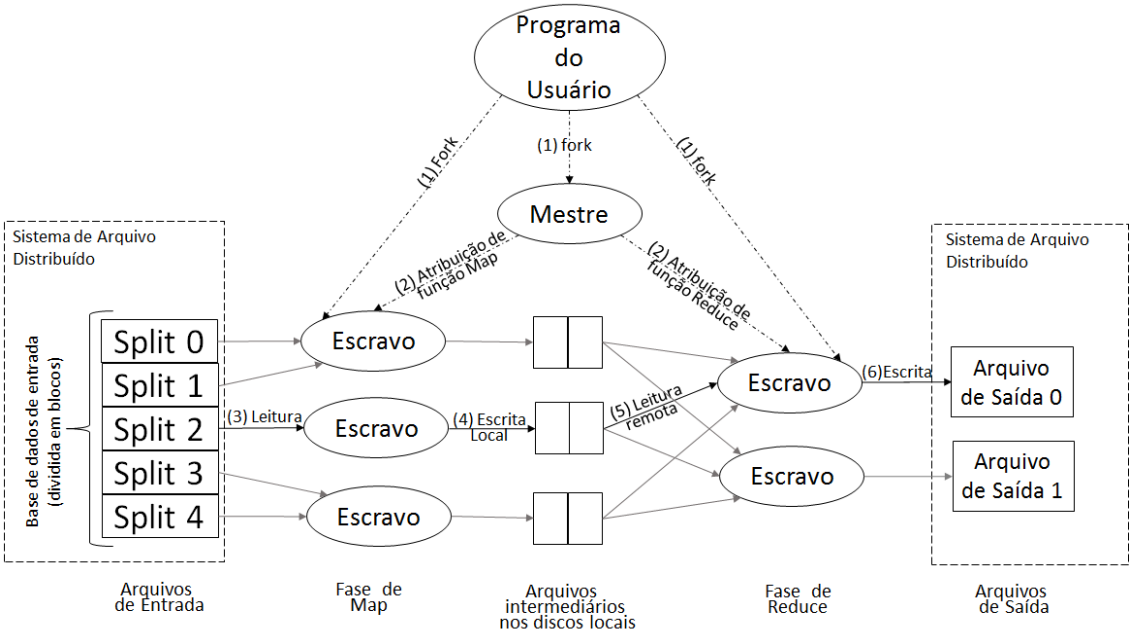


Figura 7 - Modelo de execução do MapReduce, adaptado de (DEAN; GHEMAWAT, 2001)

Como observado na Figura 7, os dados são lidos de um sistema de arquivos distribuído, em seguida são repassados para as funções Map e Reduce, e por fim são gravados novamente no sistema de arquivos distribuído. Entre o fluxo de dados das funções Map e Reduce, há a transferência de dados via rede, e o armazenamento dos resultados intermediários no sistema de arquivo local das máquinas responsáveis pelas execuções das funções Reduce. Nesse caso, a única forma de compartilhamento de dados entre dois processos em execução no MapReduce consiste em uma operação de escrita

no sistema de arquivos distribuído que, por sua vez, é lido pelo outro processo, representado por várias iterações. Tal operação demanda tempo de acesso a disco, que muitas vezes pode retardar a execução de uma aplicação, tendo em vista que o tempo de acesso a disco pode ser custoso.

Para tratar essa limitação, a arquitetura do Apache Spark propõe a utilização de Resilient Distributed Datasets (RDDs) (ZAHARIA et al., 2012), que são uma abstração em que os dados são armazenados na memória. Um RDD é uma coleção de registros que são distribuídos ou particionados pelos nós do clustre (GULLER, 2015). O RDD é tolerante a falha, logo, se um dado nó ou tarefa falhar, o RDD pode ser reconstruído automaticamente nos nós remanescentes para permitir a conclusão da tarefa. Dessa forma, o Spark armazena os resultados intermediários na memória, em vez de escrevê-los no disco, o que confere mais velocidade ao processamento.

Os RDDs, implementados no Spark, suportam outros tipos de operações além das tradicionais operações de Map e Reduce. Essas operações se dividem em dois grupos: Transformações e Ações. As Transformações criam novos conjuntos de dados (novos RDDs) a partir de um já existente, enquanto que as ações retornam um valor depois de executar uma computação no conjunto de dados (no RDD). A Tabela 1 demonstra algumas operações[10] suportadas pelo pelos RDDs no Spark, a título de exemplo.

Tabela 1 - Exemplos de operações suportadas pelos RDDs no Spark

Tipo de Operação	Operação	Descrição
Transformação	map(func)	Retorna um novo RDD formado pela passagem de cada elemento pela função func
Transformação	Filter(func)	Retorna um novo RDD formado pela seleção dos elementos dados pela função func
Transformação	FlatMap(func)	Similar à função map; no entanto cada elemento pode ser mapeado para 0 ou mais itens
Transformação	distinct([numTasks]))	Retorna os elementos únicos de um conjunto de dados
Transformação	union(otherDataset)	Retorna os elementos presentes na união de RDDs
Transformação	groupByKey([numTasks]))	Retorna o conjunto dos elementos de uma chave específica K

[10] Não foram listadas todas as transformações e ações suportadas pelo Spark, por esse não ser o foco dessa dissertação, porém, a lista de todas as transformações e ações podem ser encontradas na documentação oficial do Spark, disponível em http://spark.apache.org.

Transformação	reduceByKey(func, [numTasks])	Retorna o conjunto dos elementos de uma chave específica K, utilizando uma função Reduce informada
Ação	reduce(func)	Retorna a agregação dos elementos de acordo com a função func
Ação	collect()	Retorna todos os elementos de um conjunto de dados
Ação	count()	Retorna o número de elementos de um conjunto de dados
Ação	First()	Retorna o primeiro elemento de um conjunto de dados
Ação	take(n)	Retorna a quantidade n de elementos de um conjunto de dados takeSample(withReplacement, num, seed) Retorna um subconjunto do RDD

Considerando-se o grande volume de dados que compõem as bases dos portais de transparência pública, soluções que tenham o objetivo de tratar informações disponibilizadas em tais portais, sem se restringir a uma delimitação específica de tempo, ou seja, sem delimitar o tamanho da base de dados a ser tratada, carecem de aplicações que façam uso de técnicas e ferramentas específicas para o processamento de grandes volumes de dados.

2.4 Trabalhos Relacionados

Atualmente, existe uma série de trabalhos que se propõem a extrair informações relevantes de dados textuais gerados pela Administração Pública. Nesse sentido, MAIA (2015) propõe um classificador automático das denúncias que compõem o banco de denúncias recebidas pela Controladoria Geral da União. O objetivo do trabalho é ler os textos das diferentes denúncias recebidas e classificá-las dentre as 91 categorias de denúncias existentes. O autor avaliou os resultados obtidos por diferentes algoritmos de classificação: SVM (HEARST et al., 1998), Random Forest (BREIMAN, 2001), Naive Bayes (LANGLEY; IBA; THOMPSON, 1992) e Árvore de Decisão (QUINLAN, 1986).

Apesar do SVM ter sido o algoritmo que apresentou melhores resultados, esse ainda ficou aquém do esperado. Sendo assim, optou-se pela adoção de uma abordagem que utilizava indução de um classificador baseado em árvore de Huffman.

A codificação de Huffman (HUFFMAN, 1952) é um método de compactação que usa as probabilidades de ocorrência dos símbolos no conjunto de dados a ser compactado para determinar códigos de tamanhos variáveis para cada símbolo. O uso dessa

codificação permite que os símbolos sejam representados de forma binária. A árvore de Huffman é uma árvore binária utilizada para representar cada um dos símbolos de acordo com sua codificação binária.

A abordagem proposta por MAIA (2015) constrói uma árvore de Huffman para cada classe considerada. Dessa forma, o procedimento de classificação consiste em submeter os documentos a cada uma das árvores de Huffman e escolher como classe do documento aquela correspondente à árvore cuja codificação seja a mais curta (MAIA, 2015).

CARVALHO et al. (2013) e CARVALHO et al. (2014a) sugerem uma metodologia para a formulação de um banco de preço da Administração Pública Federal a partir dos dados de compras que são apresentados no Portal da Transparência do Governo Federal. Essas compras vêm descritas em formato textual e carecem do emprego de técnicas de mineração de texto para se extrair o produto correspondente a cada uma das descrições de compras.

A abordagem proposta está dividida em 6 passos. Primeiro são selecionadas, do banco de dados do portal, todas as notas de empenho[11] referentes a um determinado período de tempo. Depois, para cada uma dessas notas de empenho são recuperados os códigos de material das compras descritas. Esse código é uma informação oriunda do SIASG[12] e está presente no campo textual de descrição do produto. O passo seguinte é a filtragem do conjunto de dados referente a um código de material específico. No quarto passo, utiliza-se esses resultados da filtragem e emprega-se um novo filtro, baseado na utilização de palavras chave a fim de se determinar um produto específico. Posteriormente, filtra-se o conjunto de dados resultante por faixa de preços, e finalmente calcula-se o preço de referência para o produto em questão.

O primeiro e segundo passo são executados com o auxílio de uma ferramenta de ETL (Extract, Transform, Load). Já o terceiro passo (a filtragem pelo código de material) é executado através de consultas SQL (Structured Query Language), diretamente no banco de dados. Na filtragem por palavras chaves (quarto passo), especialistas definem quais palavras devem estar contidas e quais palavras não podem estar presentes na

[11] Documento utilizado para registrar as operações que envolvem despesas orçamentárias realizadas pela Administração Pública e que indica o nome do credor, a especificação e a importância da despesa, bem como a dedução desta do saldo da dotação própria (Brasil, 1964).
[12] Sistema Integrado de Administração de Serviços Gerais (SIASG) - sistema que realiza as operações de compras governamentais dos órgãos da Administração Pública Federal direta, autárquica e fundacional.

descrição de uma determinada compra, para que um determinado produto possa ser caracterizado. Isso permite a identificação dos produtos.

No entanto, mesmo após a caracterização do produto, ainda há uma grande variabilidade na faixa de preço paga. Essas diferenças, em muitas situações decorre das diferentes formas de se quantificar um produto (por exemplo, diferentes unidades de medidas). Sendo assim, durante o passo 5 são aplicadas técnicas de clusterização, para cada grupo de produtos identificados, considerando-se que produtos quantificados de forma igual ficam em um mesmo cluster. Ainda nesse passo, os especialistas definem rótulos para cada um dos clusters gerados, sendo que um produto será totalmente caracterizado a partir da combinação entre o nome do produto (identificado a partir da combinação de palavras chaves) com o rótulo definido pelos especialistas. Esses rótulos são escolhidos em uma lista que traz as palavras com maior probabilidade de definir um determinado cluster (definidos para cada um dos clusters resultante da técnica de clusterização). Finalmente, após a qualificação dos produtos, utiliza-se os preços pagos por tais produtos a fim de se calcular uma faixa de preço de referência para esse produto. Nessa abordagem, especialistas precisam definir qual conjunto de palavras deve ser utilizado para caracterizar cada um dos produtos definidos como identificáveis. A definição de quais produtos irão compor o banco de preços também é feita pelos especialistas. Dessa forma, a principal contribuição desta dissertação em relação a esses dois trabalhos é a criação de uma forma automatizada para a identificação dos produtos, sem a necessidade de intervenção de especialistas.

CARVALHO et al. (2014a) usam redes bayesianas para identificar e prevenir o fracionamento de compras, uma espécie de fraude utilizada para burlar o processo licitatório exigido por lei. No Brasil, compras inferiores a um determinado valor (R$ 8.000,00) são dispensadas do procedimento licitatório. No entanto, uma fraude comum, para enquadrar compras de valores superiores nesse tipo de dispensa é o particionamento de uma mesma compra em várias outras de valores inferiores ao limite definido por lei. O objetivo principal desse trabalho é tentar identificar as compras consideradas suspeitas, a fim permitir que providencias possam ser tomadas antes da consumação de um gasto irregular.

Essa identificação de compras suspeitas é feita através do uso de redes bayesianas e utiliza uma série de atributos estruturados durante o processo de classificação. No entanto, também se faz necessária a identificação dos produtos que estão sendo especificados de forma textual nos editais de compra.

MARZAGÃO (2015) apresenta uma outra abordagem para o problema de identificação de produtos e serviços que são adquiridos pela Administração Pública. Esse trabalho utiliza um cadastro de materiais e serviços adotados pelo Governo Federal no sistema SIASG como dado de treinamento, e a partir deste cadastro, tenta classificar as compras utilizando o algoritmo de máquina de suporte vetorial.

Essa abordagem atingiu uma acurácia de 83,35%, e segundo o autor os erros encontrados foram ocasionados por duas causas principais: falhas no conjunto de dados de treinamento e problemas de frequência de classes, pois algumas classes de produtos, por não serem compradas frequentemente, não forneciam informações suficiente para o algoritmo de aprendizado de máquinas.

Saindo do contexto da Administração Pública, mas ainda dentro do desafio de se extrair informações de dados textuais, algumas iniciativas têm se destacado no sentido de utilizar o modelo de *Bag of Phrases*[13], em um contraponto ao tradicional *Bag of Words*. Dentre essas iniciativas estão (REN et al., 2015), (LIU et al., 2015) e (EL-KISHKY et al., 2014). Essas abordagens, ao invés de trabalharem com os tokens de forma individualizada, consideram sequências de tokens, que formam frases, a fim de agregar mais expressividades as variáveis tratadas.

Logo, a principal contribuição dessa dissertação em relação aos demais trabalhos que também se propõem a extrair informações de dados textuais governamentais é a proposta de uma técnica de extração de conhecimento baseada no modelo *Bag of Phrases*, capaz de minerar as frases que melhor representam o conteúdo de um determinado texto e que não exige um conjunto de dados previamente rotulados e nem tem a necessidade de intervenção de especialistas durante o processo de descoberta de conhecimento.

[13] *Bag of Phrase*: modelo de representação utilizado no tratamento de dados textuais. Nesse modelo o texto é representado pela contagem das frases que o compõem, ignorando-se a gramática e a ordem das frases.

Capítulo 3 – Proposta

E ste capítulo descreve a solução proposta para o problema apresentado, bem como os passos seguidos para obtê-la. Conforme mencionado no capítulo 1, essa pesquisa utilizou a metodologia CRISP-DM (Cross Industry Standard Process for Data Mining) para o processo de mineração de texto. Dessa forma, foram seguidos os passos enunciados em (SHEARER, 2000) para se chegar ao modelo apresentado. As quatro primeiras fases do CRISP-DM: entendimento do negócio, entendimento dos dados, preparação dos dados e modelagem são apresentadas nesse capítulo. A fase 5, avaliação, é discutida no Capítulo 4. A fase de implantação não é tratada por estar fora do escopo dessa dissertação.

3.1 Entendimento do Negócio

O primeiro passo para o desenvolvimento da solução proposta foi o entendimento do negócio, ou seja, do contexto em que essa pesquisa está inserida. Sendo assim, fez-se necessário um aprofundamento nos estudos sobre os portais de transparência, bem como das obrigatoriedades de publicações dos dados orçamentários que esses portais devem atender.

O orçamento é uma lei anual que faz a previsão de receitas e a fixação de despesas para todo o ano, sendo que, cada ente federativo (União, estados, municípios e Distrito Federal) possui o seu próprio orçamento. Esse orçamento deve ser equilibrado, ou seja, a previsão de receita deve ser igual à fixação das despesas, e nenhuma despesa que não esteja prevista no orçamento anual poderá ser executada.

Visando dar mais transparência à execução desse orçamento, foi publicada a Lei complementar 131 (BRASIL, 2009), que determina a disponibilização, em tempo real, de informações pormenorizadas sobre a execução orçamentária da União, dos Estados, do

Distrito Federal e dos Municípios. Dessa forma, os entes federativos devem publicar diariamente, em sites de transparência pública, informações relativas as suas receitas e despesas.

Tal fato fez com que se criassem infraestruturas de atualização capazes de receber de forma automática dados relativos a arrecadação de receitas e a execução de despesas, a fim de se manter a carga diária desses portais.

Como a Lei exige que todas as receitas e despesas sejam evidenciadas nesses sites, isso faz com que se tenha que trabalhar com grandes volumes de dados, sendo que, este volume está diretamente ligado ao montante dos orçamentos de cada um dos entes federativos. Por exemplo, no caso do portal da Transparência do Governo Federal, que apresenta dados relativos ao Poder Executivo Federal, são inseridos anualmente um volume aproximado de 100 GB, sendo que algumas dessas informações não são estruturadas, o que dificulta a análise desses dados por parte do cidadão, e inviabiliza uma série de análises que poderiam ser feitas caso esses dados fossem devidamente tratados.
A execução orçamentária está dividida em duas partes, a Execução do orçamento da despesa e a execução do orçamento da receita. No entanto, esse trabalho está focado na parte da execução da despesa. A despesa pública possui uma sequência de três estágios que deve ser respeitada: Empenho, Liquidação e Pagamento.

O empenho, primeiro estágio da despesa, é uma garantia ao fornecedor de que a repartição pública tem autorização legal para realizar o gasto. Ele oferece um documento denominado nota de empenho como suporte para essa despesa.

O artigo 60 da Lei 4320 (BRASIL, 1964) veda a realização de despesa sem o empenho prévio. Quando um empenho é emitido, a Administração faz uma reserva do seu orçamento no valor desse empenho a fim de evitar gastos acima da sua capacidade de pagamento. É importante ressaltar que o empenho ainda não cria nenhuma obrigação de pagamento para a Administração Pública, ele apenas reserva uma parte do orçamento para um provável pagamento futuro, porém, esse pagamento pode não vir a se concretizar e a reserva pode ser cancelada.

O empenho é a fase da execução da despesa em que o gasto é descrito de forma mais detalhada. A Nota de Empenho específica o que será comprado, bem como a quantidade e o valor pago em cada uma das compras. Todas as demais fases da execução da despesa se baseiam no empenho.

O segundo estágio da despesa é a liquidação. Nesse momento, o Estado reconhece o compromisso de pagar aos seus fornecedores. A liquidação pode ser entendida como o

ateste do recebimento do material ou serviço e consequentemente o reconhecimento por parte da Administração da sua obrigação de pagar.

O Pagamento é o último estágio da execução da despesa. É nessa fase que a Administração faz o desembolso dos valores necessários para o efetivo pagamento dos fornecedores.

Todas essas fases são sequenciais, ou seja, para que um determinado estágio ocorra, é necessário que o seu estágio anterior já tenha ocorrido. Logo, não se pode liquidar despesa que ainda não tenha sido previamente empenhada e nem tão pouco é possível realizar o pagamento sem a anterior liquidação. Na Figura 8 é ilustrada a sequência de estágios da execução da despesa pública.

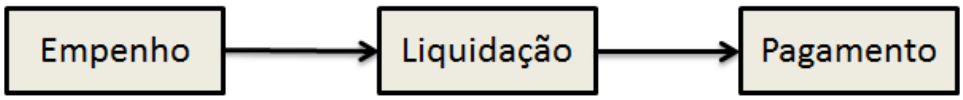


Figura 8 - Estágios da Execução da Despesa

3.2 Entendimento dos Dados

Pelo fato do empenho ser a fase em que o gasto é descrito de forma mais detalhada, essa pesquisa irá usar as informações da Nota de Empenho (documento utilizado para formalizar o estágio do empenho na execução da despesa). Os dados desse documento estão divididos em duas tabelas: tabela de Empenhos e tabela de Itens de Empenho.

Essas tabelas estão relacionadas, sendo que, uma nota de empenho corresponde a um registro na tabela de Empenho e a um ou mais registros na tabela de Itens de Empenho. Ou seja, um empenho pode ter vários itens de empenho. Cabe ressaltar que todas as despesas possuem um empenho específico, porém, nem todas as despesas correspondem a compra de materiais; visto que, existem outros tipos de gastos que não dizem respeito à compra de produtos, como por exemplo contratação de serviços e pagamento de pessoal; e por isso não estão dentro do escopo dessa pesquisa. Dessa forma, existem empenhos que devem ser desconsiderados no momento do desenvolvimento da proposta.

A tabela de empenhos traz informações gerais da Nota de Empenho, como por exemplo: quem está executando aquele gasto, quem será o favorecido da despesa, a data de emissão do empenho, classificação da despesa, e etc. Já a tabela de itens de Empenho traz informações relativas a cada um dos itens que estão sendo pagos. Dessa forma, caso um determinado órgão esteja comprando, de um mesmo fornecedor cinco itens distintos,

cada um desses itens corresponderá a uma linha na tabela de itens de empenho. Na Figura 9.a é apresentada uma tela de empenho do Portal da Transparência do Governo Federal, com a indicação das tabelas fontes de cada uma das informações apresentadas. Na Figura 9.b é apresentado um modelo de dados resumido que mostra o relacionamento entre as tabelas que armazenam os dados de uma Nota de Empenho. Nesse modelo de dados resumido só estão sendo apresentados alguns campos de cada uma das tabelas, visto que, essas tabelas possuem uma grande quantidade de campos distintos.

Data:	19/09/2016
Tipo de Empenho:	ORDINARIO — **Espécie de Empenho:** Original
Órgão Superior:	39000 - MINISTERIO DOS TRANSPORTES
Órgão / Entidade Vinculada:	39252 - DEPTO.NAC.DE INFRA+ESTRUT.DE TRANSPORTES-DNIT
Unidade Gestora Emitente:	393020 - SUPERINTENDENCIA REG. NO ESTADO MT - DNIT
Gestão:	39252 - DEPTO. NAC. DE INFRA+ESTRUTURA DE TRANSPORTES
Favorecido:	05.383.313/0001-90 - NOGUEIRA NOBRE COMERCIO E SERVICOS LTDA - ME
Valor:	R$ 6.549,90
DADOS DETALHADOS	
Observação do Documento:	EMPENHO QUE SE EFETIVA PARA COBRIR DESPESAS COM MATERIAL DE EXPEDIENTE PARA ATENDER AS NECESSIDADES DA SR/MT. PROC ORIGEM: 2016PR00280
Esfera:	1 - ORÇAMENTO FISCAL — **Tipo de Crédito:** A - INICIAL (LOA)
Grupo da Fonte de Recursos:	1 - RECURSOS DO TESOURO - EXERCÍCIO CORRENTE
Fonte de Recursos:	00 - RECURSOS ORDINARIOS
Unidade Orçamentária:	39252 - DEPTO.NAC.DE INFRA+ESTRUT.DE TRANSPORTES-DNIT
Funcional Programática	
Função:	26 - TRANSPORTE
Subfunção:	122 - ADMINISTRACAO GERAL
Programa:	2126 - PROGRAMA DE GESTAO E MANUTENCAO DO MINISTERIO DOS TRANSPORTE
Ação:	2000 - ADMINISTRACAO DA UNIDADE — **Linguagem Cidadã:** Administração de unidade
Subtítulo (localizador):	0001 - ADMINISTRACAO DA UNIDADE - NACIONAL
Plano Orçamentário - PO:	0000 - ADMINISTRACAO DA UNIDADE — **Autor da Emenda:** SEM EMENDA
Categoria de Despesa:	3 - Despesas Correntes — **Grupo de Despesa:** 3 - Outras Despesas Correntes
Modalidade de Aplicação:	90 - Aplic. Diretas (Gastos Diretos do Governo Federal)
Elemento de Despesa:	30 - MATERIAL DE CONSUMO
Processo Nº:	50611003665201621
Modalidade de Licitação:	PREGAO — **Inciso:** — **Amparo:**
Referência da Dispensa ou Inexigibilidade:	
Nº Convênio / Contrato de Repasse / Termo de Parceria / Outros:	

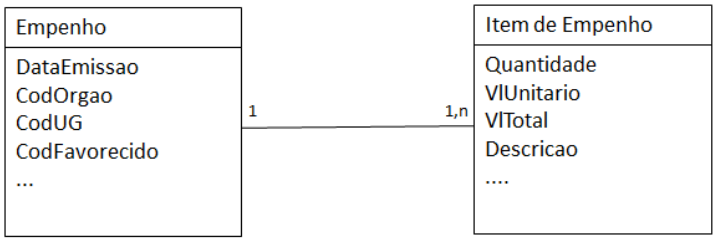
Detalhamento do Gasto

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
16 - MATERIAL DE EXPEDIENTE	50	2,80	140,00	50,00000 UNIDADE ALMOFADA CARIMBO, MATERIAL CAIXA PLÁSTICO/METAL, TAMANHO MÉDIO, COR AZUL, TIPOENTINTAMENTO PERMANENTE, COMPRIMENTO 12 CM, LARGURA 9 CM MARCA: RADEX ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000251047
16 - MATERIAL DE EXPEDIENTE	10	13,70	137,00	10,00000 UNIDADE COLA, COR BRANCA, APLICAÇÃO PAPEL, CARACTERÍSTICAS ADICIONAIS ATÓXICA, TIPO BASTÃO MARCA: LEONORA ITEM DO PROCESSO: 00002 ITEM DE MATERIAL: 000292447
16 - MATERIAL DE EXPEDIENTE	10	15,40	154,00	10,00000 UNIDADE COLA, COMPOSIÇÃO POLIVINIL ACETATO- PVA, COR BRANCA, APLICAÇÃO PAPEL, TIPO PASTOSA MARCA: FRAMA ITEM DO PROCESSO: 00003 ITEM DE MATERIAL: 000282967
16 - MATERIAL DE EXPEDIENTE	5	7,80	39,00	5,00000 UNIDADE COLA, COMPOSIÇÃO CIANIACRILATO, COR INCOLOR, APLICAÇÃO VIDRO,BORRACHA,PLÁSTICO,PVC,METAL,ACRÍLICO,NÁILON, CARACTERÍSTICAS ADICIONAIS GEL, TIPO INSTANTÂNEA MARCA: TEKBOND ITEM DO PROCESSO: 00004 ITEM DE MATERIAL: 000281629

Fonte:
Tabela de Empenho

Fonte:
Tabela de Itens de Empenho

(a) Tela de Empenho do Portal da Transparência do Governo Federal



(b) Modelo de dados de Empenho

Figura 9 - Tela de Empenho e Modelo de Dados de Empenho

O atributo mais relevante para esse estudo é o denominado “Descrição”, da tabela Item de Empenho, pois é ele que traz a especificação textual que será utilizada na mineração de texto para se identificar o produto que está sendo comprado. Os demais atributos das tabelas de empenho e item de empenho são utilizados somente no processo de validação e nas aplicações da solução proposta.

3.3 Preparação dos Dados e Modelagem

A proposta para a criação de regras de identificação de produtos está dividida em cinco passos, sendo que o primeiro, o pré-processamento, consiste da preparação dos dados, terceira fase do modelo CRISP-DM. Os quatro passos seguintes compõem um processo de descoberta de conhecimento em dados textuais e está relacionado a fase de modelagem do CRISP-DM.

Na Figura 10 é ilustrada a proposta, sendo que além dos 5 passos citados, opcionalmente, dependendo da disponibilidade de pessoal, pode-se inserir um sexto passo, que consiste de uma intervenção humana. O objetivo desse sexto passo é melhorar a forma de representação do conhecimento das regras criadas, assim como fazer algumas adaptações nas regras, de modo que essas possam ser mais adequadas para os propósitos finais da classificação gerada, bem como para melhorar os resultados obtidos.

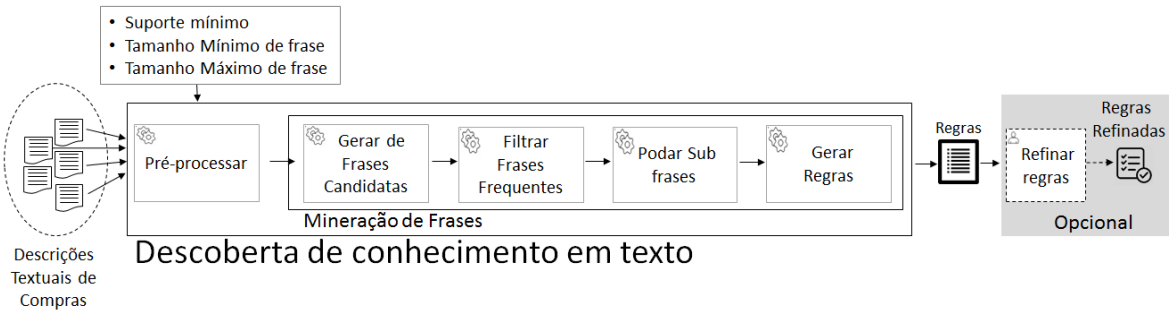


Figura 10 - Processo de geração de regras de identificação de produtos

No contexto desse trabalho, uma frase é definida como uma sequência contigua de tokens. Sendo assim, nessa dissertação, a tarefa de mineração de frases pode ser caracterizada pela agregação e contagem de todas as sequências iguais de tokens contíguos que satisfaçam a um suporte mínimo[14]. Ou seja, a mineração de frases se propõe

[14] Suporte mínimo: valor, passado como parâmetro de entrada do processo, considerado como o limite mínimo de vezes que uma determinada sequência de tokens deve aparecer no conjunto de dados textuais para ser considerada frequente.

a identificar os padrões sequenciais de palavras que mais se repetem em um determinado conjunto de dados textuais.

Dessa forma, as seguintes propriedades, definidas em (EL-KISHKY et al. 2014) e (LIU et al. 2015), devem ser atendidas no processo de mineração de frases:

- Frequência: A qualidade mais importante quando se julga se uma frase retransmite informações relevantes sobre um tópico é a sua frequência de utilização dentro do tópico. Uma frase que não é frequente dentro de um tópico, provavelmente não é importante para esse tópico.
- Completude: Se uma frase longa satisfaz ao critério da frequência, então, as sub frases dessa frase longa também irão satisfazer a este critério. Porém, serão menos informativas do que a frase mais longa, e dessa forma não precisam ser consideradas na mineração, pois a frase mais longa é mais completa.

Devido às características dos dados de portais de transparência, grandes volumes de informações com cargas diárias e incrementais. A solução proposta deve ser capaz de processar quantidades massivas de dados. Para atender a esse requisito, todo o processo foi concebido para ser executado utilizando o Apache Spark (ZAHARIA et al. 2010), um framework para processamento de grandes volumes de dados que roda de forma paralela em cluster de computadores.

3.3.1 Pré-processamento

O pré-processamento é a primeira etapa do processo, e tem o objetivo de preparar o conjunto de dados para as atividades subsequentes. Essa etapa de pré-processamento retira informações que estão presentes no campo de descrição da compra, mas que não fazem parte da especificação textual do produto. O principal objetivo desse procedimento é a eliminação de informações desnecessárias que possam prejudicar a análise das sequências de palavras geradas.

Na Figura 11 é ilustrado o resultado do pré-processamento de uma descrição de compra. Nesse procedimento, algumas informações são identificadas e extraídas, através das técnicas enunciadas em (ETZIONI et al. 2005). Essas técnicas pregam a utilização de templates na atividade de extração de informações de dados textuais. Para isso, cada template é utilizado para extrair um tipo de relação específica entre as palavras que aparecem no texto. Por exemplo, o template “tais como” na frase, “Cidades tais como Rio

de Janeiro e São Paulo" permite-nos concluir que os termos Rio de Janeiro e São Paulo são nomes de cidades.

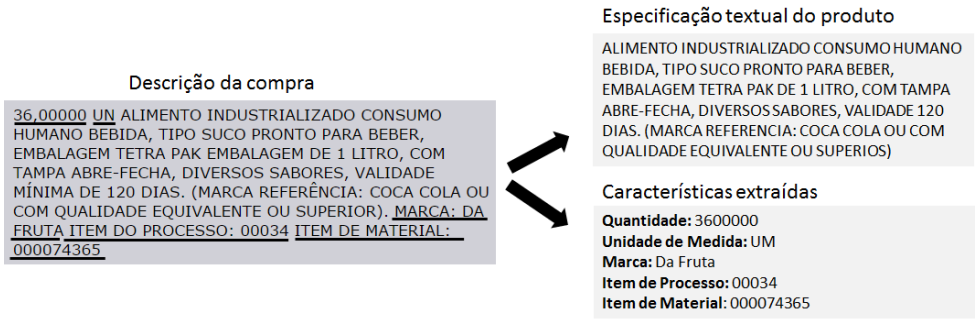


Figura 11 - Resultado do Pré-processamento

Para o caso das descrições textuais das compras, formam identificados templates específicos para esse contexto, e as relações obtidas pela aplicação dos templates ocorrem entre a compra em si e o termo referenciado pelo padrão buscado. Dessa forma a frase: "Marca: Da Fruta" evidencia que a compra que está sendo descrita se refere a um produto cuja marca é Da Fruta. Sendo assim, a utilização de simples templates de identificação permite a extração de uma série de características da compra que, após a devida classificação do produto, ao final de todo o processo, podem agregar maior conhecimento a respeito das informações apresentadas. Logo, além da retirada de termos que possam prejudicar a mineração textual, essa fase também é responsável pela extração de algumas características da compra, que ao final do processo permitem agregar maior conhecimento a respeito dos produtos que estão sendo adquiridos.

Cabe ressaltar que esse procedimento não serve para a identificação do produto que está sendo especificado na descrição da compra, pois, não há nenhum padrão na especificação textual dos produtos, uma vez que, essa especificação é feita pelo usuário, e cada usuário preenche a designação do produto de uma forma diferente.

Outra atividade realizada na etapa do pré-processamento é a filtragem dos dados que não se referem a compra de materiais, visto que, existem outros tipos de gastos que não dizem respeito à compra de produtos, como por exemplo contratação de serviços e pagamento de pessoal. Essa pesquisa não considera os contratos de prestação de serviços pelo fato desses apresentarem grande variabilidade de características, o que faz com que cada contratação seja única.

Durante o pré-processamento, também é realizado um tratamento no texto de forma que todas as letras presentes nas descrições de compras sejam passadas para o formato de letra minúscula e que todos os sinais de acentuação sejam retirados.

3.3.2 Geração de Frases Candidatas

Apesar do método proposto apresentar um enfoque estatístico, com o intuito de se diminuir o conjunto de possíveis combinações de palavras, assim como para manter a expressividade das frases geradas, algumas considerações semânticas foram feitas:

- Uma frase só pode ser formada se ela estiver contida dentro de uma determinada sentença. Nessa pesquisa, considera-se sentença como sendo uma sequência de palavras delimitada por sinais de pontuação que determinam o final de um período (ponto final, ponto de exclamação ou ponto de interrogação).
- Se uma determinada palavra W está localizada na posição n de uma sequência de palavras de uma sentença, para que essa palavra W faça parte de uma frase, é necessário que todas as demais palavras localizadas nas $(n - 1)$ posições anteriores da sequência, também façam parte dessa frase. Essa restrição foi formulada para garantir maior grau de expressividade para as frases formadas, visto que, na língua portuguesa o significado de uma frase vai se completando da esquerda para a direita.

Algoritmo 1: Geração de Frases Candidatas

Entrada:
Conjunto de Especificações de Produtos *E*,
tamanho mínimo da frase *min* e
tamanho máximo da frase *max*

Saída:
Vetor com frases construídas

```
1.  Início
2.      frases=[]
3.      Para cada especificação e em E Faça:
4.          Sentenças= SeparaSentenças(e)
5.          Para cada sentença em sentenças Faça:
6.              Para m entre (min,max) Faça:
7.                  SE (tamanho(sentença)>=m)
8.                      frases.insere(sentença[0:m])
9.                  Fim
10.             Fim
11.         Fim
12.     Fim
13.     retorna frases
14. Fim
```

Figura 12 - Algoritmo de geração de Frases Candidatas

Na Figura 12 é apresentado o algoritmo de geração das frases candidatas. Esse algoritmo recebe como entrada um conjunto de especificações textuais de produtos e como parâmetro um tamanho mínimo e outro máximo para as frases a serem geradas, sendo que, o tamanho de uma frase é medido pelo número de palavras que compõem essa frase. A saída do algoritmo proposto será o conjunto de todas as frases geradas.

Cabe ressaltar que os primeiros trabalhos que abordaram a questão de mineração de frases (REN et al., 2015), (LIU et al., 2015) e (EL-KISHKY et al., 2014) já propunham a utilização de algoritmos de geração de frases, porém, nesse trabalho foi proposto um algoritmo específico para os propósitos dessa dissertação.

Para exemplificar o funcionamento do algoritmo de geração de frases candidatas, na Figura 13 é apresentada a especificação textual de uma determinada compra, que seria um elemento pertencente ao conjunto de especificações de produtos E, que funciona como entrada para o algoritmo, e as frases resultantes da aplicação desse algoritmo para o caso de se utilizar os parâmetros de tamanhos mínimo e máximo de frases como sendo 5 e 10, respectivamente.

Figura 13 – Exemplo do processo de Geração de Frases

3.3.3 Filtragem de Frases Frequentes

Após a geração das frases candidatas, o passo seguinte é a agregação das frases iguais, a fim de se contar o número de ocorrências de cada uma das frases geradas. Sendo assim, para cada frase gerada pelo algoritmo de geração de frases candidatas é feita uma verificação e contagem de todas as frases coincidentes. Dessa forma, cada frase estará associada a um número de ocorrências, e aquelas frases que tiverem esse número de ocorrências superior a um suporte mínimo, passado como parâmetro, prosseguem no processamento, enquanto que, as frases cujo número de ocorrências for inferior a esse suporte são desconsideradas.

Essa etapa tem o objetivo de atender ao critério da frequência, e o seu funcionamento é ilustrado pela Figura 14, que apresenta um conjunto de frases agregadas, com os respectivos números de ocorrências e o resultado dessa etapa após a aplicação de um suporte mínimo de 30.

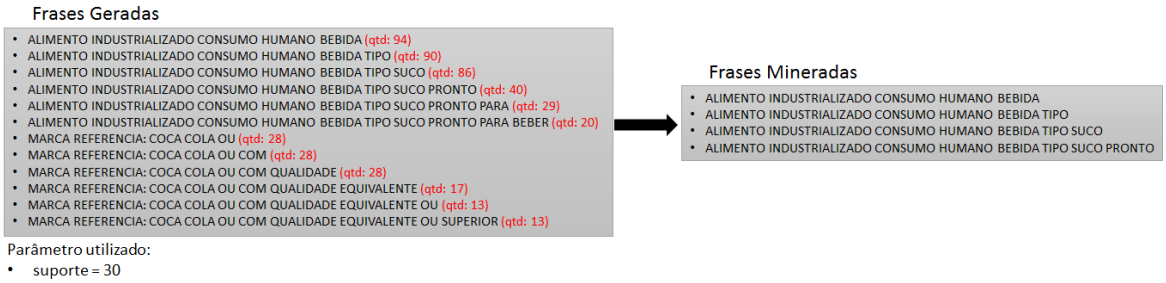


Figura 14 – Exemplo do processo de filtragem de frases

3.3.4 Poda de Sub frases

EL-KISHKY et al. (2014) definem duas propriedades na mineração de frases:

- Lema do fechamento para baixo: Se uma frase G não é frequente, então as super frases de G (frases que contêm G) também não serão.
- Antimonotonicidade dos dados: Se um documento não contém frases frequentes de comprimento n, o documento não contém frases frequentes de comprimento maior que n.

A aplicação dessas propriedades ao conjunto de frases resultante do passo anterior serve para reduzir a quantidade das frases decorrentes do processo de mineração. Sendo assim, se uma frase G, formada pela sequência de palavras $w_1\, w_2 \ldots w_n$ atende ao requisito do suporte mínimo, então, todas as suas sub frases $G' = w_1\, w_2 \ldots w_k$, com $k < n$, também atenderão a esse suporte, porém, elas não precisarão ser analisadas, uma vez que as frases maiores (em que elas estão contidas) já contemplam ao requisito necessário. Logo, é executado uma filtragem aplicando essa propriedade de forma a reduzir o número de frases mineradas.

Algoritmo 2: Poda Sub Frases

Entrada:
Vetor H, com as frases que satisfazem o suporte mínimo
Saída:
Vetor com SuperFrases

```
1.  Start
2.      frasesDeQualidade=[]
3.      Ordena(H) // ordena em ordem decrescente de tamanho
4.      Para cada frase h em H Faça:
5.              SuperFrase=Verdadeiro
6.              Para cada frase sp em frasesDeQualidade Faça:
7.                      Se h in sp Então:
8.                              SuperFrase=Falso
9.                              continua
10.                     End
11.             End
12.             If SuperFrase=Verdadeiros Then
13.                     frasesDeQualidade.insere(h)
14.             End
15.     End
16.     retorna frasesDeQualidade
17. End
```

Figura 15 - Algoritmos de Poda de Sub Frases

O algoritmo apresentado na Figura 15 mostra o processo de poda das sub frases. Esse algoritmo recebe como entrada todas as frases geradas que atenderam ao critério do suporte mínimo, e oferece como saída apenas as super frases (ou seja, frases contidas em outras frases maiores que também atendam ao requisito do suporte mínimo são desconsideradas). Essa etapa tem o objetivo de atender ao critério da completude (definido no início da seção 3.3).

Para exemplificar o funcionamento do algoritmo de poda de sub frases, a Figura 16 apresenta, no lado esquerdo, um conjunto de frases resultantes do processo de filtragem de frases, ou seja, frases que tenham atendido ao suporte mínimo passado como parâmetro. Já o lado direito da Figura 16 representa a saída do algoritmo, considerando-se como entrada as frases apresentadas do lado esquerdo da figura.

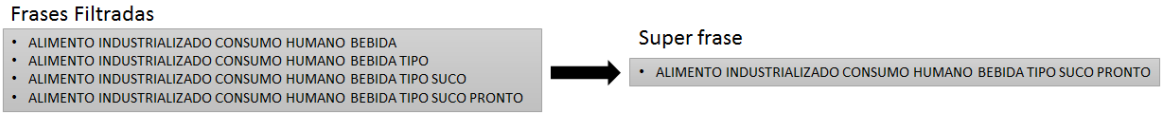


Figura 16 - Exemplo do processo de poda de sub frases

3.3.5 Geração de Regras

A última etapa do processo é a geração das regras de identificação. As regras são do tipo: “SE antecedente ENTÃO consequente”, sendo o antecedente a premissa, definida por uma determinada frase, e o consequente o produto a ser identificado a partir da premissa.

Logo, cada frase resultante do processo de poda de sub frases dá origem a uma regra distinta. Dessa forma, assume-se que todas as compras que se enquadrarem em uma determinada regra de identificação (ou seja, todas as compras cuja especificação tenha alguma frase que coincida com uma determinada frase considerada como antecedente, que resultou do processo de mineração de frases) se referem a um mesmo tipo de produto.

Os consequentes das regras são as sugestões para a designação dos produtos identificados pela análise dos antecedentes. No entanto, como as frases mineradas, que dão origem aos antecedentes das regras, tendem a ter um alto grau de expressividade, optou-se para que essas sejam consideradas como consequentes das regras.

Porém, adicionalmente, apresenta-se também uma sugestão de consequente alternativo, ou seja, rótulos alternativos para os produtos identificados. Essa consideração se faz pelo fato de que regras cujos antecedentes sejam semelhantes tendem a identificar o mesmo tipo de produto.

Para isso, utilizou-se um critério de cálculo de similaridade entre os antecedentes das regras geradas, a fim de se encontrar as regras que tenham maiores probabilidades de levar a identificação dos mesmos tipos de produtos.

Para o cálculo de similaridade entre as frases, utilizou-se o algoritmo "correspondência de padrão gestáltico", definida por (RATCLIFF; METZENER, 1988). Esse algoritmo considera que a similaridade entre duas strings é dada pela expressão $(2M)/T$, onde T é o número total de caracteres de ambas as strings, e M representa o número de caracteres correspondentes em ambas as strings (ou seja, M é o número de caracteres iguais e na mesma posição nas duas strings comparadas).

Sendo assim, todas as frases mineradas são comparadas duas a duas. Sempre que a similaridade entre duas frases é maior do que um determinado parâmetro (por exemplo 70 %) considerava-se essas frases como similares e atribuiu-se um mesmo rótulo alternativo para ambas.

3.3.6 Refinamento de Regras

Conforme dito anteriormente, a solução proposta ainda prevê uma sexta etapa no processo. Porém, a etapa de refinamento de regras é opcional e depende da disponibilidade de pessoal para a realização dessa tarefa, visto que, essa última etapa carece de uma interação humana. Essa etapa tem o objetivo de melhorar a forma de representação do conhecimento expressa pelos consequentes das regras geradas, bem como possibilitar a agregação, ou eliminação de regras de acordo com o grau de especificidade, ou generalidade, que se deseja dar no processo de identificação das compras

Durante essa etapa, os especialistas analisam as regras geradas e fazem a seleção e validação dessas regras, bem como a escolha de consequentes semanticamente mais apropriados para cada regra. Logo, o esforço dos especialistas nessa fase se resume em fazer a seleção e validação dos antecedentes e reformular os consequentes de cada regra.

- Seleção/validação dos antecedentes: Esse procedimento tem duas finalidades, a primeira se dá porque, apesar das frases tenderem a ter um alto grau de expressividade, pois, elas atingiram uma frequência alta de ocorrência, em algumas situações elas não transmitem informações capazes de discriminar um determinado produto. Outra razão que justifica o benefício da interação humana é a definição do grau de especificidade que se deseja dar a um determinado produto. Por exemplo, um produto pode ser identificado como suco de laranja ou simplesmente como suco, dependendo da análise que se deseja fazer, e a seleção dos antecedentes das regras de identificação tem importante papel nesse processo.

- Reformulação de consequentes: Um papel relevante, executado por especialistas, é a interpretação dos antecedentes das regras, a fim de definir consequentes semanticamente mais apropriados. Outra vantagem dessa atividade é a possibilidade de se definir consequentes iguais para regras diferentes, mas que tenham o mesmo conteúdo informacional. Por exemplo, um especialista pode definir um mesmo rótulo para os antecedentes “dipirona, solução oral 500 mg/ml” e “novalgina gotas 500 mg/ml”, associação essa que seria difícil de se fazer de forma automatizada. Durante essa atividade também é feita a avaliação dos rótulos alternativos sugeridos pelo processo de geração de regras, decidindo-se se esses devem ou não ser considerados como consequentes das regras.

3.3.7 Aplicação das Regras

O objetivo final de todo o processo é a identificação dos produtos que estão sendo especificados de forma textual nas descrições de compras. Sendo assim, uma vez que se tem a definição das regras de identificação, o passo seguinte é aplicá-las ao conjunto de dados que se deseja classificar.

O procedimento indicado é que se use uma massa de dados menor para se fazer a geração das regras de identificação, visto que, essa tarefa requer uma grande carga de processamento. Uma vez que se tem essas regras geradas, pode-se aplicá-las a massas de dados maiores ou até mesmo aplicá-las em tempo real, de forma concomitante com a carga diária dos dados dos portais de transparência.

A aplicação das regras está dividida em duas etapas: o pré-processamento e a aplicação das regras propriamente ditas, conforme apresentado na Figura 17. Esse procedimento recebe como entrada um conjunto de descrições textuais de compras, da mesma forma que o processo de geração de regras e como parâmetro um conjunto de regras de identificação e a saída esperada são as compras devidamente identificadas.

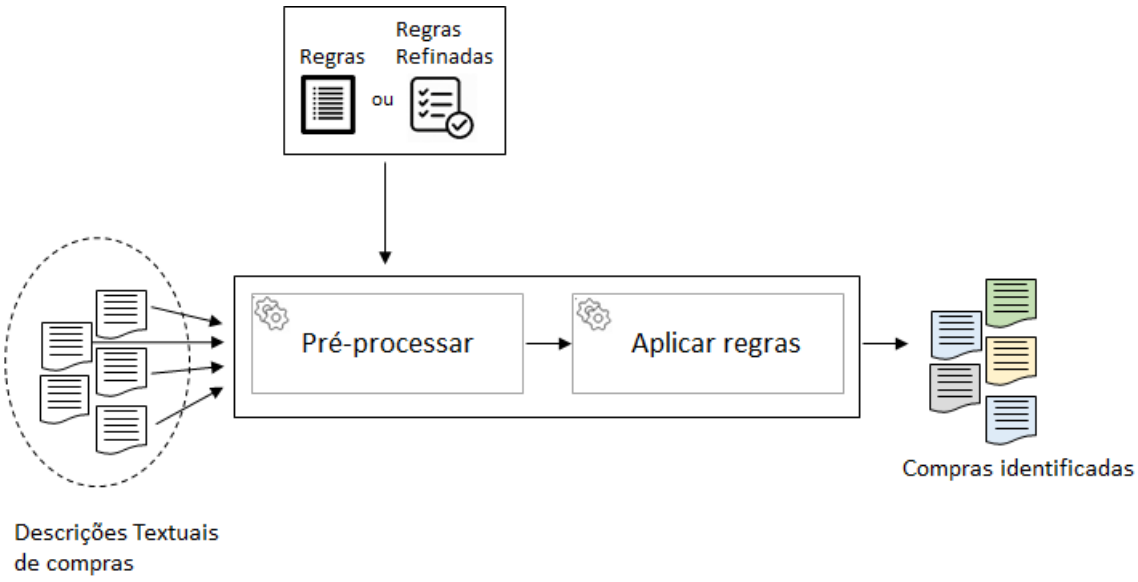


Figura 17 – Aplicação das regras

A primeira etapa do processo de aplicação das regras consiste de um pré-processamento, que é idêntico ao pré-processamento realizado durante a geração das regras, descrito na subseção 3.3.1. Logo, a exemplo do que foi descrito anteriormente,

essa etapa extrai, das descrições de compras, algumas informações que auxiliam na caracterização dos produtos, que será feita no passo seguinte.

Durante a etapa de aplicação das regras propriamente ditas, todas as regras são testadas para cada uma das compras descritas, de forma que, todas aquelas que atenderem a clausula especificada no antecedente da regra, sejam consideradas como pertencente ao grupo de produtos designado como consequente da respectiva regra. Cabe ressaltar que uma mesma descrição de compra pode atender a mais de uma regra ao mesmo tempo, sendo que, nessa situação ambos os consequentes serão atribuídos à descrição de compra. Ou seja, essa mesma descrição de compra pode estar relacionada a mais de um produto.

Esse capítulo apresentou a proposta de solução para o problema que está sendo tratado nessa dissertação. Uma validação dessa proposta é feita no Capítulo 4.

Capítulo 4 – Avaliação

O quinto passo da metodologia CRISP-DM é a validação do modelo. Esse capítulo apresenta um estudo que tem o objetivo de avaliar os resultados obtidos pela aplicação da solução proposta em um conjunto de descrições de compras que são apresentadas no Portal da Transparência do Governo Federal.

4.1 Infraestrutura

O estudo foi desenvolvido no ambiente de computação disponibilizado pela empresa Amazon Web Services (AWS) (CLOUD, 2011). O AWS é uma plataforma de infraestrutura de Tecnologia da Informação disponível na nuvem. Nessa plataforma são oferecidas opções de armazenamento, rede, banco de dados e processamento que são entregues como serviços. O AWS oferece uma ampla seleção de tipos de instâncias otimizadas para se adequarem a diferentes usos. Os tipos de instâncias consistem em várias combinações de CPU, memória, armazenamento e capacidade de rede e oferecem flexibilidade de escolha na composição de acordo com as necessidades. Esses serviços são disponibilizados sob demanda, com preço definido conforme o uso. Optou-se por essa estrutura pelo fato de tal solução apresentar todos os recursos de software e hardware necessários para a execução de aplicações que utilizam Apache Spark (ZAHARIA et al. 2010).

Os experimentos apresentados nessa dissertação foram desenvolvidos com a utilização da instância m1.large, que é composta por um cluster com 2 máquinas Linux com processador de 64 bits e 7,5 GiB de memória. Esse cluster possuía o Apache Spark instalado. Já os arquivos utilizados como fonte de dados de entrada para o processamento e os arquivos gerado após o processo de identificação das compras foram disponibilizados

no S3, um serviço de armazenamento na nuvem, também oferecido pela Amazon Web Service.

4.2 Projeto de Avaliação

O projeto de avaliação está dividido em duas partes: a primeira consiste da avaliação das regras geradas, enquanto que a segunda verifica a qualidade dos resultados obtidos no processo de identificação de compras propriamente dito. Em ambos os casos a validação é feita após a aplicação das regras a um conjunto de descrições textuais de compras.

4.2.1 Avaliação das Regras

A avaliação das regras geradas é feita através da aplicação de um método de clusterização. Dessa forma, para cada uma das regras de identificação, são utilizadas as compras que se enquadraram nessas regras. Em tal procedimento, utiliza-se os atributos da classificação da natureza de despesa detalhada[15], como variáveis para o cálculo dos clusters formados pelas compras que se enquadram nas regras.

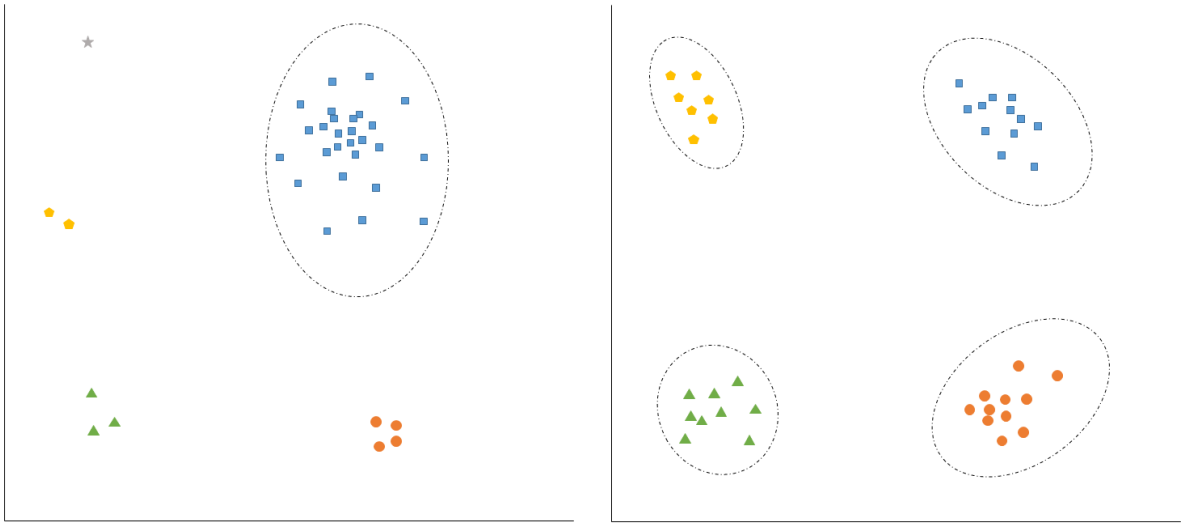
Partindo-se da premissa de que compras que se referem a um mesmo produto possuem a mesma classificação de natureza de despesa, considerou-se que em uma situação ideal, o processo de clusterização de uma regra perfeita iria gerar um único cluster com todas as compras identificadas por essa regra concentradas nesse cluster único, pois as compras identificadas de forma correta iriam corresponder a compras de um mesmo produto, e consequentemente possuiriam os mesmos atributos de natureza de despesa detalhada.

No entanto, numa situação real, outras variáveis externas interferem na avaliação do processo como um todo, como por exemplo a inserção de dados de natureza de despesa errada por parte dos usuários responsáveis por tal atividade. Dessa forma, formulou-se as seguintes considerações gerais para a análise dos clusters formados:

[15] A Natureza da Despesa Detalhada é uma classificação utilizada pela Contabilidade Pública, e é composta por um código formado por 8 dígitos numéricos, divididos em 5 grupos, com a seguinte configuração: X.Y.ZZ.MM.NN. Nessa codificação, o 1º dígito indica o código da categoria econômica, o 2º dígito indica o Grupo de natureza de despesa, os 3º e 4º dígitos indicam a modalidade de aplicação, os 5º e 6º dígitos indicam o elemento de despesa e os 7º e 8º dígitos indicam o sub elemento de despesa

- Caso haja vários clusters com quantidades equivalentes de ocorrências de compras, provavelmente a regra não é boa, pois ela está pegando muitos produtos com classificação de natureza de despesa diferentes, e provavelmente está identificando produtos diferentes como sendo iguais.
- Caso haja poucos clusters, sendo que apenas um cluster concentra a grande maioria das compras, e os demais possuam poucas ocorrências, provavelmente essa regra é boa, e as ocorrências dispersas (outros clusters) supostamente tenham sido fruto de classificações erradas da natureza de despesa, por parte da pessoa responsável por inserir essa informação no sistema.

A seguir é apresentado um exemplo do resultado do processo de clusterização das compras identificadas por uma regra boa e por uma regra ruim, nas figuras 18.a e 18.b, respectivamente. Deve ser observado que essas figuras representam situações hipotéticas, e têm apenas o objetivo de ilustrar as configurações resultantes do processo de clusterização que caracterizam uma regra como sendo boa ou ruim. Os parâmetros utilizados para determinar a qualidade das regras são especificados adiante, ainda nesse capítulo.



(a) exemplo de clusters para uma regra boa

(b) exemplo de clusters para uma regra ruim

Figura 18 - Exemplos de configurações de clusters

Conforme pode ser observado na Figura 18, quanto mais homogênea for a configuração do cluster obtido pelo processo de clusterização das compras enquadradas em uma determinada regra, melhor será a qualidade dessa regra. Por outro lado, quanto mais disperso for a configuração desses clusters, pior será a qualidade dessa regra, pois

provavelmente essa regra está identificando produtos diferentes, o que não é o comportamento esperado para as regras.

4.2.2 Avaliação da Identificação

Dada a inexistência de um conjunto de dados previamente rotulado, associada a inviabilidade de se analisar individualmente cada um dos casos, a avaliação dos resultados obtidos no processo de identificação de compras é feita de forma qualitativa. No entanto, o processo de escolha da amostra a ser analisada qualitativamente é feito por procedimentos quantitativos.

Então, a avaliação utiliza uma abordagem empírica e emprega técnicas quantitativas e qualitativas. As técnicas quantitativas são empregadas para a seleção de um conjunto amostral com maior probabilidade de ter sido classificado erroneamente. Já as técnicas qualitativas são utilizadas para analisar a amostra de dados selecionada.

A abordagem também combina características descritivas e explanatórias. A parte descritiva tem o objetivo de descrever as características dos dados relativos a cada um dos tipos de produtos identificados, a fim de possibilitar a seleção dos casos em que as características do produto diferem dos demais da mesma espécie. Já a parte explanatória tem o objetivo de analisar de forma mais detalhada esses casos selecionados, a fim de avaliar se a identificação do produto foi feita corretamente ou não.

Logo, as compras são grupadas por produtos, e para cada um dos grupos de produtos busca-se por outliers. Sendo assim, partindo-se da premissa que produtos iguais tendem a apresentar características similares, pode-se inferir que aqueles produtos cujos atributos sejam discrepantes em relação aos demais têm maior probabilidade de terem sido identificados de forma errônea. Dessa forma, pode-se selecionar uma amostra com casos mais propensos a terem sido classificados de forma errônea, otimizando assim a atividade dos especialistas responsáveis pela análise qualitativa desses resultados.

4.2.3 Estudo de Caso

A proposta de geração de regras de identificação, descrita no capítulo 3, utiliza, além dos dados de entrada (um conjunto de descrições textuais de compras), três tipos de parâmetros: tamanho mínimo de frase, tamanho máximo de frases e suporte. Sendo assim, utilizou-se como fonte de dados de entrada, para a geração das regras de identificação, as

descrições de compras dos itens de empenho, apresentados no Portal da Transparência do Governo Federal, referentes ao mês de janeiro do ano de 2015, e executou-se um estudo de caso composto por uma série de experimentos, sendo que, para cada um dos experimentos utilizou-se uma combinação diferente dos parâmetros requeridos pelo processo de geração de regras, ou seja, variou-se os valores de tamanhos máximo e mínimo de frase e o suporte requerido.

Para a verificação dos resultados obtidos pelo processo de identificação das descrições textuais de compras, utilizou-se os dados das descrições de compra dos itens de empenho, apresentados no Portal da Transparência do Governo Federal, referentes aos meses de fevereiro a dezembro do ano de 2015. Uma vez realizados os experimentos, com as diferentes configurações, analisou-se os resultados obtidos a fim de se verificar a consistência das regras geradas com cada uma das configurações de parâmetros.

4.3 Base de Dados

A solução foi avaliada com os dados referentes ao período de um ano de compras (dados do ano de 2015) apresentadas no Portal da Transparência do Governo Federal.

Dessa forma, utilizou-se os dados referentes ao mês de janeiro para se criar as regras de identificação dos produtos e os dados referentes aos demais meses para se aplicar as regras, ou seja, para se realizar a identificação dos produtos especificados nas descrições textuais das compras.

Para a avaliação, foi utilizada uma amostra das descrições de compras apresentadas no Portal da Transparência do Governo Federal. A base de dados utilizada, cuja estrutura foi apresentada na seção 3.3, era composta de 3.326.111 registos de empenho e por 5.710.039 registros de itens de empenho, dos quais 2.465.610 se referiam a compra de materiais. Os demais itens de empenho se referiam a outros tipos de despesas, como por exemplo, pagamento de pessoal e contratação de serviço. No entanto, apesar de muitos dos itens de empenho não estarem relacionados com a compra de materiais, todos eles precisam ser processados, pois não se sabe previamente a que se refere um determinado item (pré-processamento).

Os itens de empenho referentes ao mês de janeiro totalizavam 212.726 registros, enquanto que, os itens de empenhos referentes ao demais meses correspondiam a 5.497.313 registros.

4.4 Experimentos

Durante a execução do estudo de caso foram realizados seis experimentos, conforme a configuração apresentada na Tabela 2.

Tabela 2 - Parâmetros dos experimentos realizados

Identificação	Tam Mínimo Frase	Tam Máximo Frase	Suporte
Experimento 1	10	15	10
Experimento 2	10	15	15
Experimento 3	10	15	30
Experimento 4	6	9	15
Experimento 5	6	9	30
Experimento 6	6	9	50

A intenção da variação dos parâmetros utilizados é possibilitar uma análise do comportamento da solução proposta com diferentes configurações, a fim de se identificar aquela mais apropriada para a situação apresentada. Na Tabela 3 é mostrado o tempo gasto para a geração das regras (com os dados referentes ao mês de janeiro) e para a aplicação das regras (dados referentes ao período de fevereiro a dezembro), bem como o número de regras geradas em cada um dos experimentos.

Tabela 3 - Tempo de processamento e quantidade de regras geradas por experimento

Identificação	Tempo Criação das Regras	Tempo Aplicação das regras	Quantidade de regras
Experimento 1	0:57	8:49	119
Experimento 2	0:57	7:57	85
Experimento 3	1:01	7:00	31
Experimento 4	0:57	9:02	239
Experimento 5	0:58	7:28	72
Experimento 6	0:56	6:58	25

Pela análise da Tabela 3, pode-se verificar que a variação dos parâmetros de entrada não acarretou grandes diferenças nos tempos de geração das regras, porém, esses parâmetros têm uma influência direta na quantidade de regras geradas.

Analisando-se os experimentos 1, 2 e 3 e os experimentos 4, 5 e 6 de forma separada, pois possuem os mesmos tamanhos mínimo e máximo de frases, percebe-se que, para os mesmos parâmetros de tamanhos máximo e mínimo de frases, o valor do suporte está inversamente relacionado com a quantidade de regras geradas, ou seja, quanto menor o suporte, maior é o número de regras geradas.

Porém, quando se analisa os experimentos com o mesmo suporte, mas com tamanhos de frases diferentes, como é o caso dos experimentos 2 e 4 e dos experimentos 3 e 5, percebe-se que o número de regras é maior quando os tamanhos das frases são menores.

Dessa forma, pode-se concluir que suportes menores ou tamanhos de frases menores trazem como efeito um aumento no número de regras geradas, enquanto que suportes e tamanhos de frases maiores têm um efeito contrário.

Com relação ao tempo de execução da aplicação das regras, percebeu-se que quanto maior o número de regras geradas, maior o tempo gastos na execução, no entanto, esse resultado já era esperado, pois, apesar de todas as descrições de compra terem de ser analisadas, independentemente do número de regras, quanto maior o número de regras geradas, maior é o número de testes a serem feitos.

4.4.1 Análise das Regras Geradas

Conforme apresentado na subseção 4.2.1, as regras geradas são avaliadas pela aplicação de métodos de clusterização. Dessa forma, utilizou-se o algoritmo *"Density Based Spatial Clustering of Application with Noise"* – DBSCAN (ESTER et al ,1996) - para executar os testes.

O DBSCAN (ESTER et al ,1996) é um método de clusterização não paramétrico baseado em densidades . A escolha desse método se deu pelo fato dele não exigir que seja informado previamente o número de clusters a serem encontrados.

Dessa forma, para cada uma das regras de identificação, aplicou-se o método de clusterização DBSCAN em todas as compras identificadas por essa regra, utilizando-se os atributos de natureza de despesa detalhada como variáveis a serem consideradas nesse processo de clusterização. Ou seja, compras com a mesma natureza de despesa seriam enquadradas no mesmo cluster, enquanto que compras com natureza de despesa detalhada diferentes cairiam em clusters diferentes

Para auxiliar na geração dos clusters, foi utilizada a ferramenta RapidMiner , sendo que, todos os 6 experimentos executados foram considerados, porém, por uma questão de otimização dos trabalhos, apenas 25 regras de cada um dos experimentos foram analisadas.

Dessa forma, classificou-se as regras em três categorias diferentes, de acordo com a configuração dos clusters gerados. Sendo assim, uma regra pode ser classificada como sendo boa, regular ou ruim.

- Regra Boa

Para uma regra ser considerada como boa, é necessário que o processo de clusterização gere um cluster com mais de 90% das compras classificadas nessa regra.

Um exemplo de uma regra considerada como boa é a regra R_22 do experimento 6, cujo antecedente é formado pela seguinte sequência de palavras: "caneta esferografica, material plastico, quantidade cargas 1 un" e que concentra 95,38% das ocorrências em um mesmo cluster.

- Regra Regular

Para que uma regra seja considerada como regular, é necessário que o processo de clusterização gere 2 clusters cuja soma das percentagens das ocorrências supere os 90%. Essa consideração é feita para se atender aos casos em que um mesmo produto pode ser classificado em duas classificações de natureza de despesa detalhada diferentes. Por exemplo, o produto "Reagente para diagnóstico clinico" que pode ser classificado tanto na natureza de despesa detalhada 3.3.90.30.09 quanto na 3.3.90.30.36, que correspondem a material farmacológico e material hospitalar, respectivamente. Para o experimento 6, uma regra classificada como regular foi a regra cujo antecedente era formado pela seguinte sequência de palavras: "reagente para diagnostico clinico, tipo conjunto completo para" que possui os dois clusters mais populosos com 54,38% e 42,50% das ocorrências.

- Regra Ruim

Sempre que uma regra não se enquadrar em uma das duas situações anteriores (regra boa ou regular), ela é considerada como ruim. Um exemplo de uma regra considerada como ruim é a regra R_1 do experimento 6, cujo o antecedente é formado pela sequência de palavras ", numero de referencia quimica cas[16]", pois essa sequência

[16] O número de referência química CAS é um número de registro que todo os produtos químicos têm, esse número único de registro consta no banco de dados do Chemical Abstracts Service, uma divisão da Chemical American Society, e serve para identificar um determinado produto químico. Logo, todos os produtos químicos adquiridos pela Administração Pública Federal irão possuir a sequência de palavras "numero de referencia quimica cas", no entanto, tal sequência de palavras não é suficiente para identificar um determinado produto químico.

de palavras permite-nos inferir que o produto que está sendo especificado possui algum tipo de elemento químico em sua composição, porém, diferentes tipos de produtos podem ser enquadrados nessa regra. Tal regra gerou 40 clusters distintos e apresentou a seguinte distribuição de frequência de ocorrência para os clusters mais populosos: 57,04%, 14,59%, 11,44%, 4,99% e 3,79%.

Conforme apresentado na Tabela 3, os experimentos geraram diferentes quantidades de regras, sendo que o experimento 6 foi o que gerou menos regras, totalizando 25 regras distintas. Dessa forma, para se comparar a qualidade das regras geradas em cada um dos experimentos, considerou-se todas as 25 regras geradas para o experimento 6, e para os demais experimentos considerou-se as 25 regras que tenham tido mais compras classificadas

Na Tabela 4 são apresentados os clusters obtidos para as compras classificadas de acordo com as regras geradas na execução do experimento 6. Essa tabela apresenta uma linha para cada uma das 25 regras geradas, e colunas com a identificação das regras, número de clusters gerados por regra, porcentagem de itens de compra classificados nos 5 clusters mais populosos de cada regra e classificação da regra de acordo com os critérios citados acima (boa, regular ou ruim).

Tabela 4 - Clusters das regras gerados no experimento 6

Id Regra	Qtd de Clustres	Porcentagem de itens por clusters mais populosos (%)					Situação da Regra
		1º mais populoso	2º mais populoso	3º mais populoso	4º mais populoso	5º mais populoso	
R_1	40	57,04	14,59	11,44	4,99	3,79	Ruim
R_2	13	96,20	0,85	0,75	0,40	0,34	Boa
R_3	5	95,46	1,76	1,14	1,05	0,57	Boa
R_4	13	94,15	1,55	1,51	1,11	047	Boa
R_5	5	76,16	16,38	4,66	2,28	0,50	Regular
R_6	1	100,00	-	-	-	-	Boa
R_7	14	69,07	10,58	10,34	1,86	1,57	Ruim
R_8	12	79,45	6,37	3,54	3,14	2,08	Ruim
R_9	12	52,82	43,45	1,32	1,07	0,51	Regular
R_10	13	93,14	3,58	0,66	0,61	0,46	Boa
R_11	11	96,89	0,84	0,51	0,36	0,30	Boa
R_12	6	94,96	2,23	1,21	0,89	0,38	Boa
R_13	3	76,75	12,43	10,81	-	-	Ruim
R_14	7	93,41	2,61	1,62	1,06	0,67	Boa
R_15	14	97,41	0,76	0,27	0,26	0,22	Boa
R_16	5	94,28	1,87	1,58	1,48	0,78	Boa
R_17	17	53,17	14,36	13,45	8,44	2,66	Ruim
R_18	11	54,38	42,50	1,50	0,77	0,31	Regular
R_19	1	100,00	-	-	-	-	Boa
R_20	8	51,23	43,48	1,50	1,44	1,28	Regular
R_21	7	54,59	42,13	1,19	0,98	0,64	Regular
R_22	13	95,38	0,80	0,57	0,46	0,42	Boa

R_23	8	95,44	0,88	0,74	0,68	0,68	Boa
R_24	5	93,49	3,51	1,36	1,20	0,43	Boa
R_25	8	95,23	2,01	0,86	0,67	0,57	Boa

Conforme pode ser observado na Tabela 4, o experimento 6 gerou 15 regras boas, 5 regras regulares e 5 regras ruins

Já na Tabela 5 são apresentados os resultados obtidos para as 25 regras com maior número de compras classificadas de acordo com as regras geradas no experimento 5. Conforme pode ser observado na Tabela 5, dentre as 25 regras analisadas, 16 foram classificadas como boas, 6 como regulares e 3 como ruins.

Tabela 5 - Clusters das regras gerados no experimento 5

Id Regra	Qtd de Clustres	Porcentagem de itens por clusters mais populosos (%)					Situação da Regra
		1º mais populoso	2º mais populoso	3º mais populoso	4º mais populoso	5º mais populoso	
R_49	11	54,38	42,50	01,20	0,76	0,31	Regular
R_23	12	52,82	43,45	1,32	1,07	0,51	Regular
R_40	14	97,40	0,76	0,27	0,26	0,22	Boa
R_2	13	96,20	0,86	0,75	0,40	0,33	Boa
R_8	13	94,15	1,55	1,52	1,11	0,47	Boa
R_31	8	92,83	5,77	0,63	0,22	0,16	Boa
R_70	4	99,09	0,45	0,33	0,12	-	Boa
R_10	10	94,30	1,54	1,36	0,64	0,48	Boa
R_18	10	55,55	42,29	1,50	1,40	1,25	Regular
R_4	11	63,20	2,19	1,22	1,19	0,53	Ruim
R_3	12	94,26	1,35	1,27	0,88	0,62	Boa
R_14	8	96,70	1,20	0,62	0,38	0,38	Boa
R_17	6	97,16	0,83	0,69	0,53	0,47	Boa
R_47	17	53,17	14,36	13,45	8,44	2,66	Ruim
R_12	10	93,44	1,77	1,43	1,22	0,61	Boa
R_42	10	92,69	4,48	0,86	0,58	0,34	Boa
R_48	10	97,04	0,63	0,44	0,41	0,35	Boa
R_72	8	95,23	2,02	0,86	0,67	0,58	Boa
R_26	11	87,28	5,31	4,04	1,08	0,65	Regular
R_60	13	95,38	0,80	0,57	0,46	0,46	Boa
R_62	15	94,03	0,92	0,64	0,56	0,52	Boa
R_6	10	77,94	15,65	1,52	1,20	1,20	Regular
R_51	11	92,33	4,50	0,65	0,53	0,36	Boa
R_57	7	54,59	42,13	1,19	0,98	0,64	Regular
R_22	12	79,45	6,37	3,54	3,14	2,08	Ruim

O experimento 4 obteve 21 regras boas, 3 regras regulares e 1 regra ruim, e seus resultados podem ser visualizados na Tabela 6.

Tabela 6 - Clusters das regras gerados no experimento 4

Id Regra	Qtd de Clustres	Porcentagem de itens por clusters mais populosos (%)					Situação da Regra
		1º mais populoso	2º mais populoso	3º mais populoso	4º mais populoso	5º mais populoso	
R_166	11	54,38	42,50	1,50	0,77	0,31	Regular
R_231	8	99,55	3,33	0,47	0,21	0,18	Boa
R_67	12	52,82	43,45	1,32	1,07	0,51	Regular
R_118	14	97,14	0,77	0,27	0,26	0,23	Boa
R_37	13	96,20	0,85	0,76	0,40	0,34	Boa
R_14	13	94,15	1,55	1,52	1,11	0,47	Boa
R_234	4	99,09	0,46	0,33	0,12	-	Boa
R_20	10	94,30	1,54	1,36	0,63	0,48	Boa
R_7	11	93,20	2,19	1,22	1,19	0,54	Boa
R_4	12	94,26	1,35	1,27	0,88	0,62	Boa
R_109	8	93,26	3,10	1,20	0,61	0,53	Boa
R_120	9	94,77	2,39	0,78	0,54	0,48	Boa
R_27	8	96,71	1,20	0,62	0,38	0,38	Boa
R_33	6	97,16	0,83	0,69	0,52	0,47	Boa
R_150	17	53,17	14,36	13,45	8,44	2,66	Ruim
R_8	11	94,63	1,03	0,94	0,85	0,58	Boa
R_123	10	92,69	4,48	0,86	0,58	0,34	Boa
R_124	7	92,41	5,90	0,73	0,40	0,21	Boa
R_159	10	97,04	0,63	0,44	0,41	0,34	Boa
R_238	8	95,23	2,01	0,86	0,67	0,57	Boa
R_10	9	96,16	0,96	0,93	0,77	0,40	Boa
R_19	6	96,75	1,16	0,99	0,53	0,28	Boa
R_82	11	87,28	5,31	4,04	1,08	0,65	Regular
R_3	8	97,11	0,92	0,51	0,40	0,33	Boa
R_183	13	95,38	0,80	0,57	0,45	0,42	Boa

Os resultados obtidos no experimento 3 são apresentados na Tabela 7, sendo que, 17 regras foram classificadas como boas, 6 como regulares e 2 como ruins.

Tabela 7 - Clusters das regras gerados no experimento 3

Id Regra	Qtd de Clustres	Porcentagem de itens por clusters mais populosos (%)					Situação da Regra
		1º mais populoso	2º mais populoso	3º mais populoso	4º mais populoso	5º mais populoso	
R_25	10	54,70	42,38	1,41	0,73	0,28	Regular
R_29	3	99,62	0,24	0,13	-	-	Boa
R_16	17	53,74	14,42	13,25	8,54	2,57	Ruim
R_7	14	52,01	13,06	12,76	9,51	3,70	Ruim
R_12	8	50,42	44,97	1,70	0,90	0,80	Boa
R_11	9	92,18	4,26	0,87	0,71	0,60	Boa
R_22	11	95,02	1,08	0,70	0,57	0,38	Boa
R_17	9	92,64	3,71	0,83	0,83	0,44	Boa
R_24	8	48,91	44,50	3,42	1,25	0,66	Regular
R_26	8	95,41	0,90	0,76	0,69	0,69	Boa
R_5	9	92,98	2,22	1,18	0,97	0,62	Boa
R_13	8	96,51	0,65	0,65	0,58	0,43	Boa
R_28	4	55,89	42,12	10,92	0,95	-	Regular
R_23	4	97,17	1,31	0,91	0,60	-	Boa
R_19	4	94,87	1,78	1,67	1,67	-	Boa
R_10	6	51,63	43,90	1,93	1,45	0,60	Regular
R_18	5	51,48	43,67	2,73	1,36	0,74	Regular
R_31	2	99,21	0,78	-	-	-	Boa

R_30	5	72,37	23,29	2,27	1,03	1,03	Regular
R_8	1	100	-	-	-	-	Boa
R_21	1	100	-	-	-	-	Boa
R_14	1	100	-	-	-	-	Boa
R_2	2	83,91	16,08	-	-	-	Boa
R_1	1	100	-	-	-	-	Boa
R_6	1	100	-	-	-	-	Boa

Na Tabela 8 é mostrado que no experimento 2, 17 regras foram consideradas boas, 5 regulares e 3 ruins.

Tabela 8 - Clusters das regras gerados no experimento 2

Id Regra	Qtd de Clustres	Porcentagem de itens por clusters mais populosos (%)					Situação da Regra
		1º mais populoso	2º mais populoso	3º mais populoso	4º mais populoso	5º mais populoso	
R_58	10	51,38	45,19	0,89	0,33	0,18	Regular
R_65	8	95,55	3,33	0,47	0,21	0,17	Boa
R_83	3	99,62	0,24	0,13	-	-	Boa
R_24	1	100	-	-	-	-	Boa
R_30	10	93,30	1,51	1,29	0,93	0,93	Boa
R_23	8	88,90	5,92	3,19	0,80	0,38	Regular
R_15	14	52,01	13,06	12,77	9,51	3,70	Ruim
R_22	8	50,42	44,97	1,70	0,90	0,80	Regular
R_20	9	92,18	4,26	0,87	0,71	0,60	Boa
R_48	11	95,02	1,08	0,70	0,57	0,38	Boa
R_38	9	92,64	3,71	0,83	0,83	0,44	Boa
R_59	8	95,41	0,90	0,76	0,69	0,69	Boa
R_9	9	92,98	2,22	1,18	0,97	0,62	Boa
R_63	4	55,90	42,12	1,03	0,95	-	Regular
R_70	10	56,58	14,41	13,75	7,00	2,33	Ruim
R_28	9	39,17	32,23	14,21	7,78	2,28	Ruim
R_14	5	94,16	3,31	0,96	0,96	0,60	Boa
R_62	7	47,78	45,15	4,33	1,50	0,56	Regular
R_29	6	90,61	3,42	2,34	1,66	1,47	Boa
R_50	4	97,17	1,31	0,91	0,60	-	Boa
R_37	4	94,47	2,36	1,74	1,43	-	Boa
R_41	4	94,87	1,78	1,67	1,67	-	Boa
R_51	6	96,95	0,94	0,63	0,63	0,52	Boa
R_43	4	97,13	1,16	1,06	0,63	-	Boa
R_12	7	94,90	1,37	0,95	0,84	0,74	Boa

Os resultados do experimento 1 são expostos na Tabela 9, sendo que, 21 das regras foram consideradas boas, 2 regulares e 2 boas.

Tabela 9 - Clusters das regras gerados no experimento 1

Id Regra	Qtd de Clustres	Porcentagem de itens por clusters mais populosos (%)					Situação da Regra
		1º mais populoso	2º mais populoso	3º mais populoso	4º mais populoso	5º mais populoso	
158	8	95,55	3,34	0,47	0,21	0,18	Boa
211	3	99,62	0,24	0,13	-	-	Boa
51	1	100,00	-	-		-	Boa

28	14	52,01	13,06	12,77	9,51	3,69	Ruim
45	9	92,18	4,26	0,87	0,71	0,60	Boa
42	5	95,90	1,62	1,06	0,78	0,61	Boa
36	5	92,67	5,65	0,69	0,63	0,34	Boa
119	11	95,02	1,08	0,70	0,57	0,38	Boa
84	9	92,64	3,71	0,83	0,83	0,44	Boa
138	8	93,53	1,68	1,48	0,97	0,84	Boa
146	8	95,41	0,90	0,76	0,69	0,69	Boa
16	9	92,98	2,22	1,18	0,97	0,62	Boa
38	4	93,49	5,38	0,59	0,52	-	Boa
133	8	92,44	1,80	1,57	1,27	1,04	Boa
155	4	55,89	42,19	1,02	0,95	-	Regular
170	10	56,58	14,41	13,75	7,00	2,33	Ruim
62	5	96,14	1,42	1,00	0,75	0,67	Boa
24	5	94,16	3,31	0,95	0,95	0,61	Boa
77	7	94,75	1,47	1,47	0,74	0,64	Boa
153	7	47,78	45,15	4,33	1,50	0,56	Regular
59	6	90,60	3,42	2,34	1,66	1,46	Boa
114	5	97,74	0,68	0,58	0,49	0,49	Boa
128	4	97,17	1,31	0,90	0,61	-	Boa
87	4	94,87	1,78	1,67	1,67	-	Boa
131	6	96,53	0,94	0,63	0,63	0,52	Boa

Analisando-se separadamente os grupos de experimentos com mesmo tamanho de frases, conforme apresentado na Figura 19 (Grupo A e Grupo B), percebe-se que a medida em que se diminui o número do suporte, para um mesmo tamanho de frase, a tendência é que a qualidade das frases melhore.

Por outro lado, analisando-se os grupos de experimentos para um mesmo suporte, com diferentes tamanhos de frases conforme a Figura 19 (Grupos 1 e Grupo 2), esperava-se que tamanhos de frases maiores produzissem regras de melhores qualidades, porém, os experimentos realizados não foram capazes de comprovar essa hipótese.

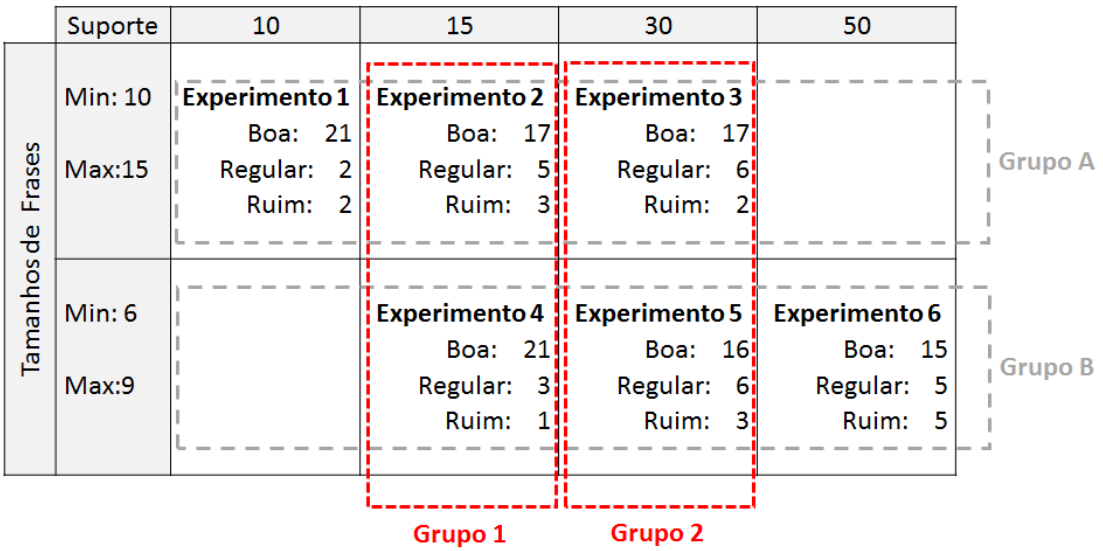


Figura 19 - Qualidade das regras geradas por experimento

4.4.2 Análise dos Resultados

Após a avaliação das regras geradas, o passo seguinte é a avaliação dos resultados obtidos pela aplicação dessas regras. Com o intuito de se deixar os trabalhos de análise mais concisos, utilizou-se como parâmetro apenas as regras geradas no experimento 6. Porém, esse procedimento pode ser replicado para qualquer um dos demais experimentos realizados. A análise também só avalia as regras consideradas como "Boas", dessa forma, o número inicial de 25 regras geradas cai para um total de 15 regras.

Conforme mencionado anteriormente, essa avaliação faz uma análise qualitativa dos resultados obtidos. A seleção do conjunto amostral a ser analisado é feita de forma a identificar aquelas compras com maior probabilidade de terem sido classificadas de forma errônea. Sendo assim, para cada uma das regras classificadas como "Boa", foram identificados outliers, considerando-se três atributos diferentes para o cálculo desses outliers: "Valor unitário", "unidade de medida" e "marca". A escolha desses atributos se deu pelo fato de que produtos iguais tendem a possuir valores unitários próximos e unidades de medida e marca semelhantes, ou seja, caso um determinado produto apresente valor, unidade de medida ou marca muito diferente dos demais produtos do mesmo tipo, ele será classificado como um outlier, e terá uma probabilidade maior de ter sido identificado de forma errônea, sendo assim selecionado para uma análise mais minuciosa. No entanto, cabe ressaltar que tal consideração é apenas uma premissa, visto que, é possível que produtos iguais tenham marcas, unidades de medida ou valores unitários totalmente diferentes. Dessa forma, esse é apenas um critério para a seleção da amostra a ser analisada, porém, a qualidade dos resultados é feita pela análise individualizada das compras selecionadas.

Quanto a obtenção dos atributos considerados no processo de clusterização, o campo "Valor unitário" já estava presente na base de dados utilizada, como um atributo estruturado. Os campos "unidade de medida" e "marca" foram obtidos durante a etapa de pré-processamento, descrita na Seção 3.3.1.

O algoritmo utilizado para a detecção dos outliers foi o apresentado em (RAMASWAMY; RASTOGI; SHIM, 2000). No algoritmo em questão, a detecção de outliers é feita através do cálculo da distância de um ponto a seus *q-ésimos* vizinhos mais próximos. Dessa forma, cada ponto é classificado com base na sua distância a seus *qésimo* vizinhos mais próximos, e os r pontos superiores, neste ranking (ou seja, os r pontos com maiores distâncias a seus vizinhos) são declarados como outliers. Logo, os valores de q e

r podem ser definidos como o número de vizinhos a serem considerados e o número de outliers a serem identificados, respectivamente.

Na execução desse estudo, utilizou-se como parâmetro q (número de pontos vizinhos a serem considerados) o valor 10, e o número de outliers a serem encontrados foi definido como sendo 2 (r =2). Sendo que, para o cálculo da distância entre os pontos foi utilizada a fórmula da distância euclidiana.

Tabela 10 - Outliers identificados por regra

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho[17]	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800889	1	R_2	Diesel	100.000.000,00	Litro	Petrobras
2015NE800586	1	R_2	Diesel	87.279.000,00	Litro	Petrobras
2015NE800803	1	R_3	Água mineral	68.611,32	Garrafao 20 L	Seiva
2015NE800466	1	R_3	Água mineral	52.602,06	Garrafao 20 L	Seiva
2015NE804597	1	R_4	Banana	16.388,00	QUILOGRAMA	CEASA
2015NE803169	1	R_4	Banana	15.000,00	QUILOGRAMA	CEASA
2015NE801459	12	R_6	Produto perecível	9,65	Kg	frigolaste
2015NE801459	8	R_6	Produto perecível	21,46	Kg	sabadini
2015NE801077	1	R_10	Gás liquefeito - glp	63.982,92	KG	GASBALL
2015NE801613	1	R_10	Gás liquefeito - glp	120.000,00	KG	GASBALL
2015NE800466	1	R_11	Água mineral	52.602,06	GALAO 20,00 L	calogi
2015NE800899	1	R_11	Água mineral	61.800,48	GALAO 20,00 L	Hydrate
2015NE800375	1	R_12	Bequer de vidro	350,00	UNIDADE	Leica Biosystems
2015NE806768	1	R_12	Bequer de vidro	900,00	UNIDADE	VELP
2015NE800075	2	R_14	Proveta de vidro	7,00	UNIDADE	rav
2015NE800736	1	R_14	Proveta de vidro	4,60	UNIDADE	Uniglass
2015NE800068	4	R_15	Gasolina Comum	553.834,29	Litros	xxxxxxxxxx
2015NE800809	1	R_15	Gasolina Comum	7.157.000,00	Litro	SHELL
2015NE800588	1	R_16	Balão volumétrico para laboratório	530,00	UNIDADE	-
2015NE802470	1	R_16	Balão volumétrico para laboratório	615,88	UNIDADE	DIOGOLAB
2015NE800001	16	R_19	Resistor filme metálico	0,02	UNIDADE	RohmRohm
2015NE800001	34	R_19	Resistor filme metálico	0,02	UNIDADE	RohmRohm
2015NE800042	7	R_22	Caneta esferográfica	1,10	CAIXA 1.200,00 UN	esferografica
2015NE800006	2	R_22	Caneta esferográfica	135,10	CAIXA 12,00 UN	slider
2015NE800089	2	R_23	Álcool etílico hidratado combustível	78.348,81	LITRO	xxxxxxxxxx

[17] O código do empenho não está completo (com os 23 dígitos) para não possibilitar a identificação da unidade que executou a compra, visto que esse trabalho não tem foco na área de auditoria, o único objetivo desse procedimento é fazer a identificação de compras consideradas como outliers e que por essa razão possam ter sido classificadas de forma errônea.

2015NE800549	3	R_23	Álcool etílico hidratado combustível	129.552,00	LITRO	IPIRANGA
2015NE800068	3	R_24	Álcool anidro combustível	553.834,29	Litros	xxxxxxxxxx
2015NE800876	3	R_24	Álcool anidro combustível	148.823,93	Litros	NACIONAL
2015NE800035	1	R_25	Peça para automóvel	560.000.000,00	MENSAL	Conforme Edital
2015NE800134	1	R_25	Peça para automóvel	1.000.000.000,00	UNIDADE - PECAS	ORIGINAL

Na Tabela 10 são apresentados os resultados obtidos pelo algoritmo para cada uma das regras analisadas. As 2 primeiras colunas trazem a identificação da compra, a terceira e quarta coluna apresentam a regra utilizada e o nome do produto identificado pela regra, respectivamente, e as demais colunas apresentam os atributos utilizados para a detecção dos outliers.

Após a definição da amostra com os registros com maiores probabilidades de terem sido classificados erroneamente em cada uma das regras selecionadas (outliers), o passo seguinte é a análise de cada um desses registros de forma individualizada, nas páginas do Portal da Transparência, a fim de verificar se esses registros realmente correspondem ao mesmo grupo de produtos que a regra se propõem a identificar, ou se eles se referem a outros tipos de produtos e consequentemente tenham sido classificados de forma errada pelas regras.

Sendo assim, todos os 30 registros apresentados na Tabela 19 (2 outliers por regra) foram analisados e verificou-se de forma detalhada os textos descritivos das especificações de compras a fim de se averiguar a que se referia cada uma das compras, possibilitando assim a comparação com o tipo de produto que as regras estavam identificando.

Nessa análise, não foram identificados produtos classificados de forma errônea, porém, algumas das regras geradas, mesmo sendo consideradas como boas, ainda que classificando as compras de maneira correta, faziam uma classificação muito genérica, que dependendo da finalidade para que se deseje utilizar o processo de mineração de frases desenvolvido, possa não atender os objetivos por completo. Exemplos desses tipos ocorrem com as regras R_6 e R_25, que identificam as compras de “produto perecível” e “peça para automóvel”, respectivamente, ou seja, uma forma muito genérica para se identificar um produto.

Para evitar situações como estas, a proposta prevê uma etapa opcional, apresentada na Seção 3.3.6 (refinamento de regras), em que especialistas podem, dentre outras atividades, ajustar as regras de acordo com o grau de especificidade que se deseja dar ao produto.

Os resultados desses experimentos também exemplificam outra atividade que pode ser desempenhada por especialista nessa etapa opcional, que é a junção de regras que identificam o mesmo tipo de produto. Por exemplo, as regras R_3 e R_11 levam ao mesmo produto (água mineral), e por isso poderiam ser agrupadas. Também dependendo do grau de especificidade que se deseja dar ao processo, os especialistas poderiam juntar as regras R_23 e R_24, que identificam respectivamente "álcool etílico hidratado combustível" e "álcool anidro combustível" para levarem a um único produto, um pouco mais genérico, denominado "álcool combustível".

Na Figura 20 são apresentados dois recortes de telas do portal da transparência, cada um com a descrição de uma das compras caracterizada como outliers obtidos na aplicação da regra R_3. Esses recortes de tela permitem verificar que apesar dos valores discrepantes, as especificações das compras realmente se referem ao produto esperado, ou seja, Água Mineral. O Apêndice I dessa dissertação apresenta uma análise mais detalhada dos outliers mostrados na Tabela 10.

2015NE800899

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
7 - GENEROS DE ALIMENTACAO	1	61.800,48	61.800,48	0000000001,00000 GALÃO 20,00 L ÁGUA MINERAL, MATERIAL ÁGUA MINERAL, TIPO EMBALAGEM PLÁSTICO POLICARBONATO TRANSPARENTE, GASEIFICAÇÃO SEM GÁS, CARACTERÍSTICAS ADICIONAIS COM TAMPA DE PRESSÃO/LACRE/ENVASADO MECANICAMENTE/, NORMAS TÉCNICAS CONFORME PORTARIA DE CORRELATOS DO MINISTÉRIO SAÚ- MARCA: Hydrate ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000304461

2015NE800466

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
7 - GENEROS DE ALIMENTACAO	1	52.602,06	52.602,06	0000000001,00000 GALÃO 20,00 L ÁGUA MINERAL, MATERIAL ÁGUA MINERAL, TIPO EMBALAGEM PLÁSTICO, GASEIFICAÇÃO SEMGÁS MARCA: calogi ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000217773

Figura 20 - Recortes das telas do Portal das Transparência para registros considerados outliers da regra R_3

Capítulo 5 – Aplicações

Uma vez identificado a que produto cada uma das descrições de compras se refere, uma série de outras análises tornam-se viáveis. O objetivo desse capítulo é fazer a apresentação de algumas aplicações possíveis de serem realizadas com os dados obtidos nos experimentos executados no Capítulo 4.

5.1 Cálculo de Preços de Referência dos Produtos Comprados Pela Administração Pública

Conforme sugerido em (CARVALHO et al., 2013), a partir do momento em que se tem os produtos devidamente identificados, torna-se possível se propor preços de referência para os diversos produtos que são comprados pela Administração Pública Federal e estão sendo apresentados em portais de transparência.

Tabela 11 - Amostra de preços de referência calculados

Produto	Unidade	Quantidade Comprada	Preço de Referência (mediana)
Água mineral	Galão 20 l	1989	R$ 7,90
Água mineral	Garrafa 500 ml	638	R$ 0,74
Água mineral	Garrafa 1,5 l	183	R$ 1,24
Água mineral	Copo 200 ml	107	R$ 0,29
Banana	KG	2783	R$ 1,96
Banana	Unidade	116	R$ 0,52
Gasolina	Litro	5410	R$ 3,68
Diesel	Litro	3887	R$ 2,94
Gás - GLP	Botijão 45 Kg	181	R$ 199
Gás - GLP	Botijão 13 Kg	230	R$ 52
Álcool combustível	Litro	1678	R$ 3,03
Caneta esferográfica	Unidade	1113	R$ 0,38
Caneta esferográfica	Caixa 50 unidades	439	R$ 20,90

Na Tabela 11 é mostrada uma sugestão de preço de referência para alguns desses produtos identificados. Nesse cálculo, considerou-se os preços de referência dos produtos como sendo a mediana dos valores unitários pagos em cada uma das compras desses produtos apresentados no Portal da Transparência, visto que, esta métrica está menos suscetível a influência de outliers.

5.2 Identificação de Compras com Preços Muito Acima do Esperado

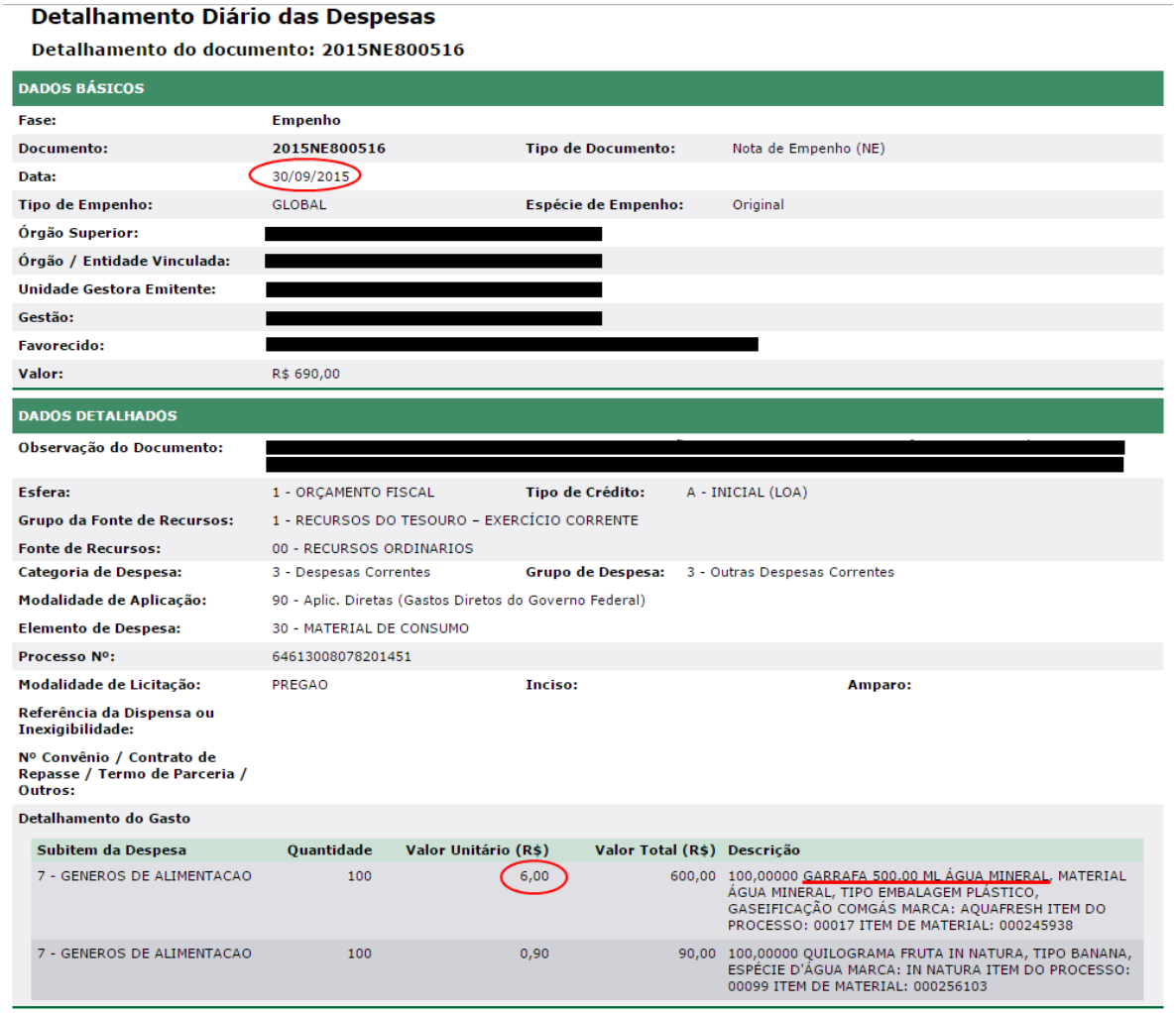
A partir do momento em que se consegue estabelecer um preço de referência para os diversos produtos comprados e apresentados nos portais de transparência, torna-se possível também se identificar compras que tenham sido feitas com valores muito acima do esperado.

Na Tabela 12 é apresentada uma amostra de 2 valores muito acima do esperado para cada um dos exemplos de preços de referência identificados na Tabela 11. Essa tabela é apenas exemplificativa, visto que, devido ao grande número de compras apresentadas no Portal da Transparência, o número de compras consideradas muito acima do preço de referência também é elevado.

Tabela 12 - Amostra de preços muito acima do esperado

Produto	Unidade	Preço de referência	Número do empenho	Valor Unitário
Água mineral	Galão 20 l	7,90	2015NE802221	R$ 29,05
			2015NE800994	R$ 38,00
Água mineral	Garrafa 500 ml	0,74	2015NE800516	R$ 6,00
			2015NE800274	R$ 25,50
Água mineral	Garrafa 1,5 l	1,24	2015NE801472	R$ 6,78
			2015NE800273	R$ 3,42
Água mineral	Copo 200 ml	0,29	2015NE800968	R$ 0,85
			2015NE800344	R$ 0,65
Banana	KG	1,96	2015NE800558	R$ 4,97
			2015NE800947	R$ 12,80
Banana	Unidade	0,52	2015NE800335	R$ 8,87
			2015NE800564	R$ 3,91
Gasolina	Litro	3,68	2015NE800825	R$ 11,99
			2015NE801011	R$ 16,21
Diesel	Litro	2,94	2015NE800014	R$ 11,70
			2015NE800089	R$ 14,99
Gás - GLP	Botijão 45 Kg	199	2015NE800419	R$ 283,45
			2015NE800066	R$ 270,00
Gás - GLP	Botijão 13 Kg	52	2015NE801546	R$ 120,00
			2015NE800416	R$ 140,62
Álcool combustível	Litro	3,03	2015NE800302	R$ 8,08
			2015NE800119	R$ 7,34
Caneta esferográfica	Unidade	0,38	2015NE800035	R$ 8,00
			2015NE800712	R$ 5,27
Caneta esferográfica	Caixa 50 unidades	20,90	2015NE800019	R$ 51,33
			2015NE800478	R$ 32,50

Cabe ressaltar que esses valores elevados apresentados na Tabela 12 não são suficientes para se dizer que tenha havido irregularidades nos referidos processos de compras, visto que, qualquer indício levantado por meio da análise de dados carece de uma averiguação mais aprofundada por meio de auditorias específicas. Também não faz parte do escopo desse trabalho qualquer tipo de análise de compras individuais. No entanto, todos os processamentos sugeridos nessa dissertação tornam viável a identificação de compras que fogem do padrão esperado, possibilitando novos tipos de análises que seriam impossíveis de serem feitas com a informação apresentada em formato original (formato textual).



Detalhamento Diário das Despesas

Detalhamento do documento: 2015NE800516

DADOS BÁSICOS

Fase:	**Empenho**		
Documento:	**2015NE800516**	**Tipo de Documento:**	Nota de Empenho (NE)
Data:	30/09/2015		
Tipo de Empenho:	GLOBAL	**Espécie de Empenho:**	Original
Órgão Superior:	[redacted]		
Órgão / Entidade Vinculada:	[redacted]		
Unidade Gestora Emitente:	[redacted]		
Gestão:	[redacted]		
Favorecido:	[redacted]		
Valor:	R$ 690,00		

DADOS DETALHADOS

Observação do Documento:	[redacted]		
Esfera:	1 - ORÇAMENTO FISCAL	**Tipo de Crédito:**	A - INICIAL (LOA)
Grupo da Fonte de Recursos:	1 - RECURSOS DO TESOURO – EXERCÍCIO CORRENTE		
Fonte de Recursos:	00 - RECURSOS ORDINARIOS		
Categoria de Despesa:	3 - Despesas Correntes	**Grupo de Despesa:**	3 - Outras Despesas Correntes
Modalidade de Aplicação:	90 - Aplic. Diretas (Gastos Diretos do Governo Federal)		
Elemento de Despesa:	30 - MATERIAL DE CONSUMO		
Processo Nº:	64613008078201451		
Modalidade de Licitação:	PREGAO	**Inciso:**	**Amparo:**
Referência da Dispensa ou Inexigibilidade:			
Nº Convênio / Contrato de Repasse / Termo de Parceria / Outros:			

Detalhamento do Gasto

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
7 - GENEROS DE ALIMENTACAO	100	6,00	600,00	100,00000 GARRAFA 500.00 ML ÁGUA MINERAL, MATERIAL ÁGUA MINERAL, TIPO EMBALAGEM PLÁSTICO, GASEIFICAÇÃO COMGÁS MARCA: AQUAFRESH ITEM DO PROCESSO: 00017 ITEM DE MATERIAL: 000245938
7 - GENEROS DE ALIMENTACAO	100	0,90	90,00	100,00000 QUILOGRAMA FRUTA IN NATURA, TIPO BANANA, ESPÉCIE D'ÁGUA MARCA: IN NATURA ITEM DO PROCESSO: 00099 ITEM DE MATERIAL: 000256103

Figura 21 - Tela do Portal da Transparência do Governo Federal com empenho de produtos com preços muito acima do esperado

Para ilustrar as diferenças que podem ser encontradas no Portal da Transparência, na Figura 21 é apresentada uma cópia da tela de um documento de empenho presente no portal. Nessa tela são apresentadas várias informações sobre o documento de empenho (as informações que permitem a identificação da unidade que emitiu esse documento de empenho foram tarjadas) e em destaque estão a descrição textual do produto a ser comprado, bem como o valor unitário a ser pago. Dessa forma, pode-se verificar que foi pago um valor de R$ 6,00 para a garrafa de 500 ml de água mineral, bem acima do valor de referência, que é de R$ 0,78.

5.3 Possibilidade de Cancelar uma Compra Superfaturada Antes de sua Concretização

Outro benefício em se obter um preço de referência para cada um dos produtos identificados com o processo sugerido, é a possibilidade de se fazer um controle preventivo de compras superfaturadas. Como todo o processo foi concebido para ser realizado utilizando o framework Apache Spark, é possível que a aplicação das regras de identificação dos produtos seja feita em tempo real, ou seja, no mesmo momento em que se faz a carga diária dos dados do Portal da Transparência.

Dessa forma, a identificação dos produtos a que se referem as compras que estão sendo carregadas torna possível a comparação entre o preço a ser pago pelo produto em questão, com o seu preço de referência, calculado anteriormente, permitindo-se assim a detecção de disparidades.

A fase de empenho (fase em que é aplicado o procedimento de identificação dos produtos) é a primeira etapa do processo de execução da despesa, nessa etapa ainda não houve a consumação da despesa. Logo, pode-se interromper esse procedimento de compra ou sugerir-se a renegociação dos preços praticados para valores economicamente mais favoráveis para a Administração Pública.

Detalhamento Diário das Despesas

Detalhamento do documento: 2016OB800192

DADOS BÁSICOS			
Fase:	**Pagamento**		
Documento:	**2016OB800192**	**Tipo de Documento:**	Ordem Bancária (OB)
Data:	02/03/2016	**Tipo de OB:**	OBC/OBB PARA TERCEIROS EM OUTROS BANCOS
Órgão Superior:	52000 - MINISTERIO DA DEFESA		
Órgão / Entidade Vinculada:	52121 - COMANDO DO EXERCITO		
Unidade Gestora Emitente:	160291 - CENTRO TECNOLOGICO DO EXERCITO		
Gestão:	00001 - TESOURO NACIONAL		
Favorecido:	17.471.773/0001-59 - ROJAO COMERCIAL DE ALIMENTOS LTDA - EPP		
Valor:	R$ 690,00		

DADOS DETALHADOS			
Observação do Documento:	PGTO DANFE 000.000.417 DE 03FEV2016 - 2015NE8000516 DE 30SET2015 - OPTANTE DO SIMPLES - 2014PR00013-UASG 160327 - PI: K1DTDEFOUTR - PTRES: 085621 - AQUISIÇÃO DE GENEROS DE ALIMENTAÇÃO.		
Processo Nº:			
Categoria de Despesa:	3 - Despesas Correntes	**Grupo de Despesa:**	3 - Outras Despesas Correntes
Modalidade de Aplicação:	90 - Aplic. Diretas (Gastos Diretos do Governo Federal)		
Elemento de Despesa:	30 - MATERIAL DE CONSUMO		

Detalhamento do Documento

Empenho	Subitem da Despesa	Cancelamento / Estorno	Convênio / Outros	Valor (R$)
2015NE800516	7 - GENEROS DE ALIMENTACAO	Não	0	690,00

Figura 22 - Tela de pagamento do Portal da Transparência do Governo Federal

Por exemplo, o empenho apresentado na Figura 21 foi emitido no dia 30/09/2015 (conforme pode ser observado no campo data, na parte superior da Figura 21). Porém, o pagamento referente a essa despesa foi feito pela Ordem Bancária 2016OB800192, cujo recorte da tela do Portal da Transparência é apresentado na Figura 22. Observando-se a data de emissão dessa ordem bancária (campo data na parte superior da Figura 22) verifica-se que o pagamento, referente a essa despesa, só foi executado no dia 02/03/2016, ou seja, aproximadamente 5 meses após a emissão do documento de empenho e da carga dessa informação no Portal da Transparência, ou seja, tempo suficiente para que fossem tomadas as devidas providências com o intuito de se evitar a concretização de uma compra superfaturada.

5.4 Comparação Entre Valores Pagos em Compras Licitadas e Não Licitadas

A Lei 8666/93 prevê que todas as compras realizadas pela Administração Pública devem ser precedidas de licitação, no entanto, essa mesma lei também prever algumas situações em que o procedimento licitatório pode ser dispensável ou inexigível.

Um outro tipo de análise que pode ser feita a partir dos resultados obtidos pela técnica proposta nessa dissertação é a comparação entre os preços praticados nas compras de produtos em situações em que houve licitação e nos casos em que esse procedimento não ocorreu.

Na Tabela 13 são apresentados os preços praticados, com e sem procedimento licitatório, para os mesmos produtos listados na Tabela 11, sendo que, nesse caso, assim como na Tabela 11, os preços foram obtidos pela mediana dos valores unitários de cada uma das compras.

Tabela 13 - Amostra de preços praticados em compras com e sem licitação

Produto	Unidade	Preço Praticado (mediana)	
		Com Licitação	Sem Licitação
Água mineral	Galão 20 l	R$ 7,48	R$ 9,90
Água mineral	Garrafa 500 ml	R$ 0,71	R$ 1,35
Água mineral	Garrafa 1,5 l	R$ 1,24	R$ 2,83
Água mineral	Copo 200 ml	R$ 0,29	--
Banana	KG	R$ 1,96	R$ 2,85
Banana	Unidade	R$ 0,52	R$ 0,58
Gasolina	Litro	R$ 3,62	R$ 3,93
Diesel	Litro	R$ 2,93	R$ 2,98
Gás - GLP	Botijão 45 Kg	R$ 199,00	R$ 225,00
Gás - GLP	Botijão 13 Kg	R$ 51,19	R$ 53,50
Álcool combustível	Litro	R$ 3,28	R$ 2,59
Caneta esferográfica	Unidade	R$ 0,36	R$ 0,55
Caneta esferográfica	Caixa 50 unidades	R$ 20,45	R$ 35,00

Como pode ser observado na Tabela 13, os produtos tendem a ser comprados por um preço maior quando não há um procedimento licitatório anterior a essa compra. A única exceção identificada ocorreu para o caso de compra de álcool combustível, pois nessa situação as compras feitas por dispensa de licitação apresentaram um valor menor do que aquelas que foram precedidas do processo licitatório. Cabe ressaltar que para o caso de compra de copo de água mineral de 200 ml todas as compras foram precedidas do procedimento licitatório. Porém, essas análises só foram possíveis de serem realizadas pelo fato dos produtos terem sido anteriormente identificados.

5.5 Análise das Marcas Mais Compradas

A identificação dos produtos também permite a descoberta das marcas mais compradas para cada tipo de produto. Na Tabela 14 são apresentadas as 3 marcas com maior número de compras para uma amostra de 7 tipos de produtos.

Pela análise da Tabela 14, pode se verificar que, via de regra, as marcas consideradas como referência, para os diversos tipos de produtos, são as que mais são compradas pela Administração Pública.

Tabela 14 - Amostra de marcas mais compradas por produtos

Produto	Marca
Açúcar Refinado	União
	Caravelas
	Alto alegre
Caneta esferográfica	Bic
	Compactos
	Jocar
Conexão hidráulica	Krona
	Copacol
	Tigre
Farinha Láctea	Nestle
	Mococa
	Mutiday
Gás (GLP)	Supergasbras
	Ultragaz
	Liquigas
Milho em conserva	Predilecta
	Goias Verde
	Quero
Papel Sulfite	Chamex
	One
	Report
Pilha pequena	Elgin
	Alfacell
	Maxprint
Pneu	Firestone
	JK
	Bridgestone
Toner para Impressora	HP
	Fastprinter
	DSI

5.6 Identificação de Fornecedores Vendendo o Mesmo Produto com Preços Diferentes

A partir do momento em que se tem os produtos devidamente identificados, pode-se juntar essas informações com outras informações que já estão estruturadas na base de dados do Portal da Transparência, a fim de se enriquecer as análises feitas. Um exemplo disso é a comparação de diferentes unidades gestoras comprando o mesmo produto, de um mesmo fornecedor, mas com preços bem diferentes.

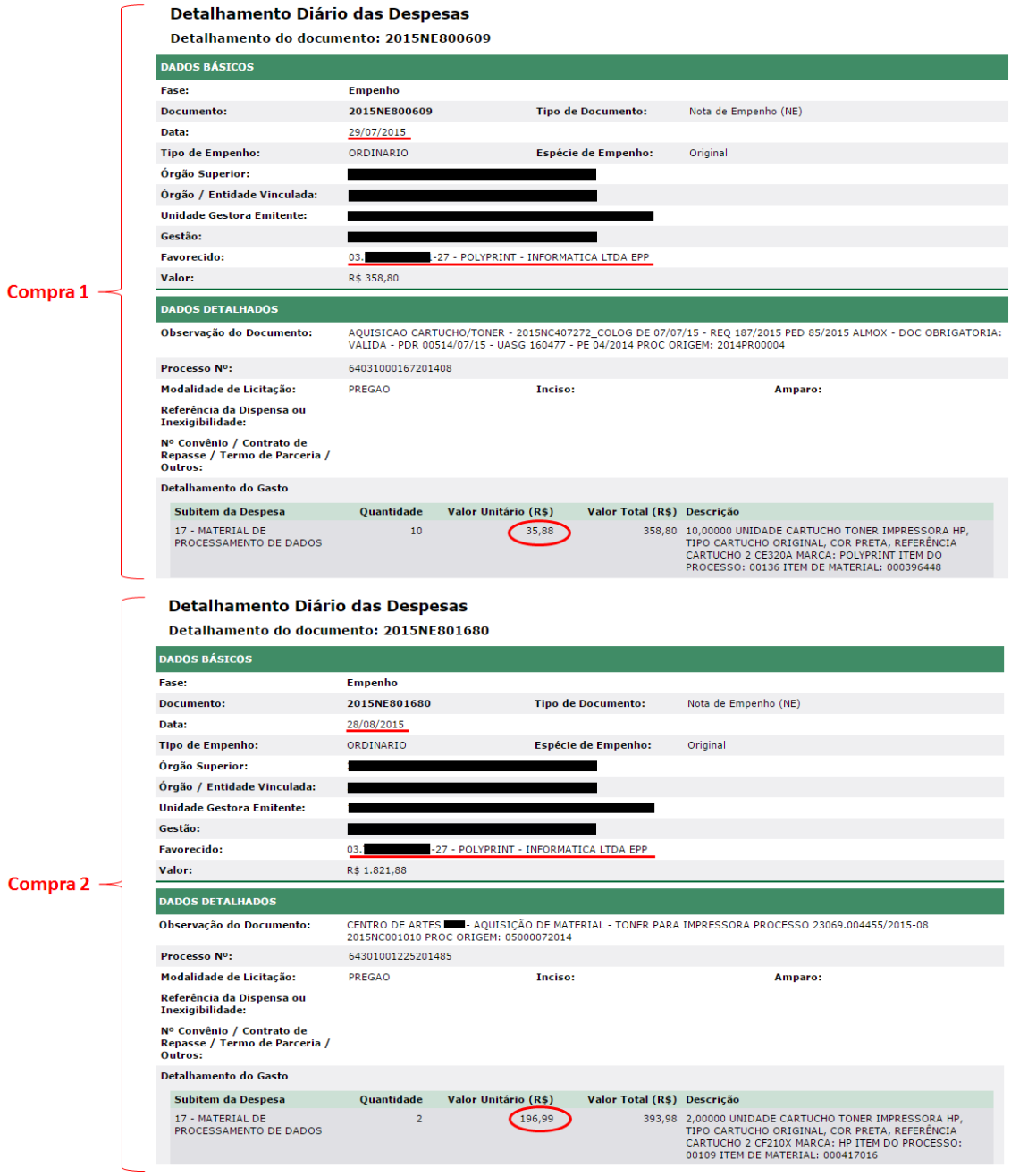


Figura 23 - Telas do Portal da Transparência com o mesmo fornecedor vendendo o mesmo produto com preços diferentes

Na Figura 23 são apresentados dois recortes de telas do portal da transparência. Cada uma dessas telas apresenta uma compra de toner para impressora, sendo que em ambas as compras o fornecedor do produto é o mesmo, porém, o preço a ser pago tem uma variação de mais de 500 %, sendo que, a segunda compra ocorreu um mês após a primeira.

Esse caso apresentado é apenas um dos muitos casos semelhantes identificados. Essas disparidades acontecem porque muitas vezes os processos de compra ocorrem de

maneira independente, sendo que, uma unidade gestora não fica sabendo do preço que uma outra unidade gestora está pagando pelo mesmo produto ao mesmo fornecedor.

A metodologia ora proposta pode ajudar na otimização das compras realizadas pela Administração Pública, permitindo que a unidade gestora possa propor renegociações de preços a serem pagos, quando identificados valores economicamente mais vantajosos sendo praticados pelo mesmo fornecedor, na venda do mesmo produto para outros compradores da Administração Pública.

5.7 Acompanhar a Tendência dos Preços de Produtos

A identificação dos produtos que estão sendo descritos também é útil para se traçar o comportamento dos preços praticados pela Administração pública ao longo de um determinado período de tempo. Na Tabela 15 é apresentado o preço do litro da gasolina pago pelo governo nos diferentes meses do ano de 2015.

Tabela 15 - Amostra dos preços do litro da gasolina durante o ano de 2015

Mês	Qtd de Compras	Preço Calculado
Janeiro	163	R$ 3,09
Fevereiro	282	R$ 3,13
Março	377	R$ 3,30
Abril	437	R$ 3,49
Maio	516	R$ 3,70
Junho	561	R$ 3,85
Julho	481	R$ 4,16
Agosto	432	R$ 3,85
Setembro	436	R$ 3,85
Outubro	396	R$ 4,02
Novembro	661	R$ 3,62
Dezembro	667	R$ 5,40

Pela análise da Tabela 15, pode-se constatar que o preço do litro da gasolina pago pelo governo teve uma tendência de alta durante praticamente todos os meses do ano, a exceção dos meses de agosto e novembro, que registraram quedas, e do mês de setembro que manteve o mesmo patamar do mês anterior.

Esse mesmo tipo de estudo pode ser feito para os demais produtos adquiridos pela Administração Pública, a fim de se verificar não só a tendência dos preços praticados, mas também compará-los com fatores externos à Administração Pública, como questões de sazonalidade de produtos, influência de atividades econômicas no Brasil e no exterior e etc.

5.8 Comparar a Relação Entre Preço Pago e Quantidade Comprada

Um outro tipo de análise que pode ser feita a partir do momento em que se tem os produtos devidamente identificados é a verificação se há algum tipo de relação entre o preço pago por um determinado produto com a quantidade que está sendo comprada, ou seja, se há algum tipo de economia de escala para as compras feitas pela Administração Pública.

As análises realizadas não demonstraram uma relação direta entre os preços pagos pelos produtos e a quantidade comprada. No entanto, os estudos evidenciaram que a maioria dos outliers de preços identificados tinha uma relação direta com a quantidade especificada.

Apesar dos cálculos dos preços de referência, realizados na seção 5.1 não indicarem valores discrepantes para os preços de referências obtidos a partir das compras realizadas pelo Governo Federal, muitas compras apresentavam valores unitário extremamente altos, caracterizando grandes diferenças de preços. Ainda que tais valores não chegassem a prejudicar os cálculos dos preços de referência, realizadas na seção 5.1, pois o número de compras realizadas dentro dos padrões esperados era muito superior a esses casos, tal fato caracterizava uma recorrência que merecia ser estudada.

Quando se estudou a relação entre os preços praticados e as quantidades compradas, verificou-se que a grande maioria dos casos em que havia valores extremamente altos (muito acima do esperado) ocorria quando a quantidade a ser comprada era igual a exatamente um (uma unidade, um quilograma, um litro, etc.).

A Tabela 16 monstra alguns casos de exemplos em que o valor unitário do produto é extremamente alto, e que a quantidade a ser comprada é exatamente 1.

Observando-se esses valores unitários extremamente altos, infere-se que esses são frutos de preenchimentos incorretos do quantitativo a ser comprado no sistema, pois o valor unitário é obtido pela divisão do valor total a ser pago ao fornecedor (na venda do referido produto) pela quantidade total que está sendo adquirira. Logo, nessas situações, o valor unitário do produto fica muito alto, pois considera-se como valor unitário o preço pago para se adquirir uma grande quantidade de produtos (quantidade essa que não se sabe exatamente qual é).

Tabela 16 - Amostra de produtos com valores unitários muito altos

Produto	Preço de Referência	Número do Empenho	Valor Unitário	Quantidade Comprada	Valor Total
Água Mineral (20 litros)	R$ 7,90	2015NE800466	R$ 52.602,06	1	R$ 52.602,06
		2015NE801154	R$ 43.314,89	1	R$ 43.314,89
Álcool Combustível (Litro)	R$ 3,03	2015NE800068	R$ 553.834,29	1	R$ 553.834,29
		2015NE800876	R$ 148.823,93	1	R$ 148.823,93
Diesel (Litro)	R$ 1,24	2015NE801025	R$ 817.921,18	1	R$ 817.921,18
		2015NE800876	R$ 239.499,28	1	R$ 239.499,28
Gasolina (Litro)	R$ 2,94	2015NE800877	R$ 453.444,49	1	R$ 453.444,49
		2015NE801023	R$ 249.836,66	1	R$ 249.836,66
Banana (Quilograma)	R$ 0,52	2015NE804597	R$ 16.388,00	1	R$ 16.388,00
		2015NE805293	R$ 7.542,00	1	R$ 7.542,00
Gás (Botijão 13 Kg)	R$ 52,00	2015NE800124	R$ 13.219,20	1	R$ 13.219,20
		2015NE800747	R$ 6.500,00	1	R$ 6.500,00

Percebe-se, que esse tipo de caso ocorre recorrentemente, o que prejudica a transparência dos gastos públicos, inviabilizando a verificação dos preços pagos pelos produtos identificados, assim como outros estudos que poderiam ser feitos a partir de tais informações. Logo, a metodologia proposta também se demonstra útil para a identificação de práticas administrativas erradas, possibilitando assim a melhora da gestão das atividades de compra executadas pelo governo.

Tabela 17 - Amostra de produtos com valores unitários muito baixos

Produto	Preço de Referência	Número do Empenho	Valor Unitário	Quantidade Comprada	Valor Total
Água Mineral (20 litros)	R$ 7,90	2015NE802221	R$ 0,00	5.075	R$ 10,00
		2015NE800994	R$ 0,00	3.000	R$ 7,90
Álcool Combustível (Litro)	R$ 3,03	2015NE800516	R$ 0,00	200.000	R$ 1.000,00
		2015NE800274	R$ 0,00	82.368	R$ 286,46
Diesel (Litro)	R$ 1,24	2015NE801472	R$ 0,00	351.510	R$ 1.180,58
		2015NE800273	R$ 0,00	291.000	R$ 9,72
Gasolina (Litro)	R$ 2,94	2015NE800113	R$ 0,00	5.777.915	R$ 21,21
		2015NE800344	R$ 0,00	123.600	R$ 0,50
Banana (Quilograma)	R$ 0,52	2015NE800558	R$ 0,01	1.500	R$ 15,60
		2015NE800947	R$ 0,00	8.000	R$ 0,80
Gás (Botijão 13 Kg)	R$ 52,00	2015NE800335	R$ 3,33	300	R$ 1.000,00

Cabe ressaltar ainda que também foi identificado outro tipo de erro recorrente no preenchimento dos quantitativos a serem comprados. Conforme pode ser visto na Tabela 17, também acontecem situações em que a quantidade a ser comprada é preenchida com um valor extremamente alto, que faz com que o valor unitário calculado (que é obtido pela divisão do valor total pela quantidade) caia para valores extremamente baixos e irreais, sendo até mesmo, em algumas situações aproximado para zero, o que é flagrantemente inviável. Essa situação, a exemplo da exposta anteriormente também tem

um reflexo negativo na transparência, visto que também torna inviável a identificação do preço a ser pago, evidenciando uma prática nociva ao controle das contas públicas.

Essas situações são exemplificadas na Figura 24 através da apresentação de 2 recortes de telas do Portal da Transparência, uma com valor unitário extremamente alto (Figura 24.a) e outro com valor extremamente baixo (Figura 24.b).

2015NE800466

Detalhamento do Gasto

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
7 - GENEROS DE ALIMENTACAO	1	52.602,06	52.602,06	0000000001,00000 GALÃO 20,00 L ÁGUA MINERAL, MATERIAL ÁGUA MINERAL, TIPO EMBALAGEM PLÁSTICO, GASEIFICAÇÃO SEMGÁS MARCA: calogi ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000217773

(a) Valor Unitário Extremamente Alto

2015NE800260

Detalhamento do Gasto

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
7 - GENEROS DE ALIMENTACAO	5.075	0,00	10,00	0000005075,00000 GALÃO 20,00 L ÁGUA MINERAL, MATERIAL ÁGUA MINERAL, TIPO EMBALAGEM GARRAFÃO DE POLIPROPILENO,GASEIFICAÇÃO SEM GÁS, CARACTERÍSTICAS ADICIONAIS TAMPA, LACRE, SEM VASILHAME EVALIDADE MÍNIMA DE, NORMAS TÉCNICAS CONFORME PORTARIA DE CORRELATOS DO MINISTÉRIO SAÚ- MARCA: FLORATA ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000332485

(b) Valor Unitário Extremamente Baixo

Figura 24 - Recorte de telas do Portal da Transparência de compras com valores unitários discrepantes

5.9 Outras Aplicações

As aplicações expostas nesse capítulo são apenas alguns exemplos de possíveis utilizações para a identificação de produtos a partir da metodologia proposta nessa dissertação. Porém, existe uma série de outras aplicações possíveis, como por exemplo:

- Aplicação de regras de associação para se identificar a probabilidade de um órgão comprar um determinado produto, dado que ele já tenha comprado um conjunto de outros tipos de produtos;
- A identificação dos órgãos que compram com melhores preços e aqueles que pagam mais caro pelos produtos
- Verificação se há algum padrão de comportamento entre empresas fornecedoras de produtos que possam caracterizar algum tipo de conluio ou combinação de preços;
- Identificação das variações de preços praticados nas diferentes regiões país, e etc.

Os dados gerados durante o processamento sugerido também podem ser integrados com outras bases de dados (dados governamentais ou dados obtidos da internet) a fim de se ampliar os tipos de análises a serem feitos, como por exemplo, utilizar a bases de dados de Notas Fiscais da Receita Federal do Brasil para verificar se as empresas estão vendendo para o governo com preços compatíveis com os que elas praticam no mercado privado.

Logo, as possibilidades de aplicações dos resultados obtidos com os procedimentos propostos nessa dissertação são inúmeras, ficando elas limitadas apenas pelas necessidades e criatividade dos analistas de dados que se propuserem a desenvolver estudos com tais informações.

Capítulo 6 – Conclusão

Este capítulo faz a conclusão da dissertação e tem como objetivo apresentar as principais contribuições desse estudo, uma análise da solução proposta e suas limitações, as possibilidades de trabalhos futuros e algumas considerações finais.

6.1 Contribuições

A principal contribuição dessa dissertação foi a proposta de um método capaz de gerar regras de identificação de produtos a partir de descrições textuais de compras, porém, outras contribuições intermediárias também resultaram dessa pesquisa, como a proposta de um algoritmo de geração de frases, a proposta de um algoritmo de poda de sub frases e o desenvolvimento de uma metodologia de avaliação dos resultados.

Além disso, durante o desenvolvimento dos estudos, alguns trabalhos preliminares geraram publicações e também trouxeram novas contribuições para o corpo de conhecimento. Essa seção apresenta uma breve descrição desses artigos, bem como as suas principais contribuições

O primeiro artigo foi publicado na Revista Brasileira de Sistemas de Informação (iSys) e seu título é "DW-CGU: Integração dos Dados do Portal da Transparência do Governo Federal Brasileiro" (PAIVA; REVOREDO; BAIÃO, 2016). Esse trabalho surgiu da necessidade de se estudar as bases de dados do Portal da Transparência, a fim de se verificar os possíveis caminhos para a solução do problema de identificação de produtos nas descrições textuais de compras apresentadas no Portal.

Durante esse estudo, constatou-se que devido ao fato do Portal da Transparência do Governo Federal apresentar informações oriundas de diversas bases de dados distintas, tais informações, mesmo quando tratavam do mesmo assunto, não possuíam nenhuma

ligação explícita, sendo apresentadas de forma isolada, sem qualquer vínculo que favorecesse a interoperabilidade dos dados oriundos de sistemas distintos.

Dessa forma, a questão de pesquisa que esse artigo aborda é como disponibilizar de forma integrada as informações disponíveis no Portais da Transparência do Governo Federal, oriundas dos diversos sistemas coorporativos.

O trabalho propõe uma solução de integração para os dados apresentados no Portal da Transparência do Governo Federal. Essa solução, além de controlar todo o processo de carga dos dados, também implementa os tratamentos necessários a perfeita harmonização dos diferentes conjuntos de dados que alimentam o Portal da Transparência do Governo Federal.

Sendo assim, a principal contribuição desse artigo é a proposta de uma arquitetura capaz de gerenciar todo o processo de integração dos dados de um grande portal coorporativo. Essa arquitetura gerencia questões ligadas à integração em si, assim como os processos de carga e validação dos dados, mantendo todo o histórico de rastreabilidade dos dados carregados em um repositório de controle, o que permite a identificação dos processos de carga e dos arquivos fonte utilizados no carregamento de um registro defeituoso, facilitando assim a manutenibilidade dos dados.

O trabalho também conclui que um projeto de integração de dados envolve muitas outras questões que vão além dos desafios tecnológicos, devendo-se considerar a interpretação semântica, o entendimento dos dados e a padronização das regras de negócio envolvidas nos diversos processos de trabalho; razão essa que impõe o comprometimento de todas as áreas de negócio envolvidas no projeto de unificação dos dados.

O segundo trabalho foi o artigo "Identificação Automática de Produtos e suas Características em Grandes Volumes de Dados Não Estruturados: Uma Proposta para Portais de Transparência Pública" (PAIVA e REVOREDO, 2016c), apresentado no IX Workshop de Teses e Dissertações em Sistemas de Informação. Esse trabalho expôs uma proposta inicial para o projeto dessa dissertação, evidenciando o contexto em que essa pesquisa está inserida, os objetivos definidos, assim como o referencial teórico e trabalhos relacionados, o problema a ser tratado, bem como a hipótese considerada, a proposta de solução e o projeto de avaliação de uma maneira formal, a fim de submetê-la a apreciação da comunidade científica e receber críticas e sugestão para melhorar os resultados a serem obtidos. Dessa forma, esse artigo, serviu de ponto de partida para o desenvolvimento dessa dissertação, e serve como um roteiro para outras pesquisas semelhantes a essa.

O terceiro artigo “Big Data e Transparência: Utilizando Funções de Mapreduce para incrementar a transparência dos Gastos Públicos” (PAIVA e REVOREDO, 2016b), apresentado no XII Simpósio Brasileiro de Sistemas de Informação, propõe uma metodologia de tratamento de informação para a obtenção de visões mais elucidativas sobre os gastos públicos. O objetivo é a aplicação de técnicas de programação paralela baseadas no paradigma de programação mapreduce, utilizando o framework Hadoop (WHITE, 2012), para fazer a identificação de um conjunto pré-determinado de produtos comprados pela Administração Pública, além de propor uma forma de consolidação dessas informações de maneira que permita a fácil visualização de disparidades encontradas no grande volume de dados apresentados. Esse mecanismo propicia uma melhor avaliação dos gastos públicos, permitindo com que o cidadão possa entender melhor como o governo está empregando seu dinheiro.

O intuito desse experimento no contexto do trabalho ora apresentado, foi testar a viabilidade de uma solução de mineração de texto para o volume e formato dos dados de portais de transparência, pois de nada adiantaria o desenvolvimento da técnica sem a possibilidade de empregá-la no conjunto de dados para o qual ela se destina.

A proposta apresentada foi testada para um conjunto de 15 produtos com a base de empenhos extraídos do Portal da transparência do Governo Federal. Esse conjunto de dados foi composto por todos os empenhos emitidos no ano de 2014.

Como resultado, obteve-se a relação de compras desses 15 produtos, bem como as principais métricas associadas com os preços unitários de tais produtos. A metodologia proposta também permitiu a identificação de compras superfaturadas que seriam impossíveis de serem identificadas sem a utilização de técnicas de processamento intensivo de dados.

Já o quarto artigo “Geração Automática de Regras de Identificação de Produtos em Descrições Textuais de Compras Governamentais” (PAIVA e REVOREDO, 2016a), apresentado no XIII Encontro Nacional de Inteligência Artificial e Computacional, tem o objetivo de propor uma metodologia capaz gerar regras de identificação de produtos de forma automatizada, em descrições de compras em portais de transparência, através da utilização de uma técnica de mineração de frases.

A metodologia sugerida faz uso de programação paralela, e roda em clusters de computadores, através da utilização do paradigma de programação mapreduce [Dean and Ghemawat 2008] e da infraestrutura do sistema de processamento Apache Spark [Zaharia et al. 2010], o que a torna escalável e permite que aumentos expressivos no volume de

dados a ser analisado possam ser compensados pela inclusão de novas máquinas no cluster, sem o comprometimento da performance.

A proposta apresentada nesse quarto artigo foi uma primeira versão para a solução do problema endereçado nessa dissertação, sendo que, a versão final, que foi explicada em detalhes no capítulo 3 apresentou alguns melhoramentos em relação a sua versão inicial.

6.2 Análise da Solução Proposta

Essa dissertação propõe um método de descoberta de conhecimento em texto voltado para dados de descrições textuais de compras apresentadas em portais de transparência. Tal método faz a geração de regras de identificação de produtos por meio da aplicação de um processo de mineração de frases composto de quatro etapas: geração de frases candidatas, filtragem de frases frequentes, poda de sub frases e geração de regras. Sendo que, antes desse processo de mineração de texto propriamente dito, as descrições de compras passam por uma etapa de pré-processamento, que tem o objetivo de preparar o conjunto de dados textuais para o processo de mineração de frases.

O método proposto utiliza um processo de descoberta de conhecimento em texto que recebe como entrada um conjunto de descrições textuais de compras, e oferece como saída um conjunto de regras de identificação de produtos, utilizando três parâmetros de referência: tamanhos mínimos e máximo de frase e suporte mínimo. Opcionalmente, dependendo da disponibilidade de pessoal, pode-se executar uma tarefa adicional, denominada refinamento de regras. Nessa atividade opcional, especialistas podem validar as regras geradas, bem como, adaptá-las de acordo com os propósitos desejados.

A solução proposta foi avaliada utilizando-se os dados de itens de empenho do Portal da Transparência do Governo federal, referentes ao período de um ano. Nessa avaliação executou-se 6 experimentos. Nesses experimentos, variou-se os parâmetros utilizados para a geração das regras: tamanhos máximo e mínimo de frases, e suporte mínimo, a fim de se analisar o comportamento da solução proposta com diferentes configurações.

Pela análise dos experimentos, pôde-se concluir, para a base analisada, que:

- A variação dos parâmetros de entrada não acarreta grandes diferenças nos tempos de geração das regras,

- Para os mesmos parâmetros de tamanhos máximo e mínimo de frases, o valor do suporte está inversamente relacionado com a quantidade de regras geradas, ou seja, quanto menor o suporte, maior é o número de regras geradas.
- Para um mesmo valor de suporte, percebe-se que o número de regras geradas é maior quando os tamanhos das frases são menores.

Com relação a qualidade das regras geradas, percebeu-se que a medida em que se diminui o valor do suporte, para um mesmo tamanho de frase, a tendência é que a qualidade das frases melhore.

Por outro lado, analisando-se os grupos de experimentos com um mesmo suporte, porém com tamanhos de frases diferentes, apesar de esperar-se que tamanhos de frases maiores produzissem regras de melhores qualidades, os experimentos realizados não foram capazes de provar essa hipótese.

Quanto à análise dos resultados, realizou-se uma avaliação qualitativa dos resultados obtidos. A seleção do conjunto amostral a ser analisado foi feita de forma a identificar aquelas compras com maior probabilidade de terem sido classificadas de forma errônea. Assim, selecionou-se dois registros considerados outliers, para cada regra avaliada, a fim de verificar se aquelas compras realmente se referiam ao produto a que a regra em questão se propõem a identificar.

Nessa análise, não foram identificados produtos classificados de forma errônea. Porém, algumas das regras geradas, ainda que classificando as compras de maneira correta, faziam uma identificação muito genérica. Portanto, dependendo da finalidade de utilização do método de mineração de frases desenvolvido, pode não se atender os objetivos por completo. Exemplos desses tipos de regras foram regras que levavam a produtos como “peça para automóvel” ou “produto perecível”, ou seja, uma maneira muito genérica para se identificar um produto. Por essa razão, o método prevê uma etapa adicional, em que especialistas podem analisar as regras geradas, a fim de eliminá-las ou adaptá-las aos seus propósitos finais.

Dessa forma, verificou-se que o método proposto foi capaz de identificar os produtos especificados nas descrições de compras apresentadas nos portais de transparência pública, possibilitando assim a melhora da usabilidade e do grau de informatividade dos portais de transparência, segundo e terceiro degraus de transparência definidos em (CAPPELLI; LEITE, 2008).

6.3 Limitações

A ausência de um conjunto de dados que contenha as descrições das compras, com a respectiva identificação dos produtos a que cada uma dessas descrições se refere dificultou a realização de uma avaliação mais objetiva para a verificação dos níveis de precisão atingidos pela técnica desenvolvida. Pois, dada a grande quantidade de registros que compõem a base de dados a ser analisada, torna-se inviável a identificação de forma manual de um conjunto de dados amostral que seja numericamente significativo para a avaliação dos resultados.

Por outro lado, qualquer tentativa de se gerar um conjunto de dados de teste de forma automatizada, o que resolveria o problema da grande dimensionalidade do conjunto de dados a ser analisado, poderia utilizar critérios tendenciosos, além do que, esses critérios de identificação também poderem identificar os produtos de forma errônea, o que distorceria os resultados da avalição realizada.

Por tais razões, optou-se por proceder-se uma validação qualitativa dos resultados encontrados, conforme apresentado no Capítulo 4. Tal procedimento não foi capaz de informar um valor exato sobre o grau de precisão atingido pela técnica proposta. Porém, foi capaz de indicar que a técnica se mostrou válida para os propósitos para o qual foi concebida. Ainda, de acordo com os critérios estipulados para a seleção das amostras a serem analisadas qualitativamente, não foram encontrados erros que pudesses refutar a validade da técnica proposta.

6.4 Trabalhos Futuros

Trabalhos futuros, tanto podem ser voltados para aprimorar o método de mineração de texto proposto, quanto podem partir dos resultados obtidos pela aplicação da metodologia desenvolvida, a fim de se realizar novos estudos com os dados originários dos processamentos realizados.

Com relação à possibilidade de melhoramentos do método proposto, pode-se incorporar o procedimento de clusterização, utilizado no Capítulo 4 para avaliar a qualidade das regras, ao processo de descoberta de conhecimento em dados textuais proposto, fazendo com que apenas as regras julgadas como boas (ou seja, aquelas cujo processo de clusterização de seus registros não gere um grande número de clusters com

registros espalhados por todos eles) sejam consideradas na etapa de geração de regras, melhorando-se assim a qualidade das regras geradas, e consequentemente, aprimorando-se os resultados do procedimento como um todo.

Outro melhoramento que pode ser feito ao processo de geração de regras de identificação de itens de compras é a aplicação de análise de similaridades entre os antecedentes das regras a serem geradas, a fim de se eliminar regras cujos antecedentes possuam similaridades superiores a um determinado valor (passado como parâmetro), quando comparada com outros antecedentes de regras geradas. O procedimento atual faz uma análise prévia dessas similaridades para propor consequentes (rótulos) alternativos para as regras. Porém, a decisão em se usar ou não, esses rótulos alternativos, fica a cargo dos especialistas, durante as análises que são feitas na etapa de refinamento de regras. Logo, um possível melhoramento seria a incorporação dessa atividade ao processo de descoberta de conhecimento, aumentando-se assim o grau de automatização do processo.

O procedimento de mineração de frases apresentado nessa dissertação também pode ser adaptado para ser utilizado em outros contextos. Por exemplo, esse procedimento pode ser ajustado para identificar as frases mais frequentes em um determinado conjunto de dados textuais, a fim de se levantar quais os tópicos mais relevantes nesses corpus de textos. Tal atividade poderia ser utilizada em análises de textos publicados em redes sociais, blogs, jornais e etc.

Quanto aos estudos que podem ser desenvolvidos a partir dos resultados obtidos pela aplicação das técnicas aqui expostas, existem diversas possibilidades de pesquisas. O processo sugerido consegue obter dados estruturados a partir de um conjunto de dados textuais, ou seja, as descrições textuais de compras são associadas a variáveis categóricas que especificam a que produtos essas descrições se referem.

Dessa forma, pode-se utilizar essa nova variável categórica com os demais dados estruturados presentes no banco de dados do portal da transparência, a fim de se aplicar técnicas de mineração de dados para se extrair conhecimentos implícitos a respeito das atividades governamentais que são apresentadas nos portais de transparência pública.

6.5 Considerações Finais

Uma das motivações para a escolha do tema dessa dissertação foi a necessidade de se dar maior efetividade aos dados que são publicados diariamente nos portais de transparência pública, possibilitando assim que novos conhecimentos fossem obtidos a

partir dos dados que vêm dos sistemas que fornecem informações para esses tipos de portais. Dessa forma, além das contribuições científicas apresentadas, o presente trabalho também traz contribuições tecnológicas e sociais.

Quanto à contribuição tecnológica, foi desenvolvido um framework que pode ser empregado em portais de transparência pública, sendo necessário, para isso, que o portal em questão apresente as descrições textuais de compras, e que haja condições de se implantar um cluster com o Apache Spark instalado, sendo que a configuração do cluster a ser utilizado dependerá do volume de dados tratado pelo portal em questão.

A ideia inicial é que essa solução seja incorporada à infraestrutura de atualização diária do Portal da Transparência do Governo Federal, que é mantido pelo Ministério da Transparência, Fiscalização e Controladoria-Geral da União, sendo que, essa implantação está condicionada a outros limitantes como questões de layout para apresentação das novas informações tratadas e aquisição de hardware necessário a implantação da solução.

Com relação à contribuição social, espera-se melhorar a qualidade das informações apresentadas nos portais de transparência pública, permitindo-se dessa forma, que o cidadão tenha melhores condições de acompanhar e monitorar as atividades governamentais, contribuindo-se assim com o aperfeiçoamento da democracia praticada no país.

Referências

AGUNE, ROBETO MEIZU; GREGORIO FILHO, ALVARO SANTOS; BOLLIGER, SERGIO PINTO. "Governo aberto SP: disponibilização de bases de dados e informações em formato aberto". *III Congresso Consad de Gestão Pública*, Brasília - DF , 2010.

AIZAWA, AKIKO. "An information-theoretic perspective of tf–idf measures". *Information Processing & Management* v. 39, n. 1, p. 45–65 , 2003.

ALEXANDRINO, MARCELO; PAULO, VICENTE. *Direito Administrativo*. 11a ed. Niterói-RJ: Impetus, 2006. 12 v. 85-7626-184-7.

ALTINEL, BERNA; GANIZ, MURAT CAN; DIRI, BANU. "A corpus-based semantic kernel for text classification by using meaning values of terms". *Engineering Applications of Artificial Intelligence* v. 43, p. 54–66, 2015.

AZEVEDO, ANA ISABEL ROJÃO LOURENÇO; SANTOS, MANUEL FILIPE. "Kdd, semma crisp-dm: a parallel overviee". *IADS-DM*, 2008.

BÄCKLUND, HENRIK; HEDBLOM, ANDERS; NEIJMAN, NIKLAS. "A density-based spatial clustering of application with noise,". *Data Mining TNM033* p. 11–30, 2011.

BATISTA, AUGUSTO HERRANN; SILVA, NITAI BEZERRA DA; MIRANDA, CHRISTIAN MORYAH CONTIERO. "Infraestrutura nacional de dados abertos". *IV Congresso Consad de Gestão Pública*, Brasília - DF, 2013.

BRASIL. Lei Complementar nº 131 de 27 de maio de 2009. Disponibilização em tempo real de Informações. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 18 abr. 2015. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/lcp/lcp131.htm. Acesso em: 18 abr. 2015.

_____. Lei nº 4320, de 18 março de 1964. Lei 4320. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 17 mar. 1964. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l4320.htm. Acesso em: 24 abr. 2015.

_____. Lei nº 8666, de 21 junho de 1993. Lei 8666. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 22 jun. 1993. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/L8666cons.htm. Acesso em: 24 abr. 2015.

_____. Lei nº 12.527, de 18 de novembro de 2011. Lei de Acesso à Informação. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 23 abr. 2015. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 23 abr. 2015.

_____, Constituição (1998). Constituição da República Federativa do Brasil. Brasília, DF: Senado 1998 Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 23 abr. 2015.

BREIMAN, LEO. "Random forests". *Machine learning* v. 45, n. 1, p. 5–32, 2001.

CAPPELLI, CLAUDIA, 2009. "Uma abordagem para transparência em processos organizacionais utilizando aspectos". Tese de Doutorado - Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC-Rio, Rio de Janeiro-RJ.

CAPPELLI, CLAUDIA; LEITE, JCSP. "Transparência de processos organizacionais". *II Simpósio Internacional de Transparência nos Negócios*, Universidade Federal Fluminense, LATEC, Niterói, RJ, Brasil, 2008.

CAPPELLI, C. A.; LEITE, JCSP; ARAÚJO, RENATA MENDES. "A importância de um Modelo de Estágios para avaliar Transparência". *Revista TCMRJ*, setembro n. 45, p. 97, 2010.

CARVALHO, ROMMEL; PAIVA, EDUARDO; ROCHA, HENRIQUE et al. "Methodology for Creating the Brazilian Government Reference Price Database". *X Encontro Nacional de Inteligência Artificial e Computacional*, 2013, Fortaleza-CE, 2013. Disponível em: http://www.lbd.dcc.ufmg.br/colecoes/eniac/2013/0033.pdf.

CARVALHO, ROMMEL; PAIVA, EDUARDO; ROCHA, HENRIQUE et al. "Using Clustering and Text Mining to Create a Reference Price Database". *Learning and NonLinear Models* v. 12, p. 38–52, 2014a.

CARVALHO, ROMMEL; SALES, LEONARDO; ROCHA, HENRIQUE et al. "Using Bayesian Networks to Identify and Prevent Split Purchases in Brazil". *In: BMA@ UAI.* 2014b. p. 70-78.

CGU, CONTROLADORIA-GERAL. "Manual da Lei de Acesso à Informação para Estados e Municípios". 1a edição. Brasília: CGU, abr, 2013.

CGU, CONTROLADORIA GERAL DA UNIÃO. "Portal da Transparência nos Recursos Públicos Federais". Disponível em: http://transparencia.gov.br. Acesso em: 18 abr. 2015.

DA LÍNGUA PORTUGUESA, "Dicionário Priberam. Lisboa: Priberam". Disponível em: http://www.priberam.pt/dlpo/default.aspx, 2016.

DEAN, JEFFREY; GHEMAWAT, SANJAY. "MapReduce: simplified data processing on large clusters". *Communications of the ACM* v. 51, n. 1, p. 107–113, 2008.

DEAN, JEFFREY; GHEMAWAT, SANJAY. “Mapreduce: Simplified data processing on large clusters”, *osdi’04: Sixth symposium on operating system design and implementation*, San Francisco, ca, december, 2004.

DERCZYNSKI, LEON; MAYNNARD, DIANA; RIZZO, GIUSEPPE et al. “Analysis of named entity recognition and linking for tweets”. *Information Processing & Management* v. 51, n. 2, p. 32–49 , 2015.

DIETRICH, DANIEL; GRAY, JONATHAN; MCNAMARA, TIM et al. “Open data handbook”. Open Kmowledge Foundation, 2012. .

EAVES, DAVID. “The three laws of open government data”. *Eaves. ca* v. 30, 2009.

EL-KISHKY, AHMED; SONG, YANGLEI; WANG, CHI et al. “Scalable topical phrase mining from text corpora”. *Proceedings of the VLDB Endowment* v. 8, n. 3, p. 305–316, 2014.

GOLDSCHMIDT, RONALDO; BEZERRA, EDUARDO. “Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações”. Elsevier Brasil, 2015. .85-352-7823-0.

GUPTA, RAJEEV; GUPTA, HIMANSHU; MOHANIA, MUKESH. “Cloud computing and big data analytics: what is new from databases perspective? Big Data Analytics”. Springer, 2012. p. 42–61. 3-642-35541-2.

HAHN, UDO; MANI, INDERJEET. “The challenges of automatic summarization”. *IEEE Computer* v. 33, n. 11, p. 29–36 , 2000.

HARMAN, DONNA. “How effective is suffixing?”. *Journal of the American Society for Information Science* v. 42, n. 1, p. 7 , 1991.

HEARST, MARTI A. “Untangling text data mining”. *Association for Computational Linguistics*, 1999. p.3–10.

HEARST, MARTI A; DUMAIS; OSUNA “Support vector machines”. *IEEE Intelligent Systems and their Applications* v. 13, n. 4, p. 18–28, 1998.

HU, XIA; LIU, HUAN. “Text analytics in social media. Mining text data”. *Springer*, 2012. p. 385–414. 1-4614-3222-7.

HUFFMAN, DAVID A. “A method for the construction of minimum-redundancy codes”. *Proceedings of the IRE* v. 40, n. 9, p. 1098–1101, 1952.

JAIN, ANIL K.; MURTY, M. NARASIMHA; FLYNN, PATRICK J. “Data clustering: a review”. *ACM computing surveys (CSUR)* v. 31, n. 3, p. 264–323, 1999.

JARDIM, JOSÉ MARIA. “A construção do e-gov no Brasil: configurações político-informacionais”. *Encontro Nacional da Ciência da Informação* v. 5, 2004.

LANGLEY, PAT; IBA, WAYNE; THOMPSON, KEVIN. “An analysis of Bayesian classifiers”. Tenth National Conference on Artificial Intelligence, San Jose, CA, 1992.

LIN, JIMMY; DYER, CHRIS. “Data-intensive text processing with MapReduce”. *Synthesis Lectures on Human Language Technologies* v. 3, n. 1, p. 1–177 , 2010.

LIU, JIALU; SHANG, JINGBO; WANG, CHI et al. “Mining Quality Phrases from Massive Text Corpora”. *ACM*, 2015. p.1729–1744. 1-4503-2758-3.

LIU, MING; CHEN, LEI; LIU, BINGQUAN et al. “VRCA: a clustering algorithm for massive amount of texts”. *AAAI Press*, 2015. p.2355–2361. 1-57735-738-8.

LOPES, MARIA CÉLIA SANTOS. “Mineração de dados textuais utilizando técnicas de Clustering para o idioma português”. Tese (Doutorado em Engenharia Civil)-Faculdade de Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2004.

LOVINS, JULIE B. “Development of a stemming algorithm”. *MIT Information Processing Group, Electronic Systems Laboratory Cambridge*, 1968.

LUHN, HANS PETER. “The automatic creation of literature abstracts”. *IBM Journal of research and development* v. 2, n. 2, p. 159–165, 1958.

MAIA, PATRÍCIA HELENA. “Aplicação de Técnicas de Mineração de Textos para Classificação de Documentos: Um Estudo da Automatização da Triagem de Denúncias na CGU”. Brasília: UNB, 2015. Dissertação (Mestrado em Ciência da Computação) - Programa de Pós-graduação Aplicada em Computação, Universidade de Brasília.

MARQUES, MARCELO. “Administração Pública: uma abordagem prática”. Editora Ferreira, 2010. 85-7842-002-0.

MARZAGÃO, THIAGO. “Using SVM to pre-classify government purchases”. *arXiv preprint arXiv*:1601.02680 , 2015.

PAIVA, EDUARDO; REVOREDO, KATE. “Geração Automática de Regras de Identificação de Produtos em Descrições Textuais de Compras Governamentais”. *XIII Encontro Nacional de Inteligência Artificial e Computacional*, Recife-PE , 2016a.

PAIVA, EDUARDO; REVOREDO, KATE. “Big Data e Transparência: Utilizando Funções de Mapreduce para incrementar a transparência dos Gastos Públicos”. *XII Simpósio Brasileiro de Sistemas de Informação*, Florianopolis-SC , 2016b.

PAIVA, EDUARDO; REVOREDO, KATE. “Identificação Automática de Produtos e suas Características em Grandes Volumes de Dados Não Estruturados: Uma Proposta para Portais de Transparência Pública”. *IX Workshop de Teses e Dissertações em Sistemas de Informação*, Florienopolis-SC, 2016c.

PAIVA, EDUARDO; REVOREDO, KATE; BAIÃO, FERNANDA ARAUJO. “DW-CGU: Integração dos Dados do Portal da Transparência do Governo Federal Brasileiro”. *iSys-Revista Brasileira de Sistemas de Informação* v. 9, n. 1, p. 6–32, 2016.

PENTREATH, NICK. “Machine Learning with Spark”. Packt Publishing Ltd, 2015. 1-78328-852-3.

PORTER, MARTIN F. “An algorithm for suffix stripping”. *Program: electronic library andinformation systems* v. 14, n. 3, p. 130–137, 1980.

QUINLAN, J. ROSS. “Induction of decision trees”. *Machine learning* v. 1, n. 1, p. 81–106, 1986.

RAMASWAMY, SRIDHAR; RASTOGI, RAJEEV; SHIM, KYUSEOK. “Efficient algorithms for mining outliers from large data sets”. *ACM*, 2000. p.427–438. 1-58113-217-4. .

RATCLIFF, JOHN W.; METZENER, DAVID E. “Pattern-matching-the gestalt approach”. *Dr Dobbs Journal* v. 13, n. 7, p. 46- , 1988.

REN, XIANG; EL-KISHKY, AHMED; WANG, CHI et al. “Clustype: Effective entity recognition and typing by relation phrase-based clustering”. *ACM*, 2015. p.995–1004. 1-4503-3664-7.

SHEARER, COLIN. “The CRISP-DM model: the new blueprint for data mining”. *Journal of data warehousing* v. 5, n. 4, p. 13–22 , 2000.

TAURION, CEZAR. “Big data”. Brasport, 2013. 85-7452-608-8.

WEISS, SHOLOM M.; INDURKHYA, NITIN; ZHANG, TONG et al. “Text mining: predictive methods for analyzing unstructured information”. *Springer Science & Business Media*, 2010. .0-387-34555-8.

WHITE, TOM. “Hadoop: The definitive guide”. O’Reilly Media, Inc., 2012. 1-4493-1152-0.

YANG, YIMING; PEDERSEN, JAN O. “A comparative study on feature selection in text categorization”. *In: Proceedings of the Fourteenth International Conference on Machine Learning (ICML’97)*, 1997. p.412–420.

YU, YANG; WANG, XIAO. “World Cup 2014 in the Twitter World: A big data analysis of sentiments in US sports fans’ tweets”. *Computers in Human Behavior* v. 48, p. 392–400 , 2015.

ZAHARIA, MATEI; CHOWDHURY, MOSHARAF; DAS, TATHAGATA et al. “Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing”. *USENIX Association*, 2012. p.2–2.

ZAHARIA, MATEI; CHOWDHURY, MOSHAARAF; FRANKLIN, MICHAEL J. et al. “Spark: cluster computing with working sets”. *HotCloud* v. 10, p. 10–10, 2010.

Apêndice I – Análise de Outliers

Durante a validação dos resultados obtidos pela a aplicação das regras geradas, apresentada no capítulo 4, foram identificados 2 outiliers por regra analisada. Esse apêndice tem o objetivo de monstrar as análises individuais feitas para cada um dos outliers levantados e apresentados na Tabela 10 da seção 4.2.

- **Outliers da Regra R_2**

A Tabela A.1 mostra os outliers levantados para a regra R_2, que identifica compras de óleo diesel, enquanto que a Figura A.1 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de diesel.

Tabela A.1: Outliers da regra R_2

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800889	1	R_2	Diesel	100.000.000,00	Litro	Petrobras
2015NE800586	1	R_2	Diesel	87.279.000,00	Litro	Petrobras

Empenho: 2015NE800889 - Item: 1

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
1 - COMBUSTIVEIS E LUBRIFICANTES AUTOMOTIVOS	0,00001	100.000.000,00	1.000,00	0000000000,00001 Litro ÓLEO DIESEL ÓLEO DIESEL, NOME OLEO DIESEL - S10 MARCA: PETROBRAS ITEM DO PROCESSO: 00004 ITEM DE MATERIAL: 000016993

Empenho: 2015NE800586 - Item: 1

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
1 - COMBUSTIVEIS E LUBRIFICANTES AUTOMOTIVOS	0,00001	87.279.000,00	872,79	0000000000,00001 Litro ÓLEO DIESEL ÓLEO DIESEL, NOME OLEO DIESEL - S10 MARCA: PETROBRAS ITEM DO PROCESSO: 00004 ITEM DE MATERIAL: 000016993

Figure A.1 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_2

- **Outliers da Regra R_3**

A Tabela A.2 mostra os outliers levantados para a regra R_3, que identifica compras de água mineral, enquanto que a Figura A.2 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de água mineral.

Tabela A.2: Outliers da regra R_3

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800803	1	R_3	Água mineral	68.611,32	Garrafao 20 L	Seiva
2015NE800466	1	R_3	Água mineral	52.602,06	Garrafao 20 L	Seiva

Empenho: 2015NE800803 - Item: 1

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
7 - GENEROS DE ALIMENTACAO	1	61.800,48	61.800,48	0000000001,00000 GALÃO 20,00 L ÁGUA MINERAL, MATERIAL ÁGUA MINERAL, TIPO EMBALAGEM PLÁSTICO POLICARBONATO TRANSPARENTE, GASEIFICAÇÃO SEM GÁS, CARACTERÍSTICAS ADICIONAIS COM TAMPA DE PRESSÃO/LACRE/ENVASADO MECANICAMENTE/, NORMAS TÉCNICAS CONFORME PORTARIA DE CORRELATOS DO MINISTÉRIO SAÚ- MARCA: Hydrate ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000304461

Empenho: 2015NE801154 - Item: 1

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
7 - GENEROS DE ALIMENTACAO	1	52.602,06	52.602,06	0000000001,00000 GALÃO 20,00 L ÁGUA MINERAL, MATERIAL ÁGUA MINERAL, TIPO EMBALAGEM PLÁSTICO, GASEIFICAÇÃO SEMGÁS MARCA: calogi ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000217773

Figure A.2 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_3

- **Outliers da Regra R_4**

A Tabela A.3 mostra os outliers levantados para a regra R_4, que identifica compras de banana, enquanto que a Figura A.3 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de banana.

Tabela A.3: Outliers da regra R_4

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE804597	1	R_4	Banana	16.388,00	QUILOGRAMA	CEASA
2015NE803169	1	R_4	Banana	15.000,00	QUILOGRAMA	CEASA

Figure A.3 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_4

- **Outliers da Regra R_6**

A Tabela A.4 mostra os outliers levantados para a regra R_6, que identifica compras de produtos perecíveis, enquanto que a Figura A.4 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de produtos perecíveis.

Tabela A.4: Outliers da regra R_6

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE801459	12	R_6	Produto perecível	9,65	Kg	frigolaste
2015NE801459	8	R_6	Produto perecível	21,46	Kg	sabadini

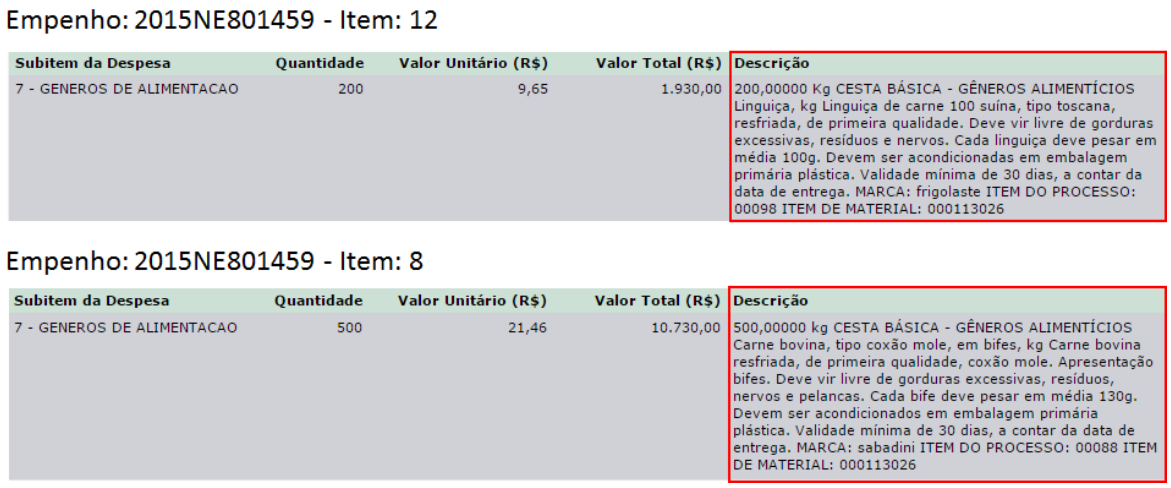


Figure A.4 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_6

- **Outliers da Regra R_10**

A Tabela A.5 mostra os outliers levantados para a regra R_10, que identifica compras de gás liquefeito do petróleo (GLP), enquanto que a Figura A.10 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de gás liquefeito do petróleo (GLP).

Tabela A.5: Outliers da regra R_10

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE801077	1	R_10	Gás liquefeito - glp	63.982,92	KG	GASBALL
2015NE801613	1	R_10	Gás liquefeito - glp	120.000,00	KG	GASBALL

Empenho: 2015NE801077 - Item: 1

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
4 - GAS E OUTROS MATERIAIS ENGARRAFADOS	1	63.982,92	63.982,92	1,00000 KG GÁS LIQUEFEITO DE PETRÓLEO - GLP GÁS LIQUEFEITO DE PETRÓLEO (GLP) À GRANEL MARCA: GASBALL ITEM DO PROCESSO: 00002 ITEM DE MATERIAL: 000047678

Empenho: 2015NE801613 - Item: 1

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
4 - GAS E OUTROS MATERIAIS ENGARRAFADOS	1	120.000,00	120.000,00	0000000001,00000 KG GÁS LIQUEFEITO DE PETRÓLEO - GLP Gás Liquefeito de Petróleo (GLP), a Granel. MARCA: GASBALL ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000047678

Figure A.5 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_10

- **Outliers da Regra R_11**

A Tabela A.6 mostra os outliers levantados para a regra R_11, que identifica compras de água mineral, enquanto que a Figura A.6 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de água mineral.

Tabela A.6: Outliers da regra R_11

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	Seq	Regra	Produto	Valor em R$	Unidade de Medida	Marca
2015NE800466	1	R_11	Água mineral	52.602,06	GALAO 20,00 L	calogi
2015NE800899	1	R_11	Água mineral	61.800,48	GALAO 20,00 L	Hydrate

Empenho: 2015NE800466 - Item: 1

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
7 - GENEROS DE ALIMENTACAO	1	52.602,06	52.602,06	0000000001,00000 GALÃO 20,00 L ÁGUA MINERAL, MATERIAL ÁGUA MINERAL, TIPO EMBALAGEM PLÁSTICO, GASEIFICAÇÃO SEMGÁS MARCA: calogi ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000217773

Empenho: 2015NE800899 - Item: 1

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
7 - GENEROS DE ALIMENTACAO	1	61.800,48	61.800,48	0000000001,00000 GALÃO 20,00 L ÁGUA MINERAL, MATERIAL ÁGUA MINERAL, TIPO EMBALAGEM PLÁSTICO POLICARBONATO TRANSPARENTE, GASEIFICAÇÃO SEM GÁS, CARACTERÍSTICAS ADICIONAIS COM TAMPA DE PRESSÃO/LACRE/ENVASADO MECANICAMENTE/, NORMAS TÉCNICAS CONFORME PORTARIA DE CORRELATOS DO MINISTÉRIO SAÚ- MARCA: Hydrate ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000304461

Figure A.6 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_11

- **Outliers da Regra R_12**

A Tabela A.7 mostra os outliers levantados para a regra R_12, que identifica compras de bequer de vidro, enquanto que a Figura A.7 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de bequer de vidro.

Tabela A.7: Outliers da regra R_12

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800375	1	R_12	Bequer de vidro	350,00	UNIDADE	Leica Biosystems
2015NE806768	1	R_12	Bequer de vidro	900,00	UNIDADE	VELP

Figure A.7 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_12

- **Outliers da Regra R_14**

A Tabela A.8 mostra os outliers levantados para a regra R_14, que identifica compras de proveta de vidro, enquanto que a Figura A.8 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de proveta de vidro.

Tabela A.8: Outliers da regra R_14

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800075	2	R_14	Proveta de vidro	7,00	UNIDADE	rav
2015NE800736	1	R_14	Proveta de vidro	4,60	UNIDADE	Uniglass

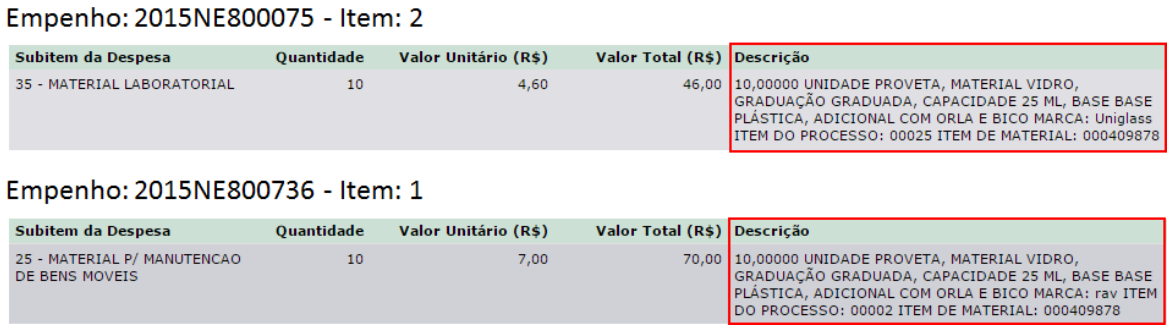


Figure A.8 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_14

- **Outliers da Regra R_15**

A Tabela A.9 mostra os outliers levantados para a regra R_15, que identifica compras de gasolina comum, enquanto que a Figura A.9 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra gasolina comum.

Tabela A.9: Outliers da regra R_15

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800068	4	R_15	Gasolina Comum	553.834,29	Litros	xxxxxxxxxx
2015NE800809	1	R_15	Gasolina Comum	7.157.000,00	Litro	SHELL

Empenho: 2015NE800068 - Item: 4

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
1 - COMBUSTIVEIS E LUBRIFICANTES AUTOMOTIVOS	1	553.834,29	553.834,29	1,00000 Litros GASOLINA COMUM GASOLINA COMUM, NOME GASOLINA - COMBUSTIVEL VEICULO MARCA: xxxxxxxxxxxxxxxx ITEM DO PROCESSO: 00004 ITEM DE MATERIAL: 000016950

Empenho: 2015NE800809 - Item: 1

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
1 - COMBUSTIVEIS E LUBRIFICANTES AUTOMOTIVOS	0,00001	7.157.000,00	71,57	0000000000,00001 litro GASOLINA COMUM GASOLINA COMUM, NOME GASOLINA - COMBUSTIVEL VEICULO MARCA: SHELL ITEM DO PROCESSO: 00001 ITEM DE MATERIAL: 000016950

Figure A.9 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_15

- **Outliers da Regra R_16**

A Tabela A.10 mostra os outliers levantados para a regra R_16, que identifica compras de balão volumétrico, enquanto que a Figura A.10 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de balão volumétrico.

Tabela A.10: Outliers da regra R_16

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800588	1	R_16	Balão volumétrico para laboratório	530,00	UNIDADE	-
2015NE802470	1	R_16	Balão volumétrico para laboratório	615,88	UNIDADE	DIOGOLAB

Empenho: 2015NE800588 - Item: 1

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
35 - MATERIAL LABORATORIAL	3	530,00	1.590,00	3,00000 UNIDADE BALÃO LABORATÓRIO, TIPO USO VOLUMÉTRICO, TIPO FUNDO FUNDO CHATO, MATERIAL VIDRO, CAPACIDADE 1000 ML, ACESSÓRIOS ROLHA DE VIDRO MARCA: - ITEM DO PROCESSO: 00005 ITEM DE MATERIAL: 000409239

Empenho: 2015NE802470 - Item: 1

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
35 - MATERIAL LABORATORIAL	1	615,88	615,88	1,00000 UNIDADE BALÃO LABORATÓRIO, TIPO USO VOLUMÉTRICO, TIPO FUNDO FUNDO CHATO, MATERIAL VIDRO, CAPACIDADE 5000 ML MARCA: DIOGOLAB ITEM DO PROCESSO: 00013 ITEM DE MATERIAL: 000409667

Figure A.10 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_16

- **Outliers da Regra R_19**

A Tabela A.11 mostra os outliers levantados para a regra R_19, que identifica compras de resistor de filme metálico, enquanto que a Figura A.11 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de resistor de filme metálico.

Tabela A.11: Outliers da regra R_19

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800001	16	R_19	Resistor filme metálico	0,02	UNIDADE	RohmRohm
2015NE800001	34	R_19	Resistor filme metálico	0,02	UNIDADE	RohmRohm

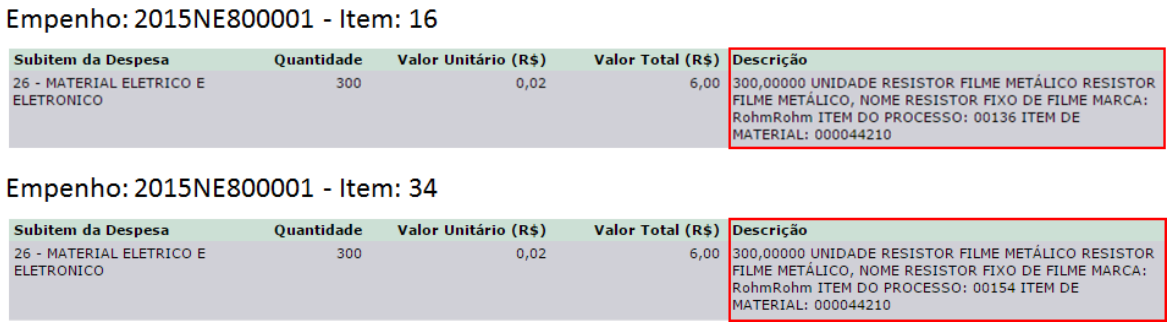


Figure A.11 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_19

- **Outliers da Regra R_22**

A Tabela A.12 mostra os outliers levantados para a regra R_22, que identifica compras de caneta esferográfica, enquanto que a Figura A.12 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de caneta esferográfica.

Tabela A.12: Outliers da regra R_22

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800042	7	R_22	Caneta esferográfica	1,10	CAIXA 1.200,00 UM	esferografica
2015NE800006	2	R_22	Caneta esferográfica	135,10	CAIXA 12,00 UN	slider

Figure A.12 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_22

- **Outliers da Regra R_23**

A Tabela A.13 mostra os outliers levantados para a regra R_23, que identifica compras de álcool etílico hidratado combustível, enquanto que a Figura A.13 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de álcool etílico hidratado combustível.

Tabela A.13: Outliers da regra R_23

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800089	2	R_23	Álcool etílico hidratado combustível	78.348,81	LITRO	xxxxxxxxxx
2015NE800549	3	R_23	Álcool etílico hidratado combustível	129.552,00	LITRO	IPIRANGA

Empenho: 2015NE800089 - Item: 2

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
1 - COMBUSTIVEIS E LUBRIFICANTES AUTOMOTIVOS	1	78.348,81	78.348,81	1,00000 LITRO ÁLCOOL ETÍLICO HIDRATADO COMBUSTÍVEL ÁLCOOL ETÍLICO HIDRATADO COMBUSTÍVEL, NOME ÁLCOOL ETÍLICO HIDRATADO COMBUSTÍVEL MARCA: xxxxxxxxxxxx ITEM DO PROCESSO: 00002 ITEM DE MATERIAL: 000150371

Empenho: 2015NE800549 - Item: 3

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
1 - COMBUSTIVEIS E LUBRIFICANTES AUTOMOTIVOS	1	129.552,00	129.552,00	1,00000 LITRO ÁLCOOL ETÍLICO HIDRATADO COMBUSTÍVEL ÁLCOOL ETÍLICO HIDRATADO COMBUSTÍVEL, NOME ÁLCOOL ETÍLICO HIDRATADO COMBUSTÍVEL MARCA: IPIRANGA ITEM DO PROCESSO: 00004 ITEM DE MATERIAL: 000150371

Figure A.13 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_23

- **Outliers da Regra R_24**

A Tabela A.14 mostra os outliers levantados para a regra R_24, que identifica compras de álcool anidro combustível, enquanto que a Figura A.14 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de álcool anidro combustível.

Tabela A.14: Outliers da regra R_24

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800068	3	R_24	Álcool anidro combustível	553.834,29	Litros	xxxxxxxxxx
2015NE800876	3	R_24	Álcool anidro combustível	148.823,93	Litros	NACIONAL

Empenho: 2015NE800068 - Item: 3

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
1 - COMBUSTIVEIS E LUBRIFICANTES AUTOMOTIVOS	1	553.834,29	553.834,29	1,00000 Litros ÁLCOOL ANIDRO COMBUSTÍVEL ÁLCOOL ANIDRO COMBUSTÍVEL, NOME ALCOOL - COMBUSTIVEL VEICULO MARCA: xxxxxxxxxxxxxxxxxxx ITEM DO PROCESSO: 00003 ITEM DE MATERIAL: 000047

Empenho: 2015NE800876 - Item: 3

Subitem da Despesa	Quantidade	Valor Unitário (R$)	Valor Total (R$)	Descrição
1 - COMBUSTIVEIS E LUBRIFICANTES AUTOMOTIVOS	1	148.823,93	148.823,93	1,00000 Litros ÁLCOOL ANIDRO COMBUSTÍVEL ÁLCOOL ANIDRO COMBUSTÍVEL, NOME ALCOOL - COMBUSTIVEL VEICULO, CONFORME ESPECIFICAÇÕES DO TERMO DE REFERÊNCIA. MARCA: NACIONAL ITEM DO PROCESSO: 00003 ITEM DE MATERIAL: 000047627

Figure A.14 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_24

- **Outliers da Regra R_25**

A Tabela A.15 mostra os outliers levantados para a regra R_25, que identifica compras de peça para automóvel, enquanto que a Figura A.15 monstra os recortes das telas do portal da transparência que trazem esses registros e comprovam que as referidas descrições realmente se referem a compra de peça para automóvel.

Tabela A.15: Outliers da regra R_25

Identificação da Compra		Regra aplicada		Atributos Considerados		
CodEmpenho	**Seq**	**Regra**	**Produto**	**Valor em R$**	**Unidade de Medida**	**Marca**
2015NE800035	1	R_25	Peça para automóvel	560.000.000,00	MENSAL	Conforme Edital
2015NE800134	1	R_25	Peça para automóvel	1.000.000.000,00	UNIDADE - PECAS	ORIGINAL

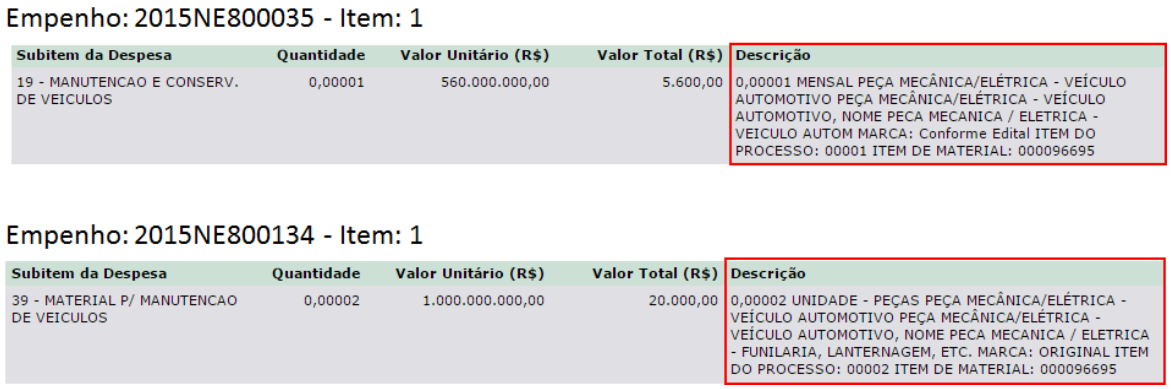


Figure A.15 – Recortes das Telas do Portal das Transparência para Registros Considerados Outliers da Regra R_25