Renato de Oliveira Caldas Madeira

Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais

Rio de Janeiro

2015

Renato de Oliveira Caldas Madeira

Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais

Dissertação apresentada à Escola de Matemática Aplicada da Fundação Getúlio Vargas, para a obtenção do Título de Mestre em Ciências, na área de Modelagem Matemática da Informação.

Orientador: Renato Rocha Souza.

Rio de Janeiro

2015

Madeira, Renato de Oliveira Caldas
Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais / Renato de Oliveira Caldas Madeira. – 2015.
66 f.

Dissertação (mestrado) – Fundação Getulio Vargas, Escola de Matemática Aplicada.
Orientador: Renato Rocha Souza.
Inclui bibliografia.

1. Mineração de dados (Computação). 2. Notas fiscais eletrônicas. 3. Fraude. I. Souza, Renato Rocha. II. Fundação Getulio Vargas. Escola de Matemática Aplicada. III. Título.

CDD – 006.312

RENATO DE OLIVEIRA CALDAS MADEIRA

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE TEXTO NA DETECÇÃO DE DISCREPÂNCIAS EM DOCUMENTOS FISCAIS.

Dissertação apresentada ao Curso de Mestrado em Modelagem Matemática da Informação da Escola de Matemática Aplicada da Fundação Getulio Vargas para obtenção do grau de Mestre em Modelagem Matemática da Informação.

Data da defesa: 29/09/2015.

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Renato Rocha Souza
Orientador (a)

Flávio Codeço Coelho

Pedro Costa Ferreira

Aos meus amores Poliana e Daniel.

Agradecimentos

Gostaria de agradecer à toda a equipe da Escola de Matemática Aplicada (EMAp) pelo suporte durante todo esse período e, em particular, aos professores pela paciência e dedicação.

Agradeço aos meus colegas de turma, pela alegria, amizade e por estarem sempre dispostos a ajudar.

Um agradecimento especial ao meu orientador, o Prof. Renato Rocha Souza, que foi o grande responsável por eu ter conseguido chegar até o final dessa empreitada.

Meu agradecimento à Secretaria Municipal de Fazenda (SMF) do Município do Rio de Janeiro pelo apoio e, especialmente, aos amigos Ricardo de Azevedo Martins e Alexandre Calvet Lima por acreditarem nesse meu projeto meio alternativo.

Aos meus pais, que sempre tiveram os estudos de seus filhos como prioridade, meu agradecimento pelo ótimo exemplo. Isso me fez, mesmo que tardiamente, encarar esse desafio.

E, finalmente, mas porque a gente sempre guarda o melhor para o final, meu muito obrigado à minha amada esposa Poliana e meu amado filho Daniel que muitas vezes foram privados do meu convívio para que eu pudesse realizar esse curso, mas que sempre souberam compreender e apoiar, e renovar minhas energias para seguir em frente.

Resumo

A implantação dos sistemas de notas fiscais eletrônicas proporcionou uma grande quantidade de dados para as administrações tributárias. Analisar esses dados e extrair informações importantes é um desafio. Esse trabalho buscou, por meio de técnicas de análise de dados e mineração de textos, identificar, a partir da descrição dos serviços prestados, notas emitidas incorretamente a fim de respaldar um melhor planejamento de fiscalizações.
Palavras-chave: Mineração de textos, Nota Fiscal de Serviços eletrônica, Detecção de fraudes.

Abstract

The implementation of electronic invoices systems provided a large amount of data for tax administrations. Analyze this data and extract important information is a challenge. This study aimed, through data analysis and text mining techniques, identify, from description of services, invoices incorrectly issued to endorse better planning of inspections.
Keywords: *Text mining, electronic invoice, fraud detection*

Lista de Figuras

Figura 1 – Exemplo de uma Nota Carioca 14
Figura 2: Processo de Descoberta de Conhecimento e Mineração de Dados (Fayyad, 1996). 21
Figura 3: Etapas do processo de mineração de textos (Corrêa, 2012)..... 23
Figura 4: A curva Zipf e os cortes de Luhn (Matsubara; Martins; Monard, 2003). O eixo das ordenadas corresponde à frequência das palavras (*f*) e o eixo das abscissas (*r*) às palavras ordenadas segundo sua frequência decrescente..... 26
Figura 5: Procedimento de agrupamento (Xu, 2008). 30
Figura 6: Exemplo de um dendrograma (adaptado de Xu e Wunsch, 2008)..... 31
Figura 7: Agrupamento pelo algoritmo *K-means* com dois clusters (Segaran, 2007)..... 32
Figura 8: Clusterização (1) e atribuição de descrição aos clusters formados (2). 35
Figura 9: Comparativo entre as empresas pelo domicílio tributário de suas notas. 45
Figura 10: Comparativo entre as notas emitidas com tributação no MRJ e com tributação em outros municípios por quantidade de notas e valor dos serviços..... 46
Figura 11: Quantidade de notas emitidas por alíquota. 46
Figura 12: Wordcloud das palavras mais relevantes nas notas tributadas no MRJ..... 50
Figura 13: Wordcloud das palavras mais relevantes nas notas tributadas em outros municípios. 51

Lista de tabelas

Tabela 1: Unigramas mais frequentes 48
Tabela 2: Lista de palavras relevantes para caracterização do tipo de serviço prestado. 49
Tabela 3: Unigramas mais frequentes dentre as palavras relevantes..... 50
Tabela 4: 30 palavras mais frequentes em cada um dos dois clusters. 57
Tabela 5: Empresas selecionadas pelo método convencional e pelos classificadores. 59

Lista de abreviaturas e siglas

ISS: Imposto sobre serviços de qualquer natureza

MRJ: Município do Rio de Janeiro

CNPJ: Cadastro Nacional de Pessoa Jurídica

Sumário

1. INTRODUÇÃO 11
1.1. OBJETIVO DO TRABALHO 11
1.2. O IMPOSTO SOBRE SERVIÇOS 11
1.3. A NOTA CARIOCA 13
1.4. A ENGENHARIA CONSULTIVA 16
2. REFERENCIAL TEÓRICO 18
2.1. TRABALHOS RELACIONADOS 18
2.2. MINERAÇÃO DE DADOS 20
2.3. MINERAÇÃO DE TEXTOS 22
2.4. AGRUPAMENTO OU CLUSTERIZAÇÃO DE DOCUMENTOS 28
2.5. CLASSIFICAÇÃO DE DOCUMENTOS 36
3. METODOLOGIA 40
3.1. DESCRIÇÃO DOS DADOS 41
3.2. O PRÉ-PROCESSAMENTO DOS DADOS 44
3.3. ELABORAÇÃO DE RELAÇÃO DE EMPRESAS SUSPEITAS PELO MÉTODO "CONVENCIONAL" 51
3.4. AGRUPAMENTO DOS DOCUMENTOS 52
3.5. CLASSIFICAÇÃO PARA DETECÇÃO DE EMPRESAS SUSPEITAS 53
4. RESULTADOS 55
5. CONCLUSÃO E CONSIDERAÇÕES FINAIS 60
REFERÊNCIAS BIBLIOGRÁFICAS 62

Capítulo 1

INTRODUÇÃO

1.1. OBJETIVO DO TRABALHO

O objetivo desse trabalho é utilizar métodos de análise de dados e mineração de textos para extrair, dos dados do sistema da Nota Carioca, informações úteis ao planejamento das ações de fiscalização do Imposto sobre Serviços (ISS) no Município do Rio de Janeiro (MRJ).

Em particular, a intenção é melhorar a seleção das empresas da programação de fiscalização, agregando às informações já utilizadas, os dados provenientes da análise do campo "DISCRIMINAÇÃO DOS SERVIÇOS", que é um campo de texto livre no qual o contribuinte descreve o serviço que foi prestado.

Esse procedimento será testado em um subconjunto dos dados da Nota Carioca, representado pelas Notas Fiscais de Serviço Eletrônicas (NFS-e) emitidas com o código de serviço 07.19.04 (engenharia consultiva - acompanhamento e fiscalização da execução de obras de engenharia, arquitetura e urbanismo (relacionados com obras de construção civil, hidráulicas, de escoramento ou de contenção de encostas)), que muitas vezes é utilizado de maneira indevida a fim de permitir o deslocamento do domicílio tributário para outros municípios, ou seja, o pagamento do imposto em outros municípios.

1.2. O IMPOSTO SOBRE SERVIÇOS

O Imposto Sobre Serviços de Qualquer Natureza (ISS), de competência dos Municípios e do Distrito Federal, tem como fato gerador a prestação de serviços constantes da lista anexa, ainda que esses não se constituam como atividade preponderante do prestador (artigo 1º da Lei complementar nº 116 de 31 de julho de 2003). A lista anexa citada é composta de 40 itens,

diversos deles divididos em subitens, e relaciona as atividades de serviço tributáveis pelo ISS como, por exemplo, informática, medicina, construção civil, ensino e hospedagem.

O ISS foi instituído no Município do Rio de Janeiro pela Lei nº 691 de 24 de dezembro de 1984 (Código Tributário Municipal). Essa lei estabelece em seu artigo 16 que a base de cálculo do imposto é o preço dos serviços e, em seu artigo 33, que o imposto é calculado aplicando-se, sobre a base de cálculo, as alíquotas de seus incisos.

Domicílio tributário é o local onde deve ser pago o imposto. A Lei complementar n° 116 de 31 de julho de 2003, em seu artigo 3°, estabelece que o imposto é devido no local do estabelecimento prestador (local onde se encontra a empresa), exceto nas hipóteses previstas nos seus incisos, nas quais o imposto é devido no local onde o serviço é prestado (por exemplo, o local de realização de uma obra de construção civil). Essa mesma lei, em seu artigo 4°, define estabelecimento prestador como o local onde o contribuinte desenvolva a atividade de prestar serviços, de modo permanente ou temporário, e que configure unidade econômica ou profissional, sendo irrelevantes para caracterizá-lo as denominações de sede, filial, agência, posto de atendimento, sucursal, escritório de representação ou contato ou quaisquer outras que venham a ser utilizadas.

Assim, quando o serviço prestado corresponde a uma das atividades de serviço elencadas nos incisos do artigo 3° da Lei complementar n° 116/2003, o imposto deve ser pago onde o serviço foi prestado, ou seja, o imposto pode ser devido em município diferente daquele onde a empresa está localizada.

Alguns contribuintes, buscando alíquotas menores ou por imposição do Município de destino (onde o serviço foi prestado), emitem documentos fiscais indicando de maneira indevida códigos de serviço que permitem o deslocamento de domicílio tributário (aqueles elencados nos incisos do art. 3º da LC 116/2003). Isso gera perda de arrecadação para o Município do Rio de Janeiro.

Considerando que, desde a implantação do sistema da Nota Carioca, o valor do ISS de serviços indicados como tributados no Município do Rio de Janeiro (MRJ) foi de 18,5 bilhões de reais e o valor do ISS de serviços indicados como tributados em outros municípios foi de 2,6 bilhões de reais, o problema do domicílio tributário afeta valores da ordem de 12% da arrecadação do ISS. Estimando a possibilidade de recuperação de ISS em 10% do valor recolhido

para outros municípios, chega-se a um montante de 260 milhões de reais, que é superior ao orçamento anual de muitas secretarias do Município.

A fim de recuperar essa receita, o primeiro passo é identificar, dentro do universo de todos os contribuintes que indicam que o imposto é devido em outros municípios, aqueles com maior chance de estarem fazendo isso indevidamente.

1.3. A NOTA CARIOCA

A Nota Fiscal de Serviços Eletrônica (NFS-e ou Nota Carioca) é um documento fiscal relativo ao imposto sobre serviços (ISS) no Município do Rio de Janeiro emitido e armazenado eletronicamente em sistema próprio da Prefeitura. Ela foi instituída pela Lei n° 5.098 de 15 de outubro de 2009 e regulamentada pelo Decreto n° 32.250 de 11 de maio de 2010 e pela Resolução SMF n° 2.617 de 17 de maio de 2010.

Antes da introdução da Nota Carioca, os documentos fiscais eram emitidos a partir de talonários de Notas Fiscais de Serviço em papel ou formulários contínuos impressos gerados nos computadores dos contribuintes. Em ambos os casos, a administração tributária não tinha visibilidade das operações que ocorriam e havia fraudes dos mais diversos tipos, como por exemplo, a “nota calçada” onde a via entregue ao cliente (1ª via) apresentava o valor efetivamente pago pelo serviço e as duas vias que ficavam no talão apresentavam valor inferior, a fim de reduzir o tributo a ser pago.

O sistema da Nota Carioca é um sistema eletrônico de emissão de documentos fiscais de serviço e controle do recolhimento do imposto sobre serviços. A nota carioca emitida pelo prestador de serviço é mostrada no sistema para o tomador de serviço e também pode ser visualizada pela administração tributária. No processo de emissão da nota são fornecidas diversas informações (por exemplo, os dados do prestador e do tomador do serviço, valor do serviço, informações sobre benefícios fiscais e código do serviço) que permitem identificar o local onde o tributo é devido, o seu valor (calculado automaticamente) e quem deve pagar o imposto (prestador ou tomador). A Figura 1 mostra o exemplo de uma Nota Carioca.

PREFEITURA DA CIDADE DO RIO DE JANEIRO

SECRETARIA MUNICIPAL DE FAZENDA

NOTA FISCAL DE SERVIÇOS ELETRÔNICA - NFS-e

- NOTA CARIOCA -

Numero da Nota **00017583**

Data e Hora da Emissão **03/01/2014 08:11:03**

Codigo de Verificação **SIJX-45EP**

PRESTADOR DE SERVIÇOS

CPF/CNPJ: **60.637.263/0777-03** Inscrição Municipal: **0.572.954-8** Inscrição Estadual: **---**

Nome/Razão Social: **ALLPARK EMPREENDIMENTOS PARTICIPACOES E SERVICOS S.A.**

Nome Fantasia: **ESTAPAR** Tel.: **2103-4096**

Endereço: **RUA BARAO DE ITAMBI 50, LOJ A - BOTAFOGO - CEP: 22231-000**

Município: **RIO DE JANEIRO** UF: **RJ** E-mail: **nota.carioca@estapar.com.br**

TOMADOR DE SERVIÇOS

CPF/CNPJ: **016.692.247-16** Inscrição Municipal: **----** Inscrição Estadual: **----**

Nome/Razão Social: **NÃO INFORMADO**

Endereço: **----** Tel.: **----**

Município: **----** UF: **----** E-mail: **-----**

DISCRIMINAÇÃO DOS SERVIÇOS

Serv. Est: FGV

VALOR DA NOTA = R$ 15,00

Serviço Prestado

11.01.01 - guarda e estacionamento de veiculos terrestres automotores

Deduções (R$)	Desconto Incond. (R$)	Base de Calculo (R$)	Alíquota (%)	Valor do ISS (R$)	Credito Gerado (R$)
0,00	0,00	15,00	5,00%	0,75	0,07

OUTRAS INFORMAÇÕES

- Esta NFS-e foi emitida com respaldo na Lei nº 5.098 de 15/10/2009 e no Decreto nº 32.250 de 11/05/2010
- PROCON-RJ: Rua da Ajuda, 5 subsolo; www.procon.rj.gov.br
- O ISS referente a esta NFS-e foi recolhido em 10/01/2014.
- Esta NFS-e substitui o RPS Nº 22002 Série 1 (Cupom), emitido em 26/12/2013.
- Esta NFS-e participa do Sorteio Carioca sob o número 9.84/53443.

Figura 1 – Exemplo de uma Nota Carioca

Até o dia 25 de agosto de 2015, já haviam sido emitidas 584.173.573 notas cariocas por 171.812 empresas. As informações geradas por esses documentos fiscais são a base de trabalho da administração tributária, que analisa esses dados a fim de garantir o correto recolhimento do tributo. Os contribuintes que estão com o imposto em atraso recebem diversas notificações solicitando o recolhimento do tributo ou a correção do documento fiscal. Caso a pendência permaneça, os débitos são constituídos por meio de auto de infração.

Outras situações podem ser identificadas por meio dos dados da Nota Carioca que indicam recolhimento do imposto menor que o devido, como empresas que se dizem optantes pelo Simples Nacional sem sê-lo, empresas que se dizem imunes ou isentas sem gozar dessa condição, empresas que se intitulam sociedades uniprofissionais indevidamente, etc. Em alguns desses casos, é possível identificar a discrepância fazendo cruzamentos de dados, como no caso da opção pelo Simples Nacional, onde é possível comparar a informação registrada pelo contribuinte no perfil da empresa com o arquivo de optantes pelo Simples Nacional fornecido pela Receita Federal. Em outros casos, é preciso uma análise mais detalhada para identificar essas discrepâncias, como no caso das sociedades uniprofissionais, onde é necessário analisar diversas características da empresa a fim de confirmar ou não essa condição.

A identificação de discrepâncias no sistema da Nota Carioca possibilita o envio de mensagens solicitando correções e, em última análise, pode resultar em ações de fiscalização a fim de garantir o correto recolhimento do tributo. Em um universo de mais de 170.000 empresas, selecionar adequadamente aquelas que serão fiscalizadas é de fundamental importância para garantir a eficiência e eficácia necessárias à administração pública.

A Nota Carioca possui diversos campos que podem ser analisados a fim de identificar discrepâncias. Entretanto, o campo “DISCRIMINAÇÃO DOS SERVIÇOS”, que é preenchido pelo contribuinte com uma descrição sucinta do serviço prestado, não tem sido utilizado nessas análises por ser um campo de texto livre. Nesse trabalho serão exploradas as possibilidades de extração de informações a partir desse campo.

1.4. A ENGENHARIA CONSULTIVA

Conforme descrito no item 1.1, o objeto deste trabalho serão as informações constantes nas notas fiscais de serviços eletrônicas emitidas com o código 07.19.04 (engenharia consultiva - acompanhamento e fiscalização da execução de obras de engenharia, arquitetura e urbanismo (relacionados com obras de construção civil, hidráulicas, de escoramento ou de contenção de encostas)), desde a implantação do sistema da nota carioca (maio de 2010).

Esse subconjunto dos dados foi selecionado por apresentar possibilidade de dúvida em relação ao correto domicílio tributário das notas fiscais e representa 0,7% (4,5 bilhões de reais) do valor total dos serviços prestados no MRJ no período. Nesse código de serviço específico foram recolhidos a título de ISS ao MRJ 44 milhões de reais, enquanto foram recolhidos a outros municípios 122 milhões de reais.

A discriminação do serviço de engenharia consultiva consta no artigo 43, Seção II, Capítulo IX do Decreto nº 10.514 de 09 de outubro de 1991, transcrito a seguir:

> Art. 43 - Os serviços de engenharia consultiva, para os efeitos do disposto no inciso II, item 1, do art. 19, são os seguintes:
>
> I - elaboração de planos diretores, estudos de viabilidade, estudos organizacionais e outros, relacionados com obras e serviços de engenharia;
>
> II - elaboração de anteprojetos, projetos básicos e projetos executivos para trabalhos de engenharia;
>
> III - fiscalização e supervisão de obras e serviços de engenharia.
>
> Parágrafo único - O tratamento fiscal previsto no *caput* deste artigo destina-se exclusivamente aos serviços de engenharia consultiva que estiverem relacionados com obras de construção civil, hidráulicas, de escoramento e de contenção de encostas.

Esse serviço está submetido a uma alíquota de 3%, conforme at. 33, II, 1 da Lei nº 691/1984 e art. 19, II, 1 do Decreto nº 10.514/1991 e o imposto é devido no local onde se encontra o estabelecimento prestador no caso dos serviços previstos nos incisos I e II do art. 43

do Decreto nº 10.514/1991 ou no local da execução da obra no caso dos serviços previstos no inciso III desse mesmo decreto.

Dessa forma, a possibilidade de deslocamento de domicílio tributário ocorre apenas no caso de fiscalização e supervisão de obras de construção civil, não sendo pertinente no caso de elaboração de projetos e atividades similares ou se o trabalho de fiscalização não estiver associado a uma obra de construção civil.

O código na Nota Carioca para atividade prevista no inciso III do art. 43 do Decreto nº 10.514/1991 é o 07.19.04 (engenharia consultiva - acompanhamento e fiscalização da execução de obras de engenharia, arquitetura e urbanismo (relacionados com obras de construção civil, hidráulicas, de escoramento ou de contenção de encostas)).

Muitas vezes observa-se a utilização indevida do código 07.19.04 para serviços previstos nos incisos I e II do art. 43 do Decreto nº 10.514/1991 associada à indicação de outro município como local de pagamento do imposto. A análise da descrição do serviço constante no campo “DISCRIMINAÇÃO DOS SERVIÇOS” pode ajudar a identificar documentos fiscais com a incorreção descrita a fim de direcionar as ações de fiscalização.

Capítulo 2

REFERENCIAL TEÓRICO

Nesse capítulo serão apresentados os métodos que foram utilizados para a consecução dos objetivos desse trabalho.

Inicialmente, será feita uma descrição de trabalhos que utilizaram técnicas similares a do presente trabalho, em particular, a mineração de textos, e de trabalhos que versaram sobre a identificação de empresas "sonegadoras".

Posteriormente, será abordada a mineração de dados, por ser a grande área a qual pertence esse trabalho, seguida pela apresentação da mineração de textos, agrupamento (clusterização) de documentos e classificação de documentos, que foram os métodos efetivamente utilizados.

2.1. TRABALHOS RELACIONADOS

Nessa seção vamos apresentar alguns trabalhos sobre mineração de textos, em especial com a utilização de técnicas de agrupamento (*clustering*), assim como os trabalhos já realizados que buscam extrair conhecimentos a partir de documentos fiscais.

Morais e Ambrósio (2007) descrevem em seu relatório técnico o estado da arte da mineração de texto, apresentando os conceitos envolvidos e detalhando as etapas do processo de descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases - KDD*) e do processo de descoberta de conhecimento em textos (*Knowledge Discovery from Text – KDT*).

Ferauche (2011) utilizou técnicas de mineração de textos para classificar as ementas que compõem a jurisprudência do Tribunal Regional do Trabalho da 2ª Região – São Paulo. Essas ementas já estavam previamente classificadas e foram utilizadas para treinar um classificador em um processo de aprendizado supervisionado. A eficácia do classificador foi avaliada comparando-se a classificação automática de documentos não previamente classificados com a classificação realizada por um especialista da área. A conclusão obtida é que seria possível, com

auxílio de técnicas de classificação de textos e aprendizado supervisionado, um sistema computacional indicar a categoria mais provável a que uma ementa pertenceria, auxiliando o trabalho do especialista classificador.

Passini (2012) aplicou técnicas de mineração de texto para organizar documentos de uma central de atendimento, a fim de facilitar a recuperação de informação, resultando em melhoria no atendimento. Foram utilizadas técnicas de agrupamento hierárquico e classificação de documentos e também de modelagem de tópicos (alocação latente de Dirichlet – LDA) na qual os documentos são representados por meio de tópicos probabilísticos contendo palavras-chave e um mecanismo de seleção imune-supressor para a seleção dos documentos mais representativos.

Wives (2002) apresentou em seu trabalho diversas técnicas de recuperação de informação, mineração de textos e inteligência competitiva, a fim de possibilitar a coleta, seleção e análise das informações mais importantes em um ambiente empresarial.

Côrrea, Marcacini e Rezende (2003) aplicaram métodos não supervisionados de agrupamento hierárquico a fim de realizar uma análise exploratória em uma base de textos de pesquisa médica sobre câncer de cabeça e pescoço. Verificou-se que esses métodos foram eficazes para a identificação de tópicos mais genéricos baseados nos termos mais frequentes, entretanto não conseguiram identificar tópicos mais específicos e inovadores. Para tanto seria necessário a utilização de um dicionário ou ontologia de domínio que apoiasse a seleção de termos mais específicos e a participação de especialistas da área em técnicas de aprendizado supervisionado.

Lopes (2004) desenvolveu uma metodologia para clusterização de documentos em português. Foram apresentadas em detalhe as técnicas de pré-processamento de dados textuais, diversos métodos para clusterização e categorização de documentos e técnicas de visualização para facilitar a interpretação dos resultados do agrupamento. Também foi discutido como a qualidade dos dados (presença ou não de atributos comuns) e a quantidade de registros prejudica a precisão dos métodos, e como a aplicação do *stemming* melhora o desempenho do agrupamento.

Segaran (2007) descreveu um processo de filtragem de e-mails a fim de evitar mensagens indesejadas (*spam*). Foi mostrado como definir e treinar um classificador Naïve Bayes, com especial ênfase no cálculo das probabilidades necessárias para a aplicação do teorema de Bayes.

Também é apresentado o método de Fisher para classificação que apresenta resultados mais precisos que o Naïve Bayes, especialmente para filtragem de *spam*.

Boavista e Silva (2011a, 2011b e 2013) elaboraram notas técnicas mostrando o impacto da implantação do sistema da Nota Carioca no desempenho da arrecadação do ISS. Além dos ganhos financeiros, é destacada a importância do Sistema da Nota Carioca na simplificação de procedimentos das empresas e na disponibilização de informações em tempo real para a administração tributária.

Andrade (2009) propôs um processo de mineração de dados a fim de identificar sonegação em contribuintes do ICMS (imposto sobre operações relativas à circulação de mercadorias e sobre prestações de serviços de transporte interestadual e intermunicipal e de comunicação). Foram analisadas informações cadastrais, de arrecadação e declarações dos contribuintes por meio de algoritmos de Redes Neurais Artificiais (RNA).

Coelho (2012) apresentou uma metodologia de detecção de fraudes fiscais no imposto sobre serviços de qualquer natureza (ISSQN) por meio mineração de dados com aplicação de ontologias e sistemas difusos. É especialmente interessante a aplicação de técnicas de sistemas difusos para modelar o conhecimento coletivos de especialistas, por vezes de natureza vaga e subjetiva. O resultado obtido com a aplicação dessa metodologia atingiu um índice de acerto de 96%.

Carvalho (2014) propôs um sistema de inferência difusa (SID) para classificação de sonegadores fiscais da Receita Federal do Brasil (RFB). Atualmente, os contribuintes são classificados em “sonegadores” ou “não sonegadores”. Esse trabalho introduziu uma classificação gradual com diversos níveis de sonegação. Os resultados obtidos por essa metodologia se mostraram bem superiores aos obtidos pela lógica clássica.

2.2. MINERAÇÃO DE DADOS

Mineração de dados é o processo de obtenção e interpretação de informações, assim como a elaboração de modelos preditivos, a partir de dados em grande escala. Ela é um campo interdisciplinar que utiliza conceitos de bancos de dados, estatística, aprendizagem por máquina, entre outros.

A mineração de dados é uma parte do processo de Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases – KDD*), que inclui tarefas de pré-processamento como extração, limpeza e redução de dados e tarefas de pós-processamento como reconhecimento de padrões e interpretação de modelos, conforme representado na Figura 2.

As principais etapas do processo de mineração de dados são a análise exploratória, a descoberta de padrões frequentes, o agrupamento e a classificação dos dados.

A análise exploratória dos dados consiste em processar os dados numéricos e categóricos a fim de extrair características chave da amostra por meio de estatísticas que fornecem medidas de centralidade, dispersão, etc.

Outro objetivo da análise exploratória é reduzir a quantidade de dados a serem minerados, que permite prover uma explicação e visualização simplificadas, suprimir ruídos de forma a obter uma melhor predição e a redução de tempo computacional. Métodos de seleção de atributos e redução de dimensionalidade são utilizados para selecionar as dimensões mais importantes, métodos de discretização são adotados para reduzir a quantidade de valores de um atributo e métodos de amostragem para reduzir o tamanho dos dados.

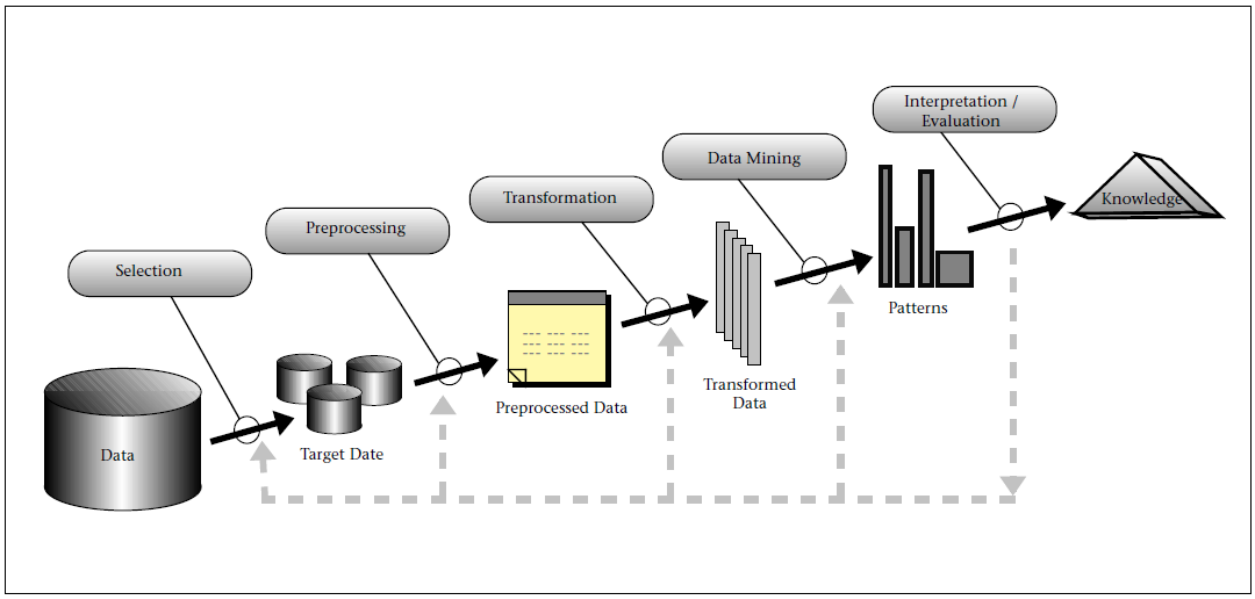

Figura 2: Processo de Descoberta de Conhecimento e Mineração de Dados (Fayyad, 1996).

A descoberta de padrões frequentes está associada à extração de padrões informativos a partir de dados complexos e em larga escala. Alguns exemplos desses padrões são conjuntos com

mesmos valores de um atributo, sequências e grafos. O ponto chave é encontrar tendências ou comportamentos ocultos nos dados, a fim de melhor compreender as interações entre pontos e atributos.

A clusterização (*clustering*) ou agrupamento é a etapa na qual os dados são separados em grupos afins chamados *clusters*. A ideia é que pontos com atributos similares estejam em um mesmo *cluster*, enquanto pontos com atributos dissonantes devem pertencer a *clusters* distintos. A similaridade entre os atributos é avaliada por meio de uma medida de similaridade, como por exemplo a distância euclidiana e o coeficiente de correlação de Pearson. O agrupamento é um processo de aprendizado não-supervisionado, assim não é necessário que haja registros previamente categorizados. Observe que o agrupamento, via de regra, não é utilizado para classificar, estimar ou predizer valores da variável, ele apenas identifica grupos de dados similares e pode ser utilizado para gerar o conjunto de treinamento de um classificador. O processo de clusterização será abordado com mais detalhes na seção 2.3.

A classificação é a etapa que consiste em predizer a categoria ou classe de um registro não classificado. A fim de construir um classificador, é necessário um conjunto de registros classificados corretamente. Esse conjunto é chamado conjunto de treinamento. Após "aprender" com os registros do conjunto de treinamento, o classificador pode prever automaticamente a classe de qualquer novo registro. Note que, diferentemente do agrupamento, a classificação é um processo de aprendizado supervisionado. Há diversos tipos de classificadores, como por exemplo, árvores de decisão, classificadores probabilísticos e máquinas de vetor de suporte (*Suport Vector Machines – SVM*).

2.3. MINERAÇÃO DE TEXTOS

A mineração de textos (*text mining*) é um processo de descoberta de conhecimento a partir de textos (*Knowledge Discovery from Text – KDT*), ou seja, a partir de dados não estruturados ou semi-estruturados por meio de técnicas de extração e análise de dados.

O objetivo da mineração de textos é identificar padrões ou tendências em grandes volumes de textos em linguagem natural, a fim de extrair informações úteis e não evidentes

nesses textos que, normalmente, não poderiam ser obtidas utilizando os métodos de consulta tradicionais.

Uma das grandes motivações para o desenvolvimento da mineração do texto foi o aumento da disponibilização de dados textuais ou não estruturados em meio eletrônico, tanto nos sistemas corporativos quanto na internet. Estima-se que 80% das informações das empresas estão armazenadas em forma de texto (Tan, 1999). Assim, essa área possui um vasto campo de estudo que inclui documentos eletrônicos (elaborados em editores de texto ou digitalizados de documentos impressos), páginas de notícias, e-mails, campos de textos em bancos de dados, etc.

Um processo de mineração de textos pode ser dividido em quatro etapas principais: identificação do problema, pré-processamento, extração de conhecimento e pós-processamento (avaliação e validação de resultados).

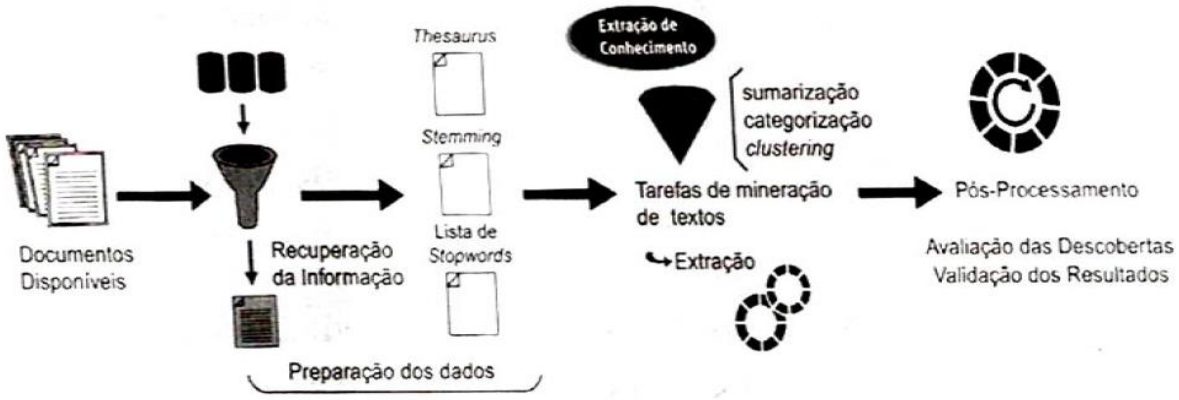

Figura 3: Etapas do processo de mineração de textos (Corrêa, 2012)

A etapa de identificação do problema consiste em selecionar as bases de textos que serão utilizadas e definir o escopo das análises que serão realizadas a fim de atingir os objetivos desejados.

Nessa etapa deve-se definir o tipo de abordagem que será feita dos textos: análise semântica ou análise estatística.

A análise semântica avalia a sequência de termos no contexto da frase, a fim de identificar a correta função de cada termo (Cordeiro, 2005), por meio de técnicas de Processamento de Linguagem Natural (*Natural Language Processing – NLP*). Essas técnicas agregam conhecimentos morfológicos, sintáticos, semânticos, do discurso, entre outros.

A análise estatística busca extrair informações com base na frequência dos termos encontrados nos textos. É comum nesse caso, modelar os documentos como um *bag of words* (“saco de palavras”). Esse modelo não leva em conta a ordem das palavras ou seu contexto, mas apenas o número de vezes que cada palavra aparece nos textos.

O objetivo da etapa de pré-processamento é extrair de textos em linguagem natural, inerentemente não estruturados, uma representação estruturada, concisa e manipulável por algoritmos de extração de conhecimento. Normalmente, os documentos são representados por vetores de atributos (*feature vectors*). Um atributo nada mais é do que uma entidade sem estrutura interna e uma dimensão no espaço de atributos. Um documento é representado como um vetor nesse espaço, ou seja, uma sequência de pesos de cada um dos atributos.

Inicialmente, devemos utilizar um *parser*, ou seja, um analisador léxico que identifique as palavras presentes nos documentos e elimine os símbolos e caracteres de controle de arquivo ou de formatação.

Adotando o modelo *bag of words* descrito acima, cada termo presente na coleção de textos representa um atributo. Assim, a dimensão do espaço de atributos é igual à quantidade de palavras diferentes em todos os documentos.

Nessa fase é importante observar que mesmo uma coleção com poucos documentos de tamanho médio pode ter uma quantidade muito grande de atributos (palavras diferentes). Assim, é importante a adoção de técnicas para a redução da quantidade de atributos (dimensão), como por exemplo o procedimento de *case folding*, a remoção de *stopwords*, o processo de *stemming*, a utilização de dicionários (*thesaurus*) e a redução de atributos por medidas de relevância.

O *case folding* consiste em converter todos os caracteres para um único tamanho, maiúsculo ou minúsculo. Esse procedimento reduz significativamente a quantidade de atributos, mas deve-se tomar cuidado para não eliminar informação que possa ser importante para o problema em análise. Um exemplo disso é a filtragem de mensagens de spam que comumente são inteiramente escritas em letras maiúsculas. Nesse caso, utilizar o *case folding* vai levar ao descarte de informações importantes para caracterizar um *spam* (Segaran, 2007).

A remoção de *stop words* consiste na eliminação de palavras que não são relevantes na análise do texto por não traduzirem sua essência, não sendo úteis para a discriminação dos textos. Normalmente, a relação de *stop words* (*stop list*) inclui preposições, pronomes, artigos, advérbios

e outras classes de palavras auxiliares, assim como palavras que aparecem em praticamente todos os documentos da coleção.

A lematização da palavra (*stemming*) é uma técnica de normalização linguística, na qual as variantes de um termo são reduzidas a uma forma comum denominada *stem* (radical). Isso resulta na eliminação se prefixos, sufixos e características de gênero, número e grau das palavras, reduzindo o número de atributos em até 50% (Morais, 2007). Por exemplo, as palavras trabalham, trabalhando, trabalhar e trabalho seriam todas transformadas no mesmo *stem* trabalh. Os algoritmos de *stemming* são extremamente dependentes da língua para o qual foram escritos. Dessa forma, é necessário utilizar um *stemmer* especialmente projetado para as palavras em português, sendo os principais a versão para português do algoritmo de Porter (Porter, 2005), o removedor de sufixo da língua portuguesa – RSLP (Orengo e Huyck, 2001) e o algoritmo STEMBR (Coelho, 2007).

Um *Thesaurus* é um dicionário que mapeia sinônimos, acrônimos e ortografias alternativas em um termo único, que expressa a ideia geral dos elementos.

Após a aplicação das técnicas descritas, o próximo passo é elaborar a tabela documento-termo, ou seja, o cálculo dos pesos dos atributos e a construção dos vetores correspondentes a cada documento da coleção, conforme o modelo *bag of words* escolhido.

Observa-se que termos que possuem alta frequência têm maior importância no documento, enquanto termos que aparecem em uma grande quantidade de documentos têm sua importância diminuída. Sendo assim, a fim de caracterizar uma coleção de documentos, é importante analisar a frequência de cada palavra nos textos e em toda a coleção, o que permite identificar a importância de cada palavra em relação aos textos analisados, bem como distinguir os textos entre si.

Diversas medidas podem ser utilizadas para atribuir pesos aos atributos. Uma possibilidade simples é utilizar a frequência da palavra no documento. Um modelo um pouco mais elaborado, que leva em consideração a frequência das palavras na coleção é a estatística TF-IDF (Feldman e Sanger, 2006) que atribui à palavra w no documento d o peso

$$TF-IDF(w,d) = TermFreq(w,d) \cdot \log\left(N / DocFreq(w)\right)$$

onde $TermFreq(w,d)$ é a frequência da palavra w no documento d e $DocFreq(w)$ é o número de documentos contendo a palavra w.

Um outro método de redução da dimensionalidade dos atributos é selecionar os termos que são mais representativos para a discriminação dos documentos. A Lei de Zipf (Zipf, 1932), também conhecida como princípio do menor esforço, estabelece que o histograma ordenado de forma decrescente do número de ocorrências de cada termo na coleção tem a forma de uma "curva Zipf", que plotada em uma escala logarítmica resulta em uma reta com inclinação -1. Isso equivale a dizer que o i-ésimo termo mais comum em uma coleção de textos ocorre com frequência inversamente proporcional a i. Baseado na lei de Zipf, Luhn elaborou um método para definir dois pontos de corte (superior e inferior), entre os quais estariam os atributos mais relevantes. Os termos que excedem o corte superior são muito frequentes na coleção e os termos abaixo do corte inferior raros, ambos não contribuem significativamente para a discriminação dos textos entre si. A Figura 4 mostra a curva Zipf (I) e os cortes de Luhn aplicados à curva Zipf.

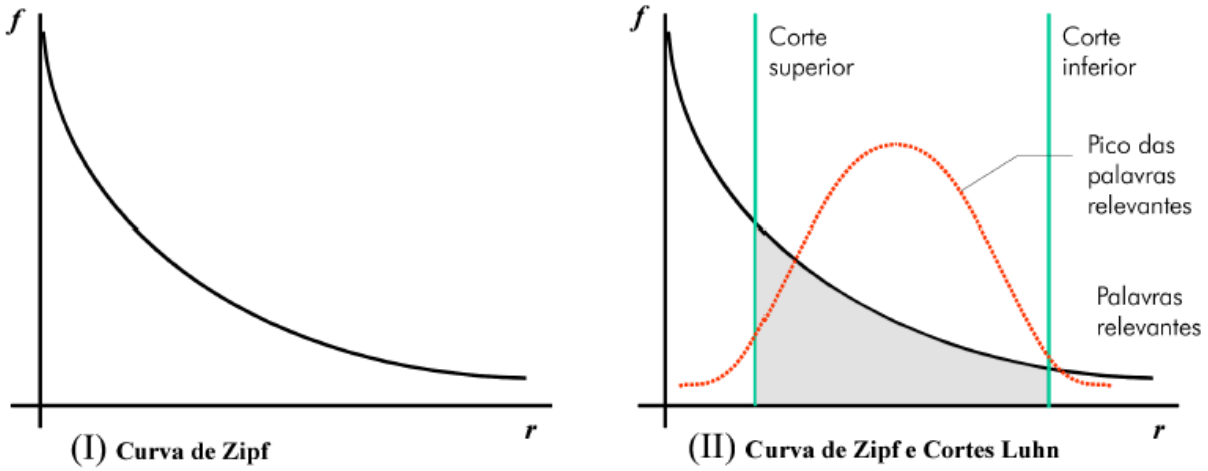

Figura 4: A curva Zipf e os cortes de Luhn (Matsubara; Martins; Monard, 2003). O eixo das ordenadas corresponde à frequência das palavras (*f*) e o eixo das abscissas (*r*) às palavras ordenadas segundo sua frequência decrescente

Encerrada a fase de pré-processamento, a nossa coleção de textos já está em uma forma estruturada e pronta para que sejam aplicados os algoritmos de extração de conhecimento.

Na etapa de extração do conhecimento busca-se a identificação de padrões que forneçam informações sobre os documentos da coleção e relacionamentos entre eles. Na análise de

coleções de textos as principais técnicas adotadas são a sumarização, a classificação (ou categorização) e o agrupamento (*clustering*).

A sumarização automática é um processo de redução do tamanho de textos, gerando um documento que retém os pontos mais importantes do documento original. A sumarização também pode ser vista também como a arte de extrair conteúdos chave de uma ou mais fontes de informação (Hahn, 2000).

A classificação ou categorização de documentos consiste em classificar um documento em um conjunto pré-especificado de categorias. Assim, dado um conjunto de categorias, assuntos ou tópicos e uma coleção de documentos de texto, a classificação é o processo de encontrar o assunto correto associado a cada documento. (Feldman; Sanger, 2006).

Há duas abordagens principais para realizar a classificação de documentos. Uma abordagem é centrada no conhecimento de especialistas, que é diretamente codificado no sistema declarativamente ou em forma de regras de classificação. A principal desvantagem desse método é que ele demanda uma grande quantidade de trabalho altamente qualificado para criar e manter o sistema de regras. A outra abordagem é a aprendizagem por máquinas (*machine learning – ML*), na qual um processo indutivo constrói um classificador com base no aprendizado obtido a partir de exemplos pré-classificados. Essa tem sido a abordagem preferida atualmente, pois requer apenas um conjunto de registros de treinamento classificados manualmente.

Na abordagem de aprendizagem por máquina (*machine learning*), o classificador é construído automaticamente a partir do aprendizado das propriedades de cada categoria obtido com base em documentos de treinamento pré-classificados. Esse é um processo supervisionado, pois é guiado pelo conhecimento da categoria correta dos registros do conjunto de treinamento. A versão não-supervisionada do problema de classificação é o problema de clusterização (*clustering*).

O agrupamento ou clusterização (*clustering*) é outra forma de extração de conhecimento em textos é, no qual uma coleção de textos não previamente classificados deve ser organizada ("dividida") em grupos afins (*clusters*), baseado em uma medida de similaridade. Os textos de um mesmo grupo devem ser altamente similares, mas dissimilares em relação aos textos de outros grupos. Por se tratar de um processo de aprendizado não supervisionado, a clusterização é especialmente útil quando há pouca informação prévia disponível sobre os textos. Uma descrição

mais detalhada do procedimento para clusterização de coleções de textos será feita na próxima seção.

Na etapa de avaliação e interpretação dos resultados (pós-processamento) de um processo de agrupamento de textos é feita a validação das informações obtidas (qualidade do agrupamento) e a análise do seu significado, que muitas vezes requer a análise por especialistas da área, a integração com outras evidências experimentais e a utilização de técnicas apropriadas de visualização de informação.

A avaliação dos resultados obtidos pode ser realizada de forma subjetiva, utilizando o conhecimento de um especialista na área, ou de forma objetiva por meio de índices estatísticos que medem a qualidade dos resultados.

A avaliação objetiva é feita por meio de índices adotados para a avaliação dos agrupamentos. O uso de técnicas para mensurar a qualidade de agrupamentos é especialmente importante porque o algoritmo sempre encontra grupos, independente destes serem reais ou não (Halkidi, 2001). A validação de um agrupamento pode ser feita por meio de critérios internos, relativos ou externos. Os critérios internos medem a qualidade de um agrupamento a partir de informações do próprio conjunto de dados, geralmente analisando se as posições dos objetos correspondem a uma matriz de proximidade. Os critérios relativos comparam diversos agrupamentos para decidir qual deles é o mais adequado aos dados. Os critérios externos utilizam o conhecimento de um especialista da área para avaliar se o agrupamento obtido corresponde a uma característica efetivamente presente nos dados (Côrrea, 2012).

Informações mais detalhadas sobre avaliação de agrupamentos podem ser encontradas em Vendramin (2010), onde é realizada a comparação de diversos índices de validação relativa de agrupamentos, e nos trabalhos de Halkidi (2001), Jain (1988) e Xu (2008) que apresentam uma visão geral sobre as técnicas de validação de agrupamentos.

2.4. AGRUPAMENTO OU CLUSTERIZAÇÃO DE DOCUMENTOS

O agrupamento ou clusterização (*clustering*) é uma técnica de extração de conhecimento que permite agrupar documentos similares em coleções de texto sem que se tenham informações prévias sobre os grupos (por exemplo, a quantidade de grupos ou o que representa cada grupo).

Nos algoritmos de agrupamento não há classes ou rótulos previamente definidos para treinamento de um modelo, ou seja, o aprendizado é realizado de forma não supervisionada. Observe que o problema de agrupamento é intrinsecamente um problema de otimização, pois envolve maximizar a similaridade interna dos grupos (intragrupo) e minimizar a similaridade entre os grupos (intergrupos). O processo de agrupamento também é conhecido como aprendizado por observação ou análise exploratória dos dados, pois a organização dos objetos em grupos é realizada apenas pela observação das regularidades (padrões) nos dados, sem uso de conhecimento externo (Xu; Wunsch, 2008).

A fim de aplicar um algoritmo de clusterização, cada texto da coleção deve ter passado pela etapa de pré-processamento descrita na seção anterior e deve estar representado por um vetor de atributos (temas dominantes ou palavras chave) composto pelas medidas de importância relativa de cada um dos atributos. A figura 5 mostra um processo completo de extração de conhecimento por meio de agrupamento.

A próxima etapa é determinar a proximidade de dois documentos baseada em seus vetores de atributos. No agrupamento de documentos, a proximidade pode ser calculada pelas métricas de distância convencionais ou por medidas de similaridades conceituais, baseadas em uma função de distância entre tópicos na hierarquia de assuntos e os pesos desses tópicos nos documentos. Calculadas as distâncias entre os documentos, eles podem ser agrupados usando diferentes métodos.

As estratégias de agrupamento podem ser divididas em dois tipos: agrupamento particional e agrupamento hierárquico. No agrupamento particional a coleção de documentos é dividida em uma partição simples de k grupos (*clusters*), enquanto no agrupamento hierárquico é produzida uma sequência de partições aninhadas, ou seja, a coleção textual é organizada em grupos e subgrupos de documentos (Feldman; Sanger, 2006).

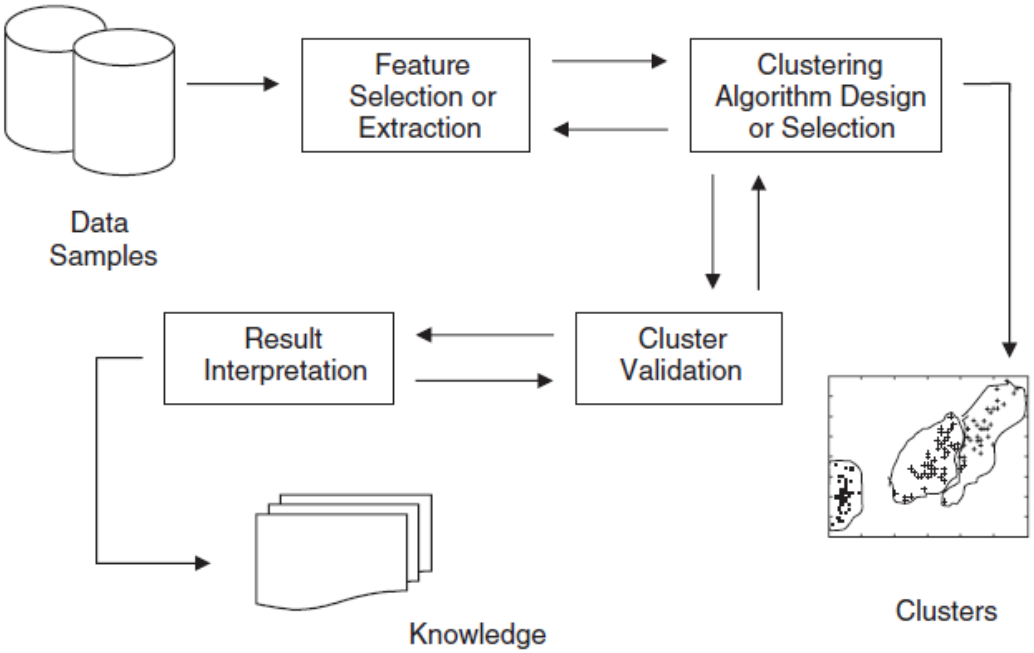

Figura 5: Procedimento de agrupamento (Xu, 2008).

Os algoritmos de agrupamento permitem sobreposição de grupos, ou seja, um mesmo documento pode pertencer a mais de um grupo ou possuir um grau de pertinência associado a cada grupo. As estratégias de agrupamento sem sobreposição são chamadas de estratégias rígidas ou *crisp* e serão nosso objeto de análise.

O agrupamento hierárquico produz uma hierarquia de partições com uma partição simples incluindo todos os documentos em um extremo e grupos (*clusters*) unitários cada um contendo um único documento no outro extremo. Cada *cluster* ao longo da hierarquia é visto como uma combinação de dois *clusters* a partir do próximo nível mais alto ou mais baixo, conforme a abordagem seja divisiva (*bottom-up*) ou aglomerativa (*top-down*), respectivamente. A árvore descrevendo a hierarquia de clusters é chamada de dendrograma (Figura 6).

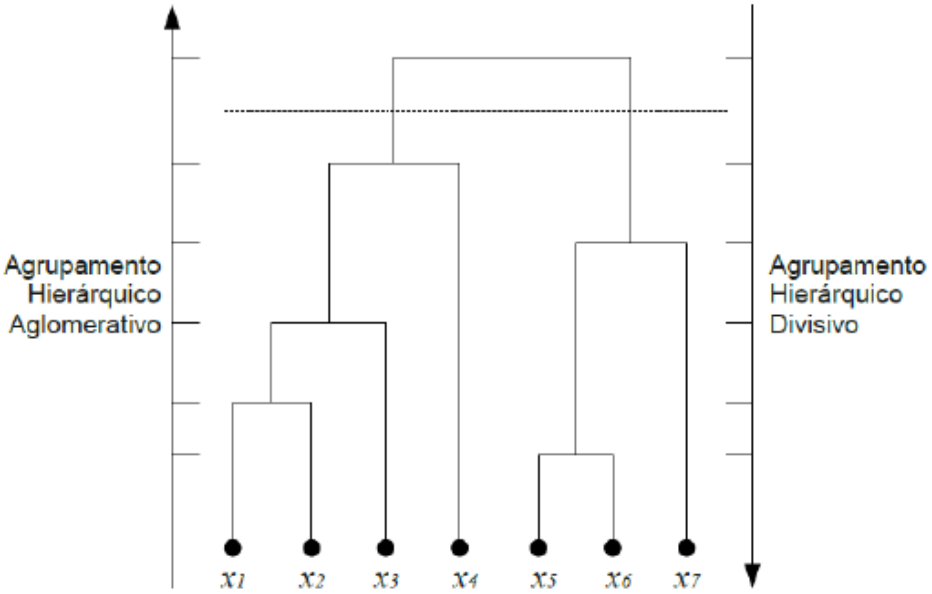

Figura 6: Exemplo de um dendrograma (adaptado de Xu e Wunsch, 2008).

O agrupamento particional ou agrupamento por otimização tem por objetivo dividir iterativamente o conjunto de objetos em k grupos, de forma a maximizar a similaridade interna dos grupos (intragrupo) e minimizar a similaridade entre os grupos (intergrupos). A quantidade k de grupos, em geral, é informada pelo usuário.

O algoritmo *K-means* (MacQueen, 1967) é o agrupamento particional mais usual e é muito utilizado em coleções de textos (Steinbach, 2000). Dado um número k, o método de agrupamento *K-means* cria um conjunto de k *clusters* (grupos) e distribui o conjunto de documentos dados nesses *clusters* usando a similaridade entre o vetor de atributo de cada documento e o centroide dos grupos. O centroide de cada *cluster* é o valor médio dos vetores de atributo de todos os documentos do *cluster*. Cada vez que um novo documento é agregado a um *cluster*, o centroide é recalculado.

Para aplicar algoritmo *K-means*, devemos ter como entrada uma coleção de documentos $X = \{x_1, x_2, \ldots, x_N\}$ e o número k de grupos. Deve-se selecionar k pontos que serão o valor inicial dos centroides, que podem ser selecionados aleatoriamente, podem ser k documentos do conjunto X ou algum outro critério. Para cada documento $x_i \in X$ é calculada a similaridade para cada um dos k centroides e o documento x_i é atribuído ao *cluster* cujo centroide corresponde à maior similaridade. Feito isso, todos os centroides são recalculados e o procedimento (cálculo das

similaridades e atribuição aos *clusters*) é repetido até que seja atingido algum critério de parada, que pode ser, por exemplo, não haver mais alterações no agrupamento ou um número máximo de interações. (Côrrea, 2012).

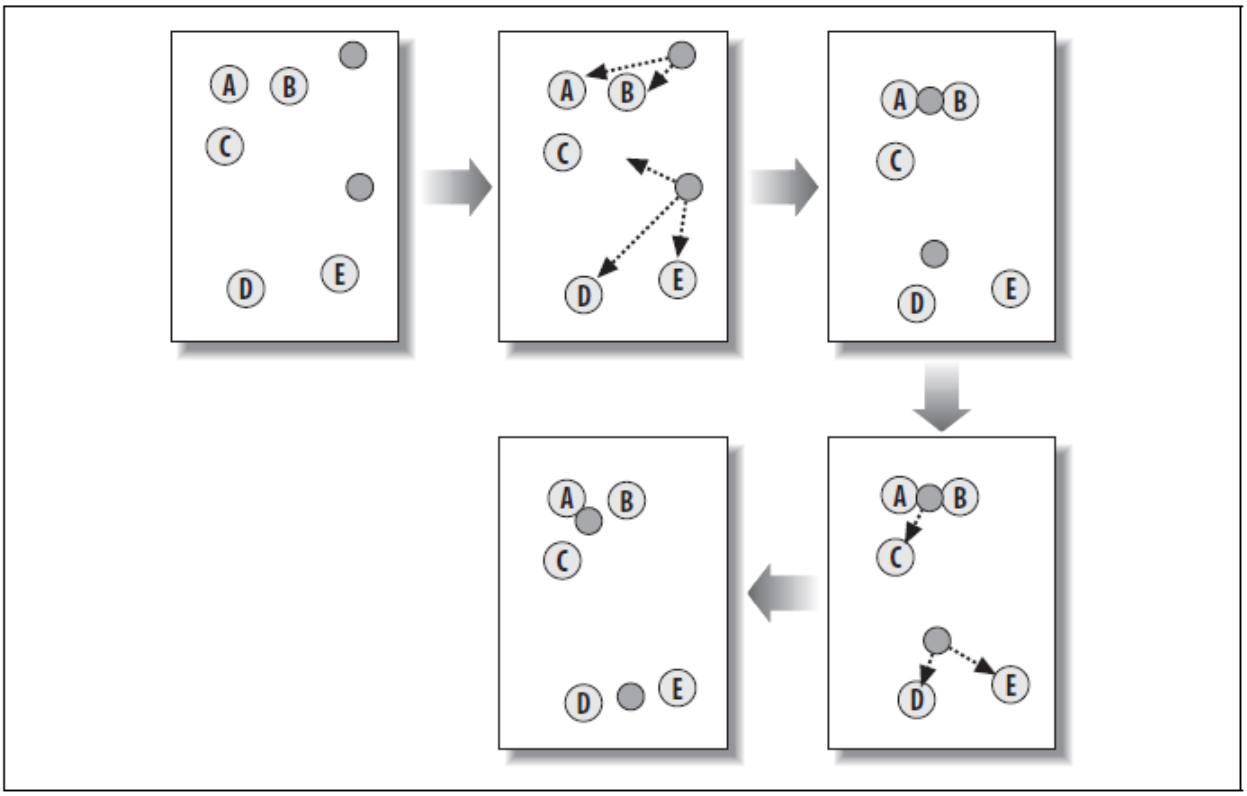

Figura 7: Agrupamento pelo algoritmo *K-means* com dois clusters (Segaran, 2007).

O agrupamento pelo *K-means*, em geral, produz agrupamentos piores que os métodos hierárquicos, mas por ter complexidade linear em relação ao número de documentos, tem custo computacional bem menor que aqueles e os resultados muitas vezes são satisfatórios.

Uma das etapas mais importantes do algoritmo *K-means* é o cálculo da similaridade entre os documentos e os centroides dos grupos. A similaridade é calculada com base em uma função de distância, que pode estar relacionada à própria distância euclidiana ou a alguma medida de similaridade como um coeficiente de correlação.

Alguns exemplos de funções de distância são a distância euclidiana, a distância euclidiana harmonicamente somada, distância *city-block*, coeficiente de correlação de Pearson, valor absoluto do coeficiente de correlação de Pearson, correlação de Pearson descentralizada e correlação de Pearson descentralizada absoluta. As três primeiras funções de distância estão

relacionadas à distância euclidiana, enquanto as outras quatro estão associadas a coeficientes de correlação.

A distância euclidiana é definida como

$$d = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i \right)^2}$$

onde x_i e y_i são a i-ésima componente do vetor de atributos que tem dimensão *n*.

Diferentemente das funções de distância baseadas em correlação, a distância euclidiana leva em consideração a magnitude das diferenças dos valores dados, preservando mais informações sobre os dados originais e pode ser preferível. Apesar de ser uma métrica usual para dados convencionais, a distância euclidiana e outras dela derivadas, quando aplicadas a dados textuais, apresentam resultados piores que o de outras métricas (Lopes, 2004).

A distância euclidiana harmonicamente somada é definida como

$$d = \left[\frac{1}{n} \sum_{i=1}^{n} \left(x_i - y_i \right)^{-2} \right]^{-1}.$$

Comparada com a distância euclidiana, a distância euclidiana harmonicamente somada é mais robusta contra dados incorretos.

A distância *city-block* ou *Manhattan* é a soma das distâncias em cada dimensão e é definida por

$$d = \sum_{i=1}^{n} \left| x_i - y_i \right|.$$

O coeficiente de correlação de Pearson é definido como

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

onde $\bar{x}$ e $\bar{y}$ são as médias das componentes dos vetores x e y, respectivamente, e σ_x e σ_y são o desvio padrão das componentes desses vetores.

O coeficiente de correlação de Pearson mede quão bem uma reta pode ser adequada a um gráfico de dispersão (*scatterplot*) de x e y. Se todos os pontos do gráfico de dispersão estão sobre uma linha reta, o coeficiente de correlação de Pearson é ± 1, dependendo se a inclinação da reta é positiva ou negativa. Se o coeficiente de correlação de Pearson é nulo, não existe correlação linear entre x e y. O coeficiente de correlação de Peason centraliza os dados pela subtração da média e os normaliza ao dividir pelo desvio padrão. Essa normalização é útil em diversas situações, mas em alguns casos é importante preservar a magnitude dos dados originais.

A distância de Pearson é definida como

$$d_P = 1 - r$$

onde $d_P \in [0,2]$.

A distância de Pearson absoluta é definida por

$$d_A = 1 - |r|$$

onde r é o coeficiente de correlação de Pearson e $d_A \in [0,1]$.

A correlação de Pearson descentralizada é dada por

$$r_U = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i}{\sigma_x^{(0)}} \right) \left(\frac{y_i}{\sigma_y^{(0)}} \right)$$

onde $\sigma_x^{(0)}$ e $\sigma_y^{(0)}$ são o módulo dos vetores de atributos x e y, respectivamente. Observe que o coeficiente de correlação de Pearson descentralizado é exatamente o produto escalar dos dois vetores de atributos dividido pelos seus módulos, ou seja, o cosseno do ângulo estre esses vetores.

A distância de Pearson descentralizada é dada por

$$d_U = 1 - r_U$$

onde $d_U \in [0,2]$

A distância de Pearson descentralizada absoluta é definida como

$$d_{AU} = 1 - |r_U|$$

onde $d_{AU} \in [0,1]$.

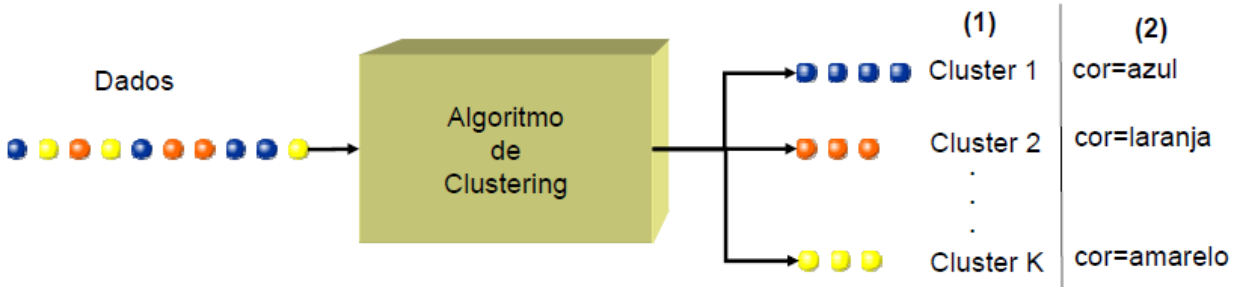

Figura 8: Clusterização (1) e atribuição de descrição aos clusters formados (2).

Realizado o agrupamento dos dados, deve-se atribuir uma descrição aos *clusters* formados de forma a auxiliar a interpretação dos resultados. Isso é importante, pois o agrupamento geralmente é utilizado em atividades exploratórias para descoberta de conhecimento, assim é necessário indicar o significado de cada grupo para que os usuários entender melhor os resultados obtidos. A atribuição de uma descrição para os clusters pode ser feita adotando-se para representa-lo o título do documento mais próximo do centroide do grupo ou os termos mais frequentes do grupo. Técnicas de aprendizado por máquina também são utilizadas para obter essa descrição (Manning, 2014).

Nesse trabalho foi utilizado o algoritmo *K-means* do pacote *scikit learning*. Ele resolve o problema pelo algoritmo de Lloyd, que foi o primeiro algoritmo proposto para esse método em 1957. Esse algoritmo vai melhorando iterativamente a posição dos centroides dos clusters a partir da posição inicial dos centroides, chamadas sementes (seeds). O algoritmo de Lloyd não define

como devem ser inicializados os centroides, entretanto o algoritmo implementado no *scikit learning* apresenta as seguintes opções: *kmeans*++, *random* e ndarray (o usuário entre os valores).

O algoritmo atualiza a posição dos centroides seguindo os seguintes passos (Eliasson e Rosén):

Passo 1: Todos os registros $x_1,\ldots,x_n$ são associados ao centroide mais próximo μ_i.

$$S_i = \left\{ x_j : \left\| x_j - \mu_i \right\| \leq \left\| x_j - \mu_c \right\|, \forall 1 \leq c \leq k \right\}$$

Passo 2: Todos os centroides $\mu_1,\ldots,\mu_k$ são atualizados calculando o valor médio de todos os registros de cada *cluster*.

$$\mu_i = \frac{1}{|S_i|}\left(\sum_{x_j \in S_i} x_j \right),\ 1 \leq i \leq k$$

Esses passos são repetidos até que algum critério de parada seja atingido, que costuma ser não haver mais mudanças na posição dos centroides, número máximo de iterações, entre outros.

O algoritmo de Lloyd proposto originalmente não definiu como seria medida a distância. O algoritmo implementado no scikit learning utiliza a distância euclidiana, que é a opção mais comum. Especialmente para mineração de textos, as distâncias de Pearson e Jaccard costumam apresentar melhores resultados que a distância euclidiana.

Além disso, o algoritmo *K-means* costuma ser muito sensível aos valores de inicialização, sendo muitas vezes reiniciar o algoritmo diversas vezes para obter um resultado adequado.

2.5. CLASSIFICAÇÃO DE DOCUMENTOS

Conforme dito anteriormente, a classificação ou categorização de documentos consiste em classificar um documento em um conjunto pré-especificado de categorias. Assim, dado um conjunto de categorias, assuntos ou tópicos e uma coleção de documentos de texto, a classificação é o processo de encontrar o assunto correto associado a cada documento. (Feldman; Sanger, 2006).

Há duas abordagens principais para realizar a classificação de documentos: uma é centrada no conhecimento de especialistas, que é diretamente codificado no sistema (declarativamente ou em forma de regras de classificação) e outra é a aprendizagem por máquinas (*machine learning – ML*), na qual um processo indutivo constrói um classificador com base no aprendizado obtido a partir de exemplos pré-classificados.

Alguns exemplos de aplicação da classificação de textos são a indexação de textos, a classificação e filtragem de documentos, e a classificação de páginas da internet. A indexação de textos consiste em atribuir a cada documento palavras chave a partir de um vocabulário controlado. O problema de classificação de documentos consiste em associar cada documento a uma caixa (categoria). Um exemplo de aplicação desse problema é a classificação de e-mails em principal, social e promoções. A filtragem de textos é um problema de classificação com apenas duas categorias: documentos relevantes ou irrelevantes. Um exemplo de aplicação desse problema é a identificação de e-mails indesejados (*spam*).

Será detalhado aqui, o processo de classificação utilizando aprendizagem por máquina (*machine learning*). Nesse caso, o classificador é construído automaticamente a partir do aprendizado das características de cada categoria obtido com base em documentos de treinamento pré-classificados, sendo, portanto, um processo supervisionado.

A classificação de documentos por aprendizagem por máquinas pode ser dividida em quatro etapas principais: definição das categorias que serão utilizadas para categorizar os registros, disponibilização de um conjunto de treinamento previamente classificado para cada uma das categorias, definição dos atributos que serão utilizados na análise e definição do algoritmo de classificação.

Os algoritmos de classificação de documentos mais comuns são os classificadores probabilísticos, a regressão logística bayesiana, as árvores de decisão, as redes neurais e as máquinas de vetor de suporte (*suport vector machines – SVM*).

Um exemplo de classificador probabilístico é o classificador bayesiano. O classificador bayesiano utiliza o teorema de Bayes para prever a classe de um novo documento, d. Para tanto, ele estima a probabilidade *a posteriori* $P(c_i \mid d)$ para cada classe c_i, e escolhe a classe de maior probabilidade. O teorema de Bayes permite calcular a probabilidade *a posteriori* $P(c_i \mid d)$ a partir da verossimilhança (*likehood*) $P(d \mid c_i)$, definida como a probabilidade de observar o

documento d assumindo que a classe c_i é a classe correta, e da probabilidade *a priori* $P(c_i)$ da classe c_i, como segue:

$$P(c_i \mid d) = \frac{P(d \mid c_i) \cdot P(c_i)}{P(d)}$$

Observe que a probabilidade marginal $P(d)$ não precisa ser calculada, pois ela é a mesma para todas as categorias. Entretanto, o cálculo da verossimilhança $P(d \mid c_i)$ é complexo e, normalmente, é feita a suposição de que os atributos do vetor $d = (w_1, w_2, \ldots)$ são independentes, ou seja,

$$P(d \mid c_i) = \prod_j P(w_j \mid c)$$

O classificador resultante dessa suposição é chamado de classificador *Naïve Bayes* (Bayes "ingênuo"), pois essa suposição nunca se verifica na prática. Observe, por exemplo, que é muito mais provável que a palavra gol apareça em documentos contendo a palavra "futebol" do que em documentos sobre programação em Python. Entretanto, os resultados da utilização desse classificador são compatíveis com os obtidos a partir de modelos que consideram a dependência entre atributos.

O classificador multinomial Naïve Bayes (multinomial NB) implementa o algoritmo Naïve-Bayes para dados com distribuição multinomial e é muito utilizado na classificação de textos, onde os dados são comumente representados por vetores com a contagem de palavras ou vetores com Tf-Idf.

A distribuição é parametrizada pelos vetores $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ para cada classe y, onde n é o número de atributos (no caso de classificação de textos, o tamanho do vocabulário) e θ_{yi} é a probabilidade $P(x_i \mid y)$ do atributo i aparecer em um registro que pertença a classe y.

O parâmetro θ_y é estimado utilizando uma versão atenuada da máxima verossimilhança, a contagem da frequência relativa:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha \cdot n}$$

onde $N_{yi} = \sum_{x \in T} x_i$ é o número de vezes que o atributo i aparece em registros da classe i no conjunto de treinamento T, e $N_y = \sum_{i=1}^{|T|} N_{yi}$ é a quantidade total de todos os atributos na classe y.

O algoritmo de classificação do gradiente de descida (*gradiente descente – GD*) busca a minimização do risco empírico $E_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f(x_i), y_i\right)$, que é uma medida da performance do conjunto de treinamento.

O algoritmo de classificação do gradiente de descida estocástico (*stochastic gradiente descente – SGD*) é uma simplificação do algoritmo do gradiente de descida. Em vez de calcular o gradiente de $E_n(f_w)$ exatamente, cada iteração estima esse gradiente com base em um único exemplo tomado aleatoriamente z_t:

$$w_{t+1} = w_t - \gamma_t \nabla_w Q(z_t, w_t).$$

O processo estocástico $\{w_t, t = 1, \ldots\}$ depende dos exemplos tomados aleatoriamente a cada iteração.

Capítulo 3

METODOLOGIA

Nesse capítulo será descrito o procedimento adotado para analisar os dados do sistema da Nota Carioca, a fim de elaborar uma relação de empresas a serem fiscalizadas ("malha" de fiscalização) com ênfase no problema de indicação indevida de domicílio tributário em outros municípios, a partir da análise do campo "DISCRIMINAÇÃO DOS SERVIÇOS" de notas fiscais de serviço eletrônicas emitidas com o código de serviço 07.19.04 (engenharia consultiva - acompanhamento e fiscalização da execução de obras de engenharia, arquitetura e urbanismo (relacionados com obras de construção civil, hidráulicas, de escoramento ou de contenção de encostas)).

Inicialmente, foi feita uma descrição e caracterização dos dados para permitir uma melhor compreensão do problema em análise. Posteriormente, foi feito o pré-processamento dos dados que consistiu na identificação dos atributos que seriam utilizados, eliminação de linhas com registros nulos, separação das palavras, eliminação de sinais de pontuação e colocação de todas as palavras em letra minúscula.

Em seguida, foram criados dois *dataframes* com as notas fiscais com domicílio tributário no MRJ e em outros municípios, a fim de permitir uma análise comparativa.

A partir da análise dos unigramas (palavras) mais frequentes em cada um dos *dataframes* foi elaborada uma relação das palavras que possuíam mais relevância para a caracterização do tipo de serviço e a frequência dessas palavras passou a ser o objeto da análise.

Na próxima etapa, foi elaborada uma "malha" de fiscalização" pelo método convencional, que consiste em selecionar as maiores empresas (maiores receitas de serviços) dentre aquelas que possuem maior razão entre o valor das notas tributadas fora do município e o valor das notas tributadas no MRJ.

Em seguida, foi utilizado um algoritmo de agrupamento (clusterização) a fim de separar as notas em dois grupos e identificar se ficariam caracterizados um grupo de notas tributadas em outros municípios e outro grupo de notas tributadas no MRJ.

O próximo passo foi a realização da classificação das notas utilizando um classificador multinomial Naïve-Bayes e um classificador SGD (*Stochastic Gradient Descent*). Os classificadores foram treinados com todas as notas adotando-se a indicação de tributada no município ou fora do município como classe. Posteriormente, a informação das classes foi omitida e as notas classificadas pelos classificadores já treinados. Foram, então, identificadas as notas tributadas fora do município e que haviam sido classificadas como tributadas no município pelos dois classificadores. As empresas com a maior incidência de notas nessa situação foram selecionadas para a "malha" de fiscalização.

Finalmente, foi feita uma análise qualitativa das notas das empresas selecionadas na "malha", a fim de identificar se a indicação foi acertada ou não. Além disso, foi feita a comparação da relação de empresas obtida pelo método convencional e a obtida utilizando o procedimento ora proposto.

3.1. DESCRIÇÃO DOS DADOS

Os dados analisados correspondem a Notas Fiscais de Serviço Eletrônicas (NFS-e's ou Notas Cariocas) emitidas nas competências junho de 2010 (implantação do sistema) até dezembro de 2014 com o código de serviço 07.19.04 (engenharia consultiva - acompanhamento e fiscalização da execução de obras de engenharia, arquitetura e urbanismo (relacionados com obras de construção civil, hidráulicas, de escoramento ou de contenção de encostas)) e não canceladas.

Esses dados foram disponibilizados em planilhas de Excel e possuíam os seguintes atributos:

INSCRICAO_PRESTADOR

CPF_CNPJ_PRESTADOR

NOME_PRESTADOR

ENDERECO_PRESTADOR

CEP_PRESTADOR
INSCRICAO_TOMADOR,
CPF_CNPJ_TOMADOR,
NOME_TOMADOR,
ENDERECO_TOMADOR,
CEP_TOMADOR,
REGIME_TRIBUTACAO,
REGIME_ESPECIAL_TRIBUTACAO,
LOCALIDADE_PRESTACAO,
NOTA_FISCAL,
DATA_EMISSAO,
DATA_COMPETENCIA,
NUMERO_RPS,
SERVICO,
DISCRIMINACAO,
VALOR_SERVICOS,
VALOR_DEDUCAO,
VALOR_DESCONTO_INCONDICIONADO,
VALOR_BASE_CALCULO,
ALIQUOTA,
VALOR_ISS,
VALOR_CREDITO,
ISS_RETIDO,
STATUS,
SIMPLES, e
GUIA.

A seguir serão apresentados os valores possíveis de alguns atributos, a fim de esclarecer melhor o seu significado:

REGIME_TRIBUTACAO: “Fora do município”, “No município”, “Imune” e “Isento”.

REGIME_ESPECIAL_TRIBUTACAO: “Art. 33, inc. II, item 8, Lei nº 691/84” (empresa júnior), “Microempreendedor individual (MEI)”, “Microempresa Municipal” e “Sociedade de Profissionais”.

ISS_RETIDO: “0” (não) e “1” (sim)

STATUS: “0” (ativo)

SIMPLES: “1” (optante pelo Simples Nacional) e “0” (não optante pelo Simples Nacional).

A ênfase do presente trabalho é extrair informações do campo DISCRIMINACAO que permitam identificar indícios de que a nota havia sido emitida indevidamente como “tributada fora”. Para esclarecer melhor essa situação, serão apresentados alguns exemplos de valores desse campo, onde informações que eventualmente possam identificar o prestador ou o tomador do serviço foram ocultadas com *.

Exemplo 1: Campo discriminação de uma NFS-e “tributada fora”

PRESTAÇÃO DE SERVIÇOS DE ENGENHARIA DE TRANSITO E PROCEDIMENTOS AO EXCESSO DE VELOCIDADE, AVANÇO DE SINAL E PARADA NA FAIXA DE PEDESTRE DE ACORDO COM O PROCESSO ADMINISTRATIVO Nº ****/2009, EDITAL DA TOMADA DE PREÇOS Nº ***/2011 *** - VALOR UNITÁRIO R$ 36,00

528 MULTAS REF. AO MÊS DE NOV/2014 - VALOR R$ 19.008,00

A análise do exemplo 1 mostra que não se trata de um serviço de acompanhamento ou fiscalização de obra de construção civil e sim engenharia de trânsito, provavelmente instalação de radares de velocidade. Logo essa nota deveria pagar imposto ao MRJ.

Exemplo 2: Campo discriminação de uma NFS-e “tributada fora”

SERVIÇOS PRESTADOS CONFORME CONTRATO **-***** - MODELO II

A análise do exemplo 2 mostra que não é possível tirar nenhuma conclusão desse campo por falta de informações. O texto apresentado no campo DISCRIMINAÇÃO é genérico.

Exemplo 3: Campo discriminação de uma NFS-e “tributada fora”

Medição Referente ao BM *** (16/10/2014 A 15/11/2014)

Elaboração de Projeto Executivo de Engenharia, Serviços de Consultoria de Engenharia, Análise e Comentários de Documentos Técnicos e Propostas de Fornecedores e Assistência Técnica à Obra -ATO, Adequação do Projeto aos Equipamentos Efetivamente Fornecidos (Revisão de Projetos), para o empreendimento denominado "*****", situado no município de ****, Estado ****, a fim de prever a ampliação de capacidade do **** para ****, dando subsídios a construção do empreendimento.

A análise do exemplo 3 mostra que não se trata de serviço de acompanhamento ou fiscalização de obra de construção civil, mas sim de elaboração de projeto e consultoria. Logo essa nota deveria pagar imposto ao MRJ.

Conhecidos os dados, a próxima etapa é prepara-los para a aplicação dos algoritmos e posterior extração de conhecimento.

3.2. O PRÉ-PROCESSAMENTO DOS DADOS

Inicialmente, foram identificados os atributos importantes para a análise em questão, ou seja, aqueles cuja a informação pudesse gerar um indício de emissão indevida do documento fiscal.

Foram selecionados os seguintes atributos para análise:

CPF_CNPJ_PRESTADOR,

CPF_CNPJ_TOMADOR,

CEP_TOMADOR,

LOCALIDADE_PRESTACAO,

REGIME_TRIBUTACAO,

DISCRIMINACAO,

VALOR_SERVICOS, e

ALIQUOTA.

Em seguida foi efetuada uma limpeza dos dados eliminando linhas com registros nulos e uma linha com repetição do cabeçalho.

O próximo passo foi mapear os valores do atributo REGIME_TRIBUTAÇÃO para valores numéricos da seguinte forma:

No município: 0

Fora do município: 1

Isento: 2

A fim de ter uma visão mais clara de como são os dados, eles foram agrupados por regime de tributação.

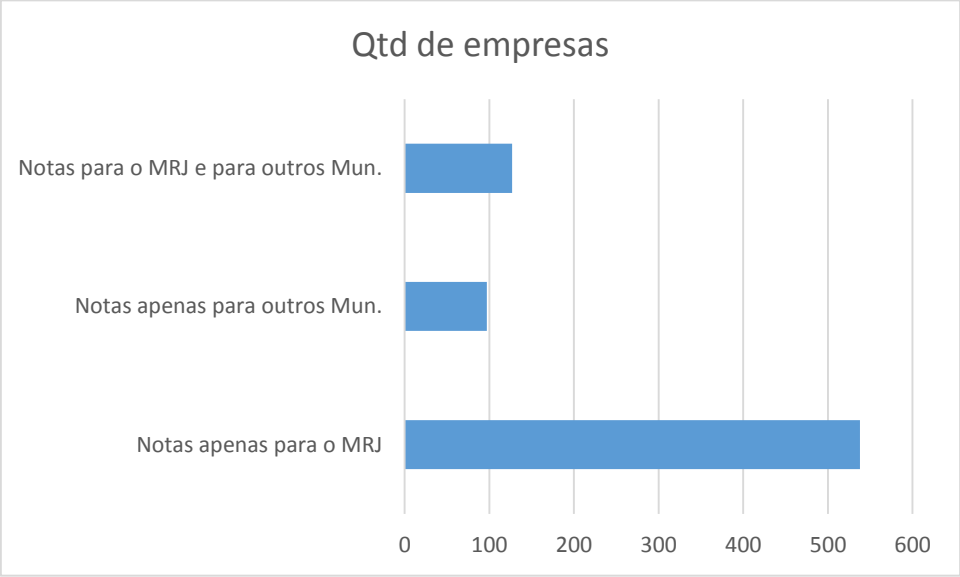

Figura 9: Comparativo entre as empresas pelo domicílio tributário de suas notas.

Observa-se que em um total de 762 empresas, 665 empresas emitem notas para o MRJ e 224 empresas que emitem notas para outros municípios, o que implica que 127 empresas emitem notas tanto para o MRJ quanto para outros municípios, 97 empresas emitem nota apenas para outros municípios e 538 empresas emitem nota apenas para o MRJ.

Esse mesmo resultado permite concluir que foram emitidas 30661 notas eletrônicas com tributação para o MRJ e 20187 notas eletrônicas com tributação para outros municípios. Observe

ainda que o valor total dos serviços das notas com tributação no MRJ foi R$ 1.068.805.293,48 e o valor total dos serviços das notas com tributação em outros municípios foi R$ 2.178.687.609,55.

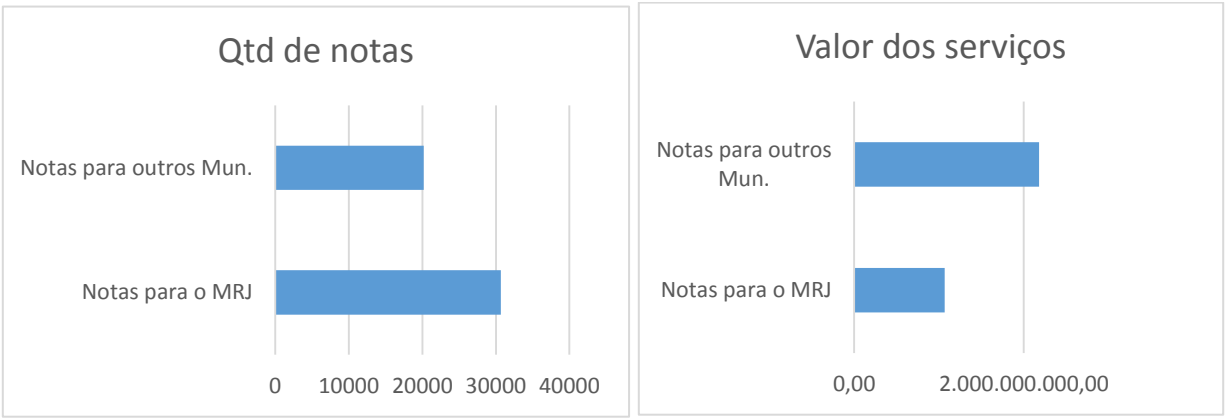

Figura 10: Comparativo entre as notas emitidas com tributação no MRJ e com tributação em outros municípios por quantidade de notas e valor dos serviços

A figura 11 mostra a distribuição das alíquotas nas notas emitidas. Observa-se uma concentração de notas nas alíquotas de 3% (alíquota de construção civil no MRJ) e 5% (alíquota padrão do MRJ).

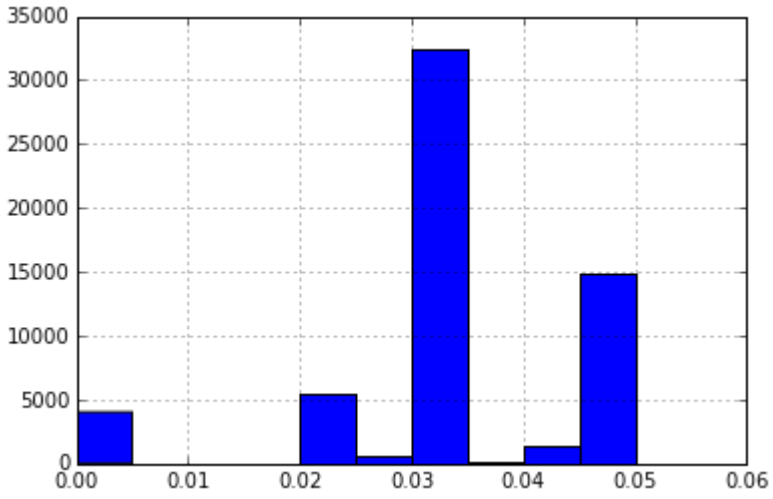

Figura 11: Quantidade de notas emitidas por alíquota.

Para poder comparar as características das notas emitidas com tributação no MRJ com as características das notas emitidas com tributação em outros municípios, foram criados dois *dataframes*, cada um com um dos tipos de nota. O *dataframe* “df_nomum” contém as notas tributadas no MRJ, enquanto o *dataframe* “df_foramun” contém as notas tributadas em outros municípios.

Os textos do campo DISCRIMINAÇÃO de todas as notas tributadas no município e de todas as notas foras do munícipio foram reunidos em dois grandes textos usando o método *aggregate(np.sum)*.

Os sinais de pontuação foram eliminados e todas as palavras colocadas em letras minúsculas nesses dois textos usando a função *elimina_pontuacao* cujo código está transcrito a seguir:

```
def elimina_pontuacao(texto):
        for s in string.punctuation:
                texto = texto.replace(s,’’)
                texto = texto.lower()
        return texto
```

Os n-gramas de cada um dos dois textos foram identificados por meio da função *ngram* e listados os mais comuns utilizando o método *most_common*().

20 unigramas mais frequentes nas notas tributadas no MRJ	**20 unigramas mais frequentes nas notas tributadas fora do município**
[(u'de', 110736), (u'do', 27049), (u'da', 23144), (u'e', 22068), (u'a', 21859), (u'r', 16304), (u'para', 14832), (u'no', 14251), (u'rio', 12652), (u'obra', 12489), (u'n\xba', 12337), (u'vencimento', 11630), (u'servi\xe7os', 10839), (u'banco', 9501), (u'rj', 8420),	[(u'de', 130707), (u'e', 38356), (u'da', 37431), (u'a', 36278), (u'do', 33277), (u'servi\xe7os', 18052), (u'vencimento', 17065), (u'r', 16918), (u'banco', 16725), (u'n\xba', 16549), (u'contrato', 15717), (u'cc', 15021), (u'no', 13982), (u'obras', 13325), (u'para', 13199),

(u'cc', 8026), (u'engenharia', 7229), (u'janeiro', 7073), (u'obras', 6982), (u'rua', 6887)]	(u'valor', 11738), (u'ao', 11481), (u'creditar', 11253), (u'ag', 11217), (u'projeto', 10894)]

Tabela 1: Unigramas mais frequentes

Esse é um ponto crucial do trabalho, pois analisando-se os unigramas (palavras) mais frequentes nos dois textos (Tabela 1), identifica-se que há uma predominância de palavras sem relevância para a determinação do tipo de serviço prestado. Dos 40 unigramas listados, apenas obra, obras, engenharia e projeto trazem informações sobre o serviço prestado.

Sendo assim, optou-se por não utilizar o método tradicional de eliminação de *stopwords* e de palavras muito frequentes e tentou-se buscar alternativas para a caracterização dos documentos.

Uma opção para a seleção de palavras seria a utilização de um *thesaurus* específico para essa área, porém o mais próximo que foi o de "construção" (Jorge, 2004) mostrou-se inadequado.

A opção encontrada foi a elaboração de um dicionário de palavras frequentes e que possuíam relevância para a caracterização do tipo de serviço (Tabela 2). Essas palavras foram extraídas da relação das 400 palavras mais frequentes em cada um dos grupos e escolhidas as mais relevantes para a caracterização do tipo de serviço que estava sendo prestado, com base no conhecimento técnico de um especialista.

Lista de palavras relevantes
abastecimento, acompanhamento, adequação, adutora, agua, ambientais, ambiental, ambiente, ampliação, ampliacao, analise, apoio, apresentação, apresentacao, aprovação, assessoramento, assessoria, atividades, básico, boletim, canal, cei, civil, civis, concessão, concreto, construção, construcao, consultiva, consultoria, contratação, controle, dados, declaratório, desenvolvimento, duplicação, eólico, elaboração, elaboracao, empreendimento, empreitada, energia, engenharia, ensaios, entroncamento, eolico, esgotamento, especializada, especializados, esquadrias, estrada, estudos, etapa, execução, executados, executivo, executivos, exigência, extensão, físico, ferroviária, fisc, fiscal, fiscalização, fiscalizacao, fisico, gerenciamento, gestão, gestao, gestor, habilitada, hotel, implantação, implantacao,

infraestrutura, integração, linha, medição, medicao, melhoramento, monitoramento, montagem, obra, obras, operação, parques, pavimentação, pista, planejamento, porto, programa, programas, projeto, projetos, qualidade, recuperação, rede, refinaria, relatório, restauração, rodoviário, rodovia, rodovias, sanitário, sanitario, sistema, sistemas, subsolo, supervisão, supervisao, técnica, técnico, técnicos, tecnica, tecnico, tecnicos, tecnológico, tecnologico, terminal, transmissão, transmissao, transportes, via, vidros, vistoria, volantes

Tabela 2: Lista de palavras relevantes para caracterização do tipo de serviço prestado.

Utilizando a função *filtra_palavras_uteis*, cujo código está descrito abaixo, os dois textos foram transformados em duas listas das palavras selecionadas, onde cada palavra aparece na lista tantas vezes quantas aparecia no texto original.

```
def filtra_palavras_uteis(texto, uteis):
    lista = texto.split()
    lista = [l.lower().strip(string.punctuation) for l in lista]
    lista = [l for l in lista if l in uteis]
    texto = ' '.join(lista)
    return texto
```

Com base nessas duas listas, foram identificadas as palavras relevantes mais frequentes originárias de notas tributadas no MRJ e de notas tributadas em outros municípios (Tabela 3).

20 unigramas mais frequentes dentre as palavras relevantes nas notas tributadas no MRJ	**20 unigramas mais frequentes dentre as palavras relevantes nas notas tributadas fora do município**
[(u'obra', 12489), (u'engenharia', 7229), (u'obras', 6982), (u'projeto', 5923), (u'dados', 5295), (u'cei', 4550), (u'consultoria', 3718), (u'gerenciamento', 3401), (u'medi\xe7\xe3o', 3020), (u'projetos', 2687),	[(u'obras', 13325), (u'projeto', 10894), (u'medi\xe7\xe3o', 9827), (u'gerenciamento', 9141), (u'engenharia', 7669), (u'obra', 6447), (u'supervis\xe3o', 5619), (u'cei', 4160), (u'fiscaliza\xe7\xe3o', 3942), (u'ambiental', 3932),

(u'subsolo', 2536), (u'fiscaliza\xe7\xe3o', 2488), (u'medicao', 2277), (u'acompanhamento', 2275), (u't\xe9cnica', 2232), (u'consultiva', 2092), (u'gest\xe3o', 1976), (u't\xe9cnico', 1931), (u'ensaios', 1796), (u'tecnol\xf3gico', 1712)]	(u'medicao', 3259), (u't\xe9cnico', 2883), (u'rodovia', 2833), (u'gest\xe3o', 2820), (u'planejamento', 2702), (u'apoio', 2668), (u'consultiva', 2608), (u'controle', 2550), (u'implanta\xe7\xe3o', 2517), (u'execu\xe7\xe3o', 2516)]

Tabela 3: Unigramas mais frequentes dentre as palavras relevantes

Essa contagem de unigramas permitiu a elaboração de dois *wordclouds* ("nuvens de palavras", onde importância de cada palavra está associada a seu tamanho ou cor) para as notas tributadas no MRJ e tributadas em outros municípios (Figuras 12 e 13).

Figura 12: Wordcloud das palavras mais relevantes nas notas tributadas no MRJ.

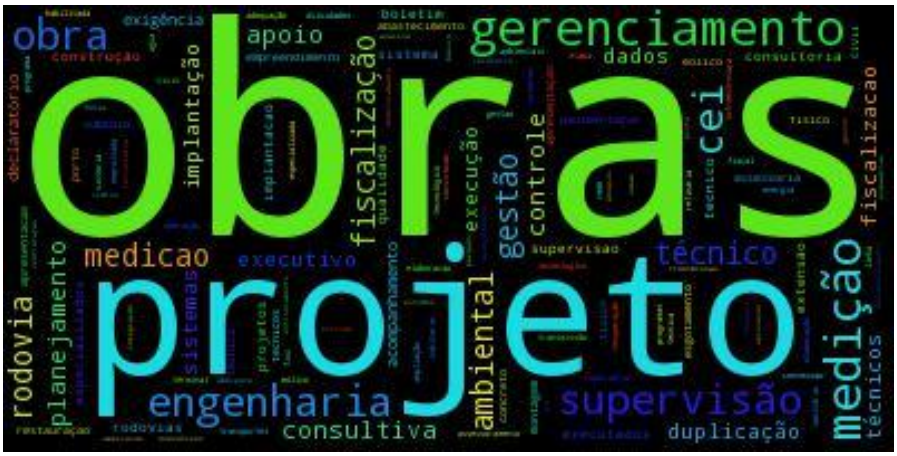

Figura 13: Wordcloud das palavras mais relevantes nas notas tributadas em outros municípios.

Observe que nas notas tributadas em outros municípios a palavra projeto aparece em destaque, ou seja, é uma palavra frequente. Entretanto, isso é um indício de que haja notas incorretas, visto que o projeto de engenharia não desloca o domicílio tributário, apenas o gerenciamento e fiscalização de obras.

3.3. ELABORAÇÃO DE RELAÇÃO DE EMPRESAS SUSPEITAS PELO MÉTODO "CONVENCIONAL"

Nessa etapa foi elaborada uma relação de empresas suspeitas pelo método atualmente adotado no MRJ para a programação das fiscalizações, tendo como foco o domicílio tributário no código de serviços 07.19.04.

Os critérios determinantes para a inclusão da empresa na relação de empresas suspeitas seriam o valor da receita de serviços e a relação entre o valor dos serviços prestados para outros municípios e os prestados para o MRJ. A adoção do valor da receita de serviços como critério justifica-se porque empresas com maior porte resultam maiores valores de sonegação e maior recuperação de valores na ação fiscal. A razão entre o valor dos serviços prestados para outros municípios e os prestados para o MRJ muito elevada seria um indício de que a empresa poderia estar indicando como tributável fora do município valores que deveriam ser tributados no MRJ. Observe que nenhum desses critérios está relacionado diretamente a um erro na indicação do domicílio tributário. Na prática, pode haver empresas grandes que prestam muito mais serviços em outros municípios do que no MRJ e entrariam na relação de empresas suspeitas, mesmo estando corretas. Por outro lado, uma empresa de menor porte e que devesse recolher todo seu imposto no MRJ, mas indicasse uma parcela desse indevidamente para outros municípios, teria uma boa chance de não entrar na relação de empresas suspeitas.

As empresas candidatas a entrar na relação de empresas suspeitas são as 224 que emitem notas tributáveis em outros municípios. Nesse conjunto, serão selecionadas 22 empresas, ou seja, cerca de 10%, seguindo os dois critérios descritos anteriormente.

A fim de identificar essa relação de empresas foi criado um novo *dataframe* “df_ambos” cujas colunas são o CNPJ do prestador, o valor dos serviços tributados no MRJ, o valor dos serviços tributados em outros municípios e a razão o valor dos serviços tributados fora e o valor dos serviços tributados no MRJ (razão fora-dentro). Esses dados foram ordenados em ordem decrescente da razão fora-dentro. Foram selecionados os 40 primeiros registros a fim de selecionar as 22 empresas que comporiam a relação de empresas suspeitas, levando em consideração também o valor dos serviços tributados em outros municípios.

Foram selecionadas as seguintes empresas identificadas apenas pelo seu sequencial na relação de 40 empresas a fim de preservar sua identidade: 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 14, 16, 19, 23, 24, 26, 29, 30, 35, 36, 38.

Essa relação será posteriormente comparada com a obtida com auxílio das técnicas de mineração de textos.

3.4. AGRUPAMENTO DOS DOCUMENTOS

A aplicação de algoritmos de agrupamento e classificação será utilizada para extrair conhecimento a partir dos textos do campo DISCRIMINAÇÃO dos documentos fiscais, que já foram devidamente pré-processados.

O objetivo nessa etapa do trabalho é separar os documentos fiscais em dois grupos e identificar se é possível associar um dos grupos a notas que deveriam ser tributadas no MRJ e o outro grupo a notas que deveriam ser tributadas em outros municípios. Havendo dois grupos bem definidos e associados a essas características, é esperada uma maior eficácia nos resultados dos algoritmos de classificação que serão utilizados posteriormente.

O primeiro passo foi construir o vetor de atributos a partir do campo DISCRIMINAÇÃO de cada um dos documentos fiscais, usando o modelo *bag of words*. Primeiro utilizou-se a função *filtra_palavras_uteis*, já descrita na seção 3.2, para transformar cada texto em uma lista das palavras relevantes nele contidas, repetidas tantas vezes quantas elas aparecerem. Sobre esses textos foram utilizados o *TfidfVectorize*r e o método *fit_transform* do pacote *scikit learning*, que retornam a matriz de termos e documentos, composta pelo Tf-Idf de cada termo (cada uma das palavras relevantes escolhidas para o nosso vocabulário).

Construída a matriz de termos e documentos, os dados estão prontos para a aplicação do algoritmo de agrupamento. Será aplicado o algoritmo *K-means* (também do pacote *scikit learning*), descrito em 2.3, com $k = 2$ tendo em vista que espera-se a formação de dois *clusters*: um representativo das notas tributadas no MRJ e outro das notas tributadas em outros municípios.

Para inicialização dos centroides foi adotado o método *k-means*++, que busca acelerar a convergência.

Os 30 termos mais frequentes em cada um dos clusters foram obtidos, a fim de identificar as características de cada grupo.

3.5. CLASSIFICAÇÃO PARA DETECÇÃO DE EMPRESAS SUSPEITAS

Foram utilizados dois classificadores, um multinomial Naïve-Bayes e um SGD, a fim de detectar notas mal classificadas.

Foi gerada novamente a matriz de termos e documentos com os valores do Tf-Idf de cada termo pela aplicação sequencial dos métodos *CountVectorizer* e *TfidfTransformer* sobre os dados agrupados por regime de tributação.

Foram treinados os dois classificadores (SGDClassifier e MultinomialNB) com o método *fit* usando os dados classificados em “No município” e “Fora do município”. Depois os classificadores foram utilizados para classificar esse mesmo conjunto de dados (sem classificação prévia).

Como a intenção é detectar discrepâncias na emissão das notas e partindo do pressuposto que as notas erradas são bem menos do que as corretas, os classificadores treinados tendem a aprender a identificar as notas “corretamente”. Sendo assim, ao classificar o mesmo conjunto de dados novamente, espera-se que as notas que estavam indevidamente na classe “Fora do município” sejam classificadas na classe “No município”.

Foi, então, construída uma relação de empresas suspeitas, considerando as empresas que possuíam maior quantidade de notas tributadas “Fora do município”, mas que ambos os classificadores classificaram como “No município”.

Capítulo 4

RESULTADOS

Nesse capítulo vamos analisar os resultados obtidos nos procedimentos descritos nas seções 3.3 (Elaboração de relação de empresas suspeitas pelo método "convencional"), 3.4 (Agrupamento dos documentos) e 3.5 (Classificação para detecção de empresas suspeitas).

Inicialmente, vamos analisar o resultado do agrupamento realizado, conforme procedimento descrito na seção 3.4. O algoritmo *K-means* foi executado com $k = 2$, a fim de identificar os dois *clusters* correspondentes às notas tributadas no MRJ e em outros municípios. As 30 palavras mais frequentes em cada um dos clusters foram contadas e estão apresentadas na Tabela 4 por ordem decrescente de frequência.

É importante relembrar a definição dos serviços de engenharia consultiva, que consta no artigo 43, Seção II, Capítulo IX do Decreto nº 10.514 de 09 de outubro de 1991, transcrito a seguir:

> Art. 43 - Os serviços de engenharia consultiva, para os efeitos do disposto no inciso II, item 1, do art. 19, são os seguintes:
>
> I - elaboração de planos diretores, estudos de viabilidade, estudos organizacionais e outros, relacionados com obras e serviços de engenharia;
>
> II - elaboração de anteprojetos, projetos básicos e projetos executivos para trabalhos de engenharia;
>
> III - fiscalização e supervisão de obras e serviços de engenharia.
>
> Parágrafo único - O tratamento fiscal previsto no *caput* deste artigo destina-se exclusivamente aos serviços de engenharia consultiva que

> estiverem relacionados com obras de construção civil, hidráulicas, de escoramento e de contenção de encostas.

Além do fato de que o domicílio tributário desse serviço é o local do estabelecimento do prestador no caso dos incisos I e II, e no caso do inciso III, o local da prestação do serviço (local da obra). Pode-se identificar que palavras como planos, estudos, projetos, consultoria, planejamento, elaboração, assessoria, por exemplo, são palavras indicativas de serviços que deveriam ser tributados no MRJ. Essas palavras serão grifadas na Tabela 4.

Top terms per cluster	
Cluster 0	Cluster 1
obras	**consultoria**
gerenciamento	técnica
obra	obra
projeto	engenharia
engenharia	acompanhamento
fiscalização	**assessoria**
medição	dados
cei	**projeto**
dados	**projetos**
supervisão	**elaboração**
consultiva	medição
gestão	esquadrias
execução	sistema
técnico	tecnica
controle	vidros
medicao	cei
planejamento	**planejamento**
acompanhamento	gestão
ambiental	implantação
subsolo	obras
projetos	sistemas
fiscalizacao	apoio

apoio	técnico
duplicação	civil
técnicos	execução
empreendimento	executivo
tecnológico	gerenciamento
executivo	construção
concreto	empreendimento
implantação	fisico

Tabela 4: 30 palavras mais frequentes em cada um dos dois clusters.

Observa-se que no *cluster* 0 há menos palavras grifadas (3) e a primeira grifada aparece na quarta posição. Além disso, palavras como gerenciamento e fiscalização aparecem entre as primeiras posições, indicando uma preponderância de serviços do inciso III que podem ser tributados fora do município. Por outro lado, o *cluster* 1 apresenta na primeira posição a palavra consultoria e uma grande quantidade de palavras grifadas (6), o que indica uma preponderância de serviços dos incisos I e II que deveriam ser tributados no MRJ.

Dessa forma, conclui-se que o *cluster* 0 deve estar associado a notas que são tributadas fora do município, enquanto o *cluster* 1, a notas tributadas no MRJ.

O próximo passo é comparar a relação de empresas obtida pelo método convencional nas seções 3.3 e a obtida por meio dos classificadores na seção 3.4.

Na Tabela 5 são mostradas as 40 empresas selecionadas pelo método convencional e estão sombreadas as que foram selecionadas para a “malha” (22 empresas) com base no valor da razão fora-dentro e no valor dos serviços prestados para outros municípios, e as 30 empresas selecionadas pelos classificadores que apresentaram a maior quantidade de notas mal classificadas. Os CNPJ’s das empresas foram substituídos por um sequencial para preservar a sua identidade.

Inicialmente, observa-se que as duas relações possuem 6 empresas em comum. A fim de avaliar a qualidade das indicações da relação obtida a partir dos classificadores, foi feita uma análise por amostragem das notas dessas empresas com o intuito de identificar se realmente há indícios de erro na indicação do domicílio tributário, cujo resultado está na Tabela 5. O valor na coluna “CF” indica se a empresa realmente possui indícios de emissão indevida de notas fiscais.

Observe que, a análise das notas por um especialista indicou que, das 30 empresas, 16 realmente deveriam ser fiscalizadas. Um índice de acerto próximo a 50%. Cabe frisar que, mesmo o índice não sendo alto, essa seleção possibilitou que a análise humana se tornasse viável. O universo inicial de empresas era de 768 com 58 mil notas. Foi necessário analisar apenas algumas notas de 30 empresas. Além disso, essas empresas não teriam sido vislumbradas no método de seleção convencional.

Seq.	Valor dos serviços tributados no MRJ	Valor dos serviços tributados em outros municípios	Razão fora-dentro
1	1.000,00	294.861,10	294,9
2	2.800,00	490.040,00	175,0
3	425.472,57	66.658.140,00	156,7
4	28.371,04	2.015.919,00	71,1
5	10.064,52	472.708,00	47,0
6	144.404,76	6.656.626,00	46,1
7	29.000,00	1.166.368,00	40,2
8	776.812,22	28.303.820,00	36,4
9	3.121.610,08	102.855.200,00	32,9
10	2.283,22	74.137,74	32,5
11	2.133,43	63.068,56	29,6
12	2.622.729,88	72.618.210,00	27,7
13	24.949,99	563.614,60	22,6
14	3.776.566,37	57.375.580,00	15,2
15	1.075,00	15.050,36	14,0
16	32.000,00	419.718,30	13,1
17	8.500,00	108.800,60	12,8
18	12.300,00	155.970,00	12,7
19	2.545.943,67	28.910.530,00	11,4
20	29.520,40	326.402,40	11,1
21	37.500,00	368.437,00	9,8
22	15.250,00	138.129,90	9,1
23	95.080.223,69	801.406.900,00	8,4
24	1.663.415,23	13.968.990,00	8,4
25	2.405,00	20.000,00	8,3
26	347.511,82	2.701.495,00	7,8
27	56.100,00	361.150,90	6,4
28	2.400,00	15.306,52	6,4
29	17.260.679,95	99.125.960,00	5,7
30	136.667,03	726.055,90	5,3
31	112.103,14	544.096,70	4,9
32	87.000,00	365.500,00	4,2
33	5.500,00	22.866,67	4,2
34	182.793,57	704.108,20	3,9
35	3.433.865,68	13.053.400,00	3,8
36	778.654,95	2.648.144,00	3,4
37	18.980,00	62.842,11	3,3
38	2.370.863,83	7.843.508,00	3,3
39	15.500,00	50.000,00	3,2
40	19.200,00	55.530,00	2,9

Seq.	Qtd de notas mal classificadas	CF.
41	601	sim
42	501	não
23	431	sim
43	278	não
44	160	sim
45	159	não
46	153	sim
47	148	sim
48	135	sim
49	131	sim
50	130	não
51	89	não
29	85	sim
52	83	não
24	69	sim
53	55	sim
30	50	não
54	49	não
55	48	sim
56	48	não
3	46	sim
57	42	não
58	41	não
59	40	sim
60	39	não
61	38	não
62	38	sim
21	37	não
63	36	sim
64	35	sim

Tabela 5: Empresas selecionadas pelo método convencional e pelos classificadores.

Capítulo 5

CONCLUSÃO E CONSIDERAÇÕES FINAIS

Esse trabalho representou uma primeira tentativa de extração de informações a partir do campo discriminação dos serviços da Nota Carioca.

Foi obtida uma relação de empresas suspeitas, utilizando-se técnicas de mineração de texto e de classificação, que teve um índice de acertos superior a 50% e, além disso, reduziu o universo de análise a um número viável para um trabalho não automatizado.

A possibilidade de elaborar uma "malha" de fiscalização mais acurada permite a racionalização no uso de recursos humanos e aumenta a probabilidade de recuperação de receitas. Tratando especificamente do serviço de engenharia consultiva, pode-se estimar a possibilidade de recuperação de ISS da ordem de 12 milhões de reais. O problema geral de domicílio tributário (para todos os códigos de serviço) levaria a recuperação de ISS próximo a 260 milhões de reais. Além disso, essa metodologia pode ser aplicada a outros problemas, como por exemplo a utilização indevida de alíquotas reduzidas e benefícios fiscais.

Um aspecto que dificultou a realização de uma análise mais apurada foi não ser possível pré-classificar algumas notas fiscais a fim de prover um conjunto de treinamento para um método supervisionado. Isso ocorreu porque muitas notas têm textos genéricos que não permitem extrair informação relevante, pois não há uma obrigação de que se faça uma descrição adequada do serviço que foi prestado.

Apesar dos bons resultados obtidos, há diversas possibilidades de melhorias em trabalhos futuros que aumentariam a acurácia do método, como por exemplo o refinamento do vocabulário utilizado na análise (lista de palavras relevantes), o acompanhamento da série histórica, a possibilidade de ter um conjunto de teste validado após a fiscalização das empresas selecionadas

neste trabalho, a agregação de outros atributos e informações colaterais à análise e a aplicação de métodos mais elaborados de classificação.

REFERÊNCIAS BIBLIOGRÁFICAS

Brasil. Lei Complementar nº 116, de 31 de julho de 2003. Dispõe sobre o Imposto Sobre Serviços de Qualquer Natureza, de competência dos Municípios e do Distrito Federal, e dá outras providências. Diário Oficial da União, de 01 de agosto de 2003.

Rio de Janeiro (Município). Lei nº 691, de 24 de dezembro de 1984. Aprova o Código Tributário do Município do Rio de Janeiro e dá outras providências. Diário Oficial do Município do Rio de Janeiro, de 26 de dezembro de 1984.

Rio de Janeiro (Município). Lei n° 5.098, de 15 de outubro de 2009. Institui a Nota Fiscal de Serviços Eletrônica e dá outras providências. Diário Oficial do Município do Rio de Janeiro, de 16 de outubro de 2009.

Rio de Janeiro (Município). Decreto n° 32.250, de 11 de maio de 2010. Dispõe sobre a Nota Fiscal de Serviços Eletrônica – NFS-e – NOTA CARIOCA – e dá outras providências. Diário Oficial do Município do Rio de Janeiro, de 12 de maio de 2010.

Rio de Janeiro (Município). Decreto nº 10.514, de 09 de outubro de 1991. Regulamenta as disposições legais relativas ao Imposto sobre Serviços de Qualquer Natureza. Diário Oficial do Município do Rio de Janeiro, de 09 de outubro de 1991.

Rio de Janeiro (Município). Resolução SMF n° 2.617 de 17 de maio de 2010. Dispõe sobre os procedimentos relativos à emissão da Nota Fiscal de Serviços Eletrônica – NFS-e – NOTA CARIOCA e dá outras providências. Diário Oficial do Município do Rio de Janeiro, de 18 de maio de 2010.

Andrade, H. S. (2009). Um processo de mineração de dados aplicado ao combate à sonegação fiscal do ICMS. Tese de M.Sc. Universidade Estadual do Ceará.

Aranha, C. e Passos, E. (2006). A Tecnologia de Mineração de Textos. RESI – Revista Eletrônica de Sistemas de Informação, nº 2.

Bird, S.; Klein, E. e Loper, E. (2009) Natural Language Processing with Python. O'Reilly Media, Inc.

Boavista, J. M. S. e Silva, F. S. (2011a). Nota Carioca: uma avaliação preliminar da implantação. Nota Técnica Nº 1. SMF-RJ.

Boavista, J. M. S. e Silva, F. S. (2011b). Nota Carioca: um ano de implantação. Nota Técnica Nº 2. SMF-RJ.

Boavista, J. M. S. e Silva, F. S. (2013). Nota Carioca: impactos financeiros diretos após 3 anos de implantação. Nota Técnica Nº 3. SMF-RJ.

Bottou, L. (2010) Large-Scale Machine Learning with Stochastic Gradient Descent. *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pgs. 177–187, Paris, France, Springer.

Carvalho, S. C. (2014). MIDCS – Um Método de Inferência Difusa para Classificação de Sonegadores Fiscais. 12º Prêmio de Criatividade e Inovação da RFB.

Coelho, A. R. (2007). Stemming para Língua Portuguesa: estudo, análise e melhoria do algoritmo RSLP. Tese de M.Sc., Universidade Federal do Rio Grande do Sul.

Coelho, E. M. P. (2012). Ontologias difusas no suporte à mineração de dados: aplicação na Secretaria de Finanças da Prefeitura Municipal de Belo Horizonte. PhD thesis. Universidade Federal de Minas Gerais.

Cordeiro, A. D. (2005). Gerador Inteligente de Sistemas com Auto-aprendizagem para Gestão de Informações e Conhecimento. PhD thesis, Universidade Federal de Santa Catarina, Departamento de Engenharia da Produção.

Côrrea, G.N., Marcacini, R. M. e Rezende, S. O. (2012). Uso da mineração de texto na análise exploratória de artigos científicos. Relatórios Técnicos do ICMC nº 383. São Carlos.

Eliasson, P. e Rosén, N. Eficiente K-means clustering and the importance of seeding. BSc. Thesis. KTH Computer Science and Communication.

Fayyad, U.; Piatetsky-Shapiro, G. e Padhraic, S. (1996) From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, pgs. 37 a 54.

Feldman, R. e Sanger, J. (2006). The Text Mining Handbook: Advanced Approaches in Analysing Unstructured Data. Cambridge University Press.

Hahn, U. e Mani, I. (2000). The challenges of Automatic Summarization. IEEE Computer, Vol. 33, No. 11.

Halkidi, M., Batistakis, Y. e Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems, 17, pgs. 107 a 145.

Jain, A. K. e Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall, Upper Saddle River, NJ, USA.

Janert, P. K. (2011). Data Analysis with Open Source Tools. O'Reilly Media, Inc., pgs. 293 a 325.

Jorge, V. (2004). Thesaurus da Língua Portuguesa do Brasil. Montreal, CA. Disponível em: http://alcor.concordia.ca/~vjorge/Thesaurus/indices.html. Acesso em: 24/09/2015.

Lopes, M. C. S. (2004). Mineração de dados textuais utilizando técnicas de clustering para o idioma português. PhD thesis, Universidade Federal do Rio de Janeiro.

MacQuenn, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. e Neyman, J. editors, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pgs. 281-297. University of California Press.

Manning, C. D.; Raghavan, P. e Schütze, H. (2014). An Introduction to Information Retrieval. Cambridge University Press.

Matsubara, E. T.; Martins, C. A. e Monard, M. C. (2003). PreText: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. Relatório Técnico do ICMC nº 209. São Carlos.

Morais, E. A. M. e Ambrósio, A. P. L. (2007) Mineração de Textos. Relatório Técnico. Universidade Federal de Goiás.

Orengo, V. M. e Huyck, C. (2001). A Stemming Algorithm for Portuguese Language. Em: Symposium on String Processing and Information Retrieval, 8. Chile.

Passini, M. L. C. (2012). Mineração de Textos para Organização de Documentos em Centrais de Atendimento. Dissertação de Mestrado, Universidade Federal do Rio de Janeiro.

Porter, M. F. (2005). The stemming algorithm. Disponível em http://snowball.tartarus.org/algorithms/portuguese/stemmer.html. Acesso em 21 de setembro de 2015.

Segaran, T. (2007). Programming Collective Intelligence. O'Reilly Media, Inc.

Soares, M .V. B.; Prati, R. C. e Monard, M. C. (2008). Pretext ii: Descrição da reestruturação da ferramenta de pré-processamento de textos. Relatório Técnico do ICMC nº 333. São Carlos.

Steinbach, M., Karypis, G. e Kumar, V. (2000). A comparison of document clustering techniques. In KDD'2000: Workshop on Text Mining, pgs. 1-20.

Tan, A.-H. (1999). Text mining: "The state of the art and the challenges." In Proceeddings, PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, Beijing, pgs. 65 a 70.

Vendramin, L. Campello, R. J. G. B. e Hruschka, E. R. (2010) Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining, 3(4), pgs. 209-235.

Wives, L. K. (2002). Tecnologias de Descoberta de Conhecimento em Textos Aplicadas à Inteligência Competitiva. Exame de Qualificação. Universidade Federal do Rio Grande do Sul.

Xu, R. e Wunsch, D. (2008). Clustering. Wiley-IEEE Press, IEEE Press Series on Computational Intelligence.

Zaki, M. J.; Meira Jr.; W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.