eib0885838909

Descoberta de conhecimento com uso de técnicas de mineração de textos aplicadas em documentos textuais da investigação policial brasileira

Marcio Ponciano da Silva*
Angel Freddy Godoy Viera**

Artículo recibido:
17 de diciembre de 2020
Artículo aceptado:
5 de abril de 2021

Artículo de investigación

Resumo

O objetivo deste estudo é analisar como técnicas de mineração de textos aplicadas em documentos textuais da investigação policial brasileira pode promover descoberta de conhecimento. A pesquisa coletou documentos da investigação policial e submeteu ao processo de mineração de textos. O estudo utilizou as técnicas de *case folding*, tokenização, *stopwords* personalizada, *bag of words* e TF-IDF para extrair resultados em *n-grams*. Os resultados foram apresentados com *word clouds*. Na pesquisa foi usado o *k-means* para clusterizar os conjuntos de trigramas, identificando em cada clusters os termos mais representativos dos clusters. O uso de

* Universidade Federal de Santa Catarina, Brasil mponcianos@gmail.com
** Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Brasil a.godoy@ufsc.br

técnicas de mineração de texto sobre esses documentos teve como propósito a extração de conhecimento não trivial. As técnicas de mineração de texto, ou descoberta de conhecimento em base de dados textual, tem a finalidade de descobrir padrões não observáveis quando analisados por manipulação humana de grande volume de documentos. Os resultados encontrados favoreceram a descoberta de conhecimentos na identificação de entidades e conexões, como também categorias temáticas da investigação.

Palavras chave: Investigação Policial; Descoberta de Conhecimento; Mineração de Textos

Descubrimiento de conocimientos mediante técnicas de minería de textos aplicadas a documentos textuales de la investigación policial brasileña

Marcio Ponciano da Silva y *Angel Freddy Godoy Viera*

Resumen

El objetivo de este estudio es analizar cómo las técnicas de minería de textos aplicadas a documentos textuales de la investigación policial brasileña pueden promover el descubrimiento de conocimiento. La investigación recopiló documentos de la investigación policial y los sometió al proceso de minería de textos. El estudio utilizó las técnicas de plegado de casos, tokenización, palabras vacías personalizadas, bolsa de palabras y TF-IDF para extraer los resultados en n-gramas. Los resultados se presentaron con nubes de palabras. En la investigación, se utilizaron k-medias para agrupar los conjuntos de trigramas, identificando en cada grupo los términos más representativos de los grupos. El uso de técnicas de minería de textos en estos documentos tenía como objetivo extraer conocimientos no triviales. Las técnicas de minería de texto, o descubrimiento de conocimiento en una base de datos textual, tienen el propósito de descubrir patrones inobservables cuando se analizan mediante manipulación humana de grandes volúmenes de documentos. Los resultados encontrados favorecieron el descubrimiento de conocimientos en la identificación de entidades y conexiones, así como categorías temáticas de la investigación.

Palabras clave: Investigación Policial; Descubrimiento del Conocimiento; Extracción de Textos

Discovery of knowledge using text mining techniques applied to textual documents of Brazilian police investigation

Marcio Ponciano da Silva and *Angel Freddy Godoy Viera*

Abstract

The aim of this study is to analyze how text mining techniques applied to textual documents of Brazilian police investigation can promote knowledge discovery. The research collected documents from the police investigation and submitted them to the text mining process. The study used the techniques of case folding, tokenization, custom stopwords, bag of words and TF-IDF in order to extract results in n-grams. The results were presented with word clouds. In the research, k-means were used to cluster the sets of trigrams, identifying in each clusters the most representative terms of the clusters. The use of text mining techniques on these documents was intended to extract non-trivial knowledge. The techniques of text mining, or discovery of knowledge in a textual database, have the purpose of discovering unobservable patterns when analyzed by human manipulation of large volumes of documents. The results found favored the discovery of knowledge in the identification of entities and connections, as well as thematic categories of the investigation.

Keywords: Police Investigation; Discovery of Knowledge; Text Mining

Introdução

O processo de descoberta de conhecimento tem sido empregado para encontrar informações ou padrões de informações, utilizando técnicas de análise e extração de dados. A descoberta de conhecimento em documentos textuais está inserida no contexto da descoberta de conhecimento em dados não estruturados.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389

Este estudo coletou documentos textuais produzidos no processo de investigação policial, com propósito de submete-los a técnicas de mineração de textos para extrair conhecimento não observáveis apenas pela análise humana, devido ao volume de dados. Os resultados apresentados neste estudo fazem parte de trabalho de pesquisa de mestrado.

Esses documentos compõem volumes de inquérito policial, que no Brasil é um procedimento administrativo escrito, resultado da investigação policial e da coleta de provas com a finalidade de apurar a infração criminal. O inquérito é um instrumento que as polícias brasileiras utilizam na fase de investigação criminal para documentar a apuração de diligências produzidas em uma investigação de crime (M. Silva, 2019).

Por ano, a Polícia Federal Brasileira produz cerca de 70 mil inquéritos policiais. O somatório ao ano alcança 14 milhões de páginas de textos (Agência, 2016). Os documentos textuais do inquérito policial recebem o nome de "peças".

O objetivo geral deste estudo é investigar quais técnicas de mineração de textos oferecem melhorias à investigação policial, examinando algumas dessas técnicas e aplicando-as sobre o conjunto de documentos textuais do inquérito policial, para extrair conhecimento não trivial que possa contribuir com o processo de investigação. O estudo propõe a análise de abordagem da Ciência da Informação, por meio do uso de métodos e técnicas da Recuperação de Informação sobre as Ciências Policiais, a partir do acervo textual do inquérito policial.

Essa aproximação busca agregar conhecimento ao processo das polícias judiciárias. Borko (1968: 2) tratou essa relação simbiótica, tendo proposto que técnicas e procedimentos empregados por bibliotecários e documentaristas deveriam se basear nos resultados teóricos da Ciência da Informação. Este estudo tem aplicação de relevância social, vez que contribui com a sociedade propondo um resultado mais eficaz à investigação policial, por meio do uso de técnicas de mineração de textos.

Revisão da literatura

Esta pesquisa utilizou o método de pesquisa bibliográfica, elaborando uma revisão da literatura sobre o tema abordado. Essa revisão de literatura buscou principalmente trabalhos relacionados ao tema da mineração de textos.

O tema da mineração de textos está relacionado ao crescente uso de dados textuais. Segundo Zhou, Peng e Liu (2010), o enorme crescimento de documentos textuais aumentou a demanda por novos métodos de mineração de

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389

dados para o processamento de texto. Por isso, a mineração de textos pode se chamar de descoberta de conhecimento de bases de dados textuais (Aranha e Passos, 2006).

Como os dados de texto codificam grande parte de nosso conhecimento acumulado, eles geralmente não podem ser descartados. Como consequência, gera o acúmulo de uma grande quantidade de dados que agora estão além da capacidade de qualquer indivíduo examinar (Zhai e Massung, 2016: 4).

A descoberta de conhecimento em documentos textuais está inserida no contexto da descoberta de conhecimento em dados não estruturados. O processo de descoberta de conhecimento tem pelo menos quatro fases: a coleta de dados, o pré-processamento de dados, a mineração e o pós-processamento (L. Silva, Peres e Boscarioli, 2017: 11). A *Figura 1* apresenta a ordem dessas etapas.

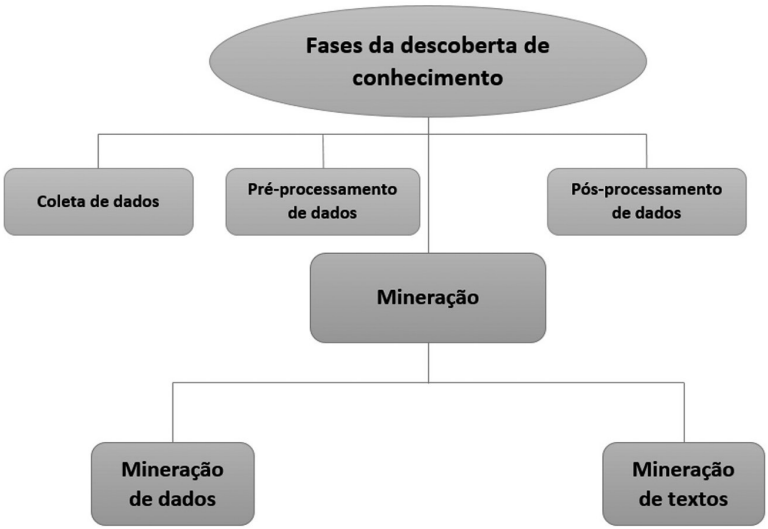

Figura 1. Etapas da mineração ou descoberta de conhecimento

A coleta de dados (textos) é a primeira etapa da mineração de textos. De acordo com Weiss *et al.* (2010: 8) o processo de recuperação de textos inicia com uma coleção de documentos. A etapa que se segue é o pré-processamento, onde são eliminados os fatores dependentes da linguagem para que a estrutura da linguagem se torne mais clara (Yang, Manoharan e Barber, 2014: 52).

A etapa de pré-processamento envolve a preparação dos dados antes de aplicar técnicas de mineração. Na preparação dos dados textuais o uso da técnica conhecida por *case folding* é usada para garantir a correspondência

de strings. Essa técnica consiste em passar todas as palavras para caixa alta ou caixa baixa. Ela deve ser considerada na análise lexical (Baeza-Yates e Ribeiro-Neto, 2013: 211). Outra preparação é a remoção de acentuação. De acordo com Orengo e Huyck (2001: 188), a remoção de acentos é necessária porque há casos em que algumas formas variantes da palavra são acentuadas e outras não.

Nessa etapa, é comum o uso da tokenização, que é a primeira etapa na criação de um índice sobre qualquer tipo de dado de texto (Zhai e Massung, 2016: 61). A tokenização consiste em dividir o fluxo de caracteres em palavras, ou tokens (Weiss *et al.,* 2010: 20). Outra técnica comum nessa etapa é a eliminação de *stopwords*. Segundo Baeza-Yates e Ribeiro-Neto (2013: 213), termos frequentes que aparecem em cerca de 80% dos documentos não têm utilidade para o propósito de recuperação.

Outra técnica é a redução do termo ao seu radical. Essa técnica é chamada de *stemming*. Para Zhai e Massung (2016: 66), *stemming* é o processo de reduzir uma palavra a uma forma básica. Semelhante modo, Berry (2004: 134) define como o processo que identifica a forma raiz das palavras removendo sufixos.

Um método usado nessa fase é o *bag of words* (BoW), que de acordo com alguns estudados (Meena e Lawrance, 2019; Costantino *et al*., 2017), é utilizado para extração de informações de palavras citadas em documentos. Segundo Meena e Lawrance (2019: 2610), a *bag of words* (BoW) constrói um vocabulário do documento fornecido.

Um modelo de ponderação de termos de um documento, muito utilizado para classificação de textos utilizado em diversos estudos é o TF-IDF (Kuang, Brantingham e Bertozzi, 2017; Meena e Lawrance, 2019). Essa medida é composta por dois cálculos, o cálculo do TF (*Term Frequency*) e o cálculo do IDF (*Inverse Document Frequency*). Esse modelo foi apontado em alguns estudos pelos bons resultados alcançados na extração de recursos (Al-Saif e Al-Dossari, 2018: 382).

Salton e McGill (1983: 205) descrevem o TF como sendo o número de vezes que um termo aparece no texto de um documento (fórmula 1):

$$tf_{ik} = freq_{ik} \tag{1}$$

O IDF mede quão rara é a palavra no corpus, conforme descrevem Manning, Raghavan e Schütze (2009: 118). Ele tem a (fórmula 2):

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389

$$idf_t = log \frac{N}{df_t} \quad (2)$$

Esses dois elementos são a base de construção da medida resultado da multiplicação desses elementos (fórmula 3):

$$tfidf_{t,d} = tf_{t,d} x idf_{t,d} \quad (3)$$

Ainda nessa etapa de pré-processamento, o uso do n-gramas dá o contexto de dois (bigramas) ou três (trigramas) termos que estão próximos e que tem certa frequência que aparecem juntos. Alguns estudos atribuem ainda vantagens a métodos estatísticos como n-gramas, tendo como principais vantagens ser independente do idioma e funcionar muito bem com arquivos que contêm erros linguísticos e ruído (Al-Saif e Al-Dossari, 2018: 378).

Para apresentação de resultados, *word cloud* (ou nuvem de palavras) é uma técnica simples de visualização que permite a quem está analisando tenha uma compreensão de primeiro nível de conceitos/termos proeminentes (Cardoza e Wagh, 2017: 61). No contexto da mineração de textos, utilizando *word cloud*, quanto maior é o peso das palavras no texto a ser analisado, maior se tornam as palavras na visualização (Ramsden e Bate, 2008).

Na etapa de processamento um método utilizado para agrupar conjuntos de dados é a clusterização. Os algoritmos de clusterização são técnicas de aprendizado de máquina não supervisionada que agrupam um conjunto de documentos em subconjuntos e tem por objetivo criar clusters que sejam coerentes internamente, mas claramente diferentes um do outro (Manning, Raghavan e Schütze, 2009: 349).

O levantamento de trabalhos científicos que abordam o tema da mineração de textos auxilia no propósito de buscar pesquisas que possam contribuir para descoberta de conhecimento em investigações criminais. De acordo com Sampaio e Mancini (2007: 84), as revisões sistemáticas servem para incorporar um espectro maior de resultados relevantes.

Metodologia

A caracterização desta pesquisa é descrita quanto a sua natureza, objetivos, abordagem e procedimentos técnicos. Quanto à natureza da pesquisa, este estudo está caracterizado por uma pesquisa aplicada. Quanto aos objetivos, trata-se de uma pesquisa exploratória. Quanto à forma de abordagem, a pesquisa trata do método misto.

Com relação aos procedimentos técnicos, este estudo se trata de pesquisa bibliográfica e documental. Ainda sobre os procedimentos técnicos deste estudo, na pesquisa bibliográfica foram consultadas as bases Web of Science, IEEE Explorer, Scopus e Lisa, além de outras fontes literárias, como os livros e leis brasileiras.

As principais etapas de uma pesquisa de mineração de textos são: coleta, pré-processamento, mineração de textos e pós-processamento. Essas etapas são descritas no modelo de descoberta de conhecimento em banco de dados (Knowledge Discovery in Databases – KDD), conhecido tradicionalmente desde 1989 (Fayyad, 2001: 32-33).

Este estudo não se propõe a esboçar diferentes modelos de processo, como CRISP-DM ou SEMMA, mas emprega etapas sob a percepção de Fayyad, onde a Mineração de Dados é uma das fases do KDD (Fayyad, Piatetsky-Shapiro e Smyth, 1996: 39). A *Figura 2* apresenta essas etapas com as técnicas aplicadas neste estudo.

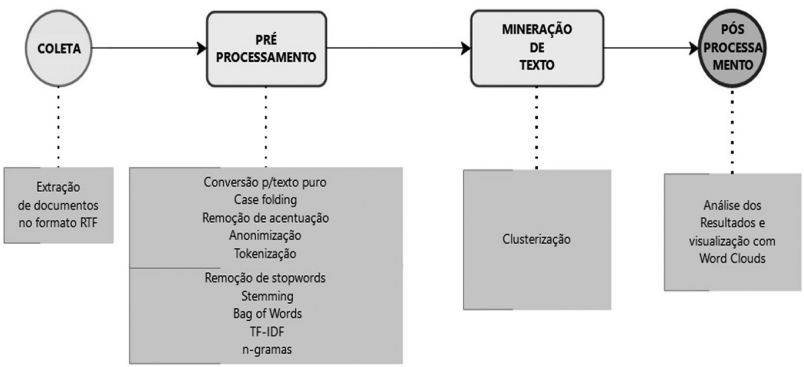

Figura 2. Etapas da mineração de textos com técnicas empregadas no estudo

Neste estudo, a coleção de 250 documentos foi coletada de um conjunto de peças textuais disponíveis em inquéritos relatados, de onde se propõe extrair conhecimento não trivial. Esses documentos foram extraídos no formato *Rich Text Format* (RTF).

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389

Ao iniciar a etapa de pré-processamento, os documentos RTF foram convertidos para o formato de texto puro, utilizando script em linguagem python escrito durante o estudo, formando um *dataset* único dos 250 documentos. Nesse processo foi empregada a codificação de caracteres no padrão utf-8 para garantir a compatibilidade de caracteres. Foi aplicada a técnica de *case folding*, sendo neste estudo convertido para letras minúsculas, além de serem removidas as acentuações de caracteres.

Esta pesquisa aplicou um processo de anonimização de dados com objetivo de salvaguardar informações de natureza pessoal. Esse processo foi dividido em 9 fases. Nela forma substituídas os atributos: UF, município, CEP, telefones, nomes, RG, matrícula, placa de veículo e CPF, conforme apresentado na *Figura 3.*

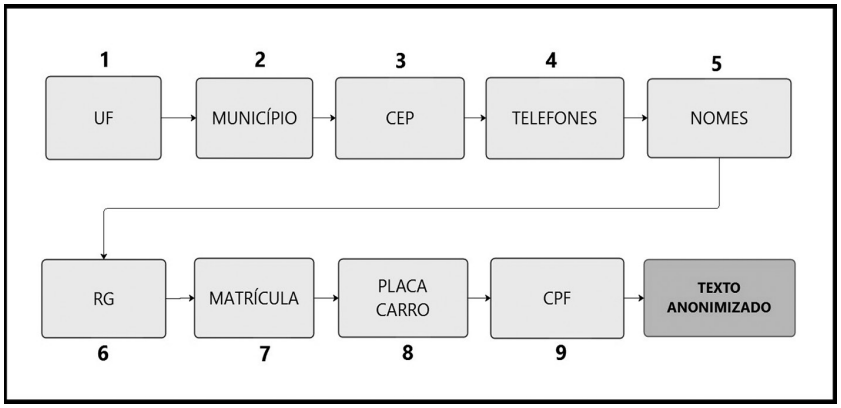

Figura 3. As 9 fases do uso da anonimização no estudo

A tokenização neste estudo foi realizada com o uso da biblioteca de código aberto do Python *Natural Language Toolkit* (NLTK). A biblioteca Pandas (biblioteca de código aberto do Python) foi utilizada para gerar um *dataframe* contendo a unidade das palavras de cada documento.

Em seguida foram eliminadas as *stopwords*. Além da lista de termos mais frequentes no idioma português do Brasil, foram incluídos termos frequentes no conteúdo de documentos textuais de inquérito policial.

Foram realizados dois experimentos neste estudo. O primeiro partiu deste ponto. Após a eliminação das *stopwords*, foi implementado o algoritmo de *bag of words* para criar um vetor booleano de frequência, indicando se cada palavra do "saco de palavras" está ou não presente no documento. Na sequência, foi aplicado o cálculo de ponderação de termos com o TF-IDF, para obter os termos mais representativos.

No uso do TF-IDF, foi utilizada a biblioteca de código aberto do Python Scikit-Learn. O método "TfidfVectorizer" da biblioteca Scikit-Learn foi parametrizado para obter esses termos representativos no modelo *n-grams,* obtendo por meio do parâmetro "ngram_range" esses valores em unigramas, bigramas e trigramas. O emprego do TF-IDF com *n-grams* foi proposto neste estudo para descobrir tópicos, tal como proposto pelo *topic model*.

No segundo experimento, após a eliminação das *stopwords*, foi aplicada também *stemming* utilizando a biblioteca NLTK. O pacote *stem* da biblioteca NLTK utiliza vários *stemmers*, dentre eles o RSLPStemmer, que é específico para língua portuguesa, sendo este utilizado neste estudo. Nesse segundo caso, só após o *stemming* foram repetidas as aplicações de *bag of words*, TD-IDF e *n-grams*.

Feitos os dois experimentos, os resultados foram comparados, sendo a principal comparação os resultados de n-gramas extraídos do cálculo do TF-IDF. Com a comparação, foi escolhido um dos experimentos para utilizar na etapa de mineração de textos, que obteve resultado mais significativo.

Nessa etapa, foi utilizada a técnica de *word clouds* como forma de visualização dos n-gramas mais representativos. Foi elaborado *script* para geração da nuvem de palavras com o uso do pacote WordCloud do Python. O *script* elaborado utilizou outros pacotes auxiliares para gerar a imagem da nuvem de palavras, como matplotlib e numpy.

Após esses procedimentos, iniciou-se a etapa de mineração de textos propriamente dita, sendo utilizada a técnica de clusterização com o método *k-means*. O algoritmo de clusterização foi utilizado para agrupar conjuntos de trigramas extraídos do experimento.

Foi aplicado o *Elbow Method* (método do cotovelo) para determinar o número de clusters a ser utilizado no método *k-means*. Também foi utilizado o indicador *Silhouette* para análise da distância de cada cluster.

Outros métodos de agrupamentos como *Agglomerative Clustering* e DBSCAN são conhecidos. O *Agglomerative Clustering* é um método hierárquico, especialmente úteis quando o objetivo é organizar os clusters em uma hierarquia natural. O DBSCAN tem boa aplicação quando as densidades não são diferentes e precisa de uma seleção cuidadosa de parâmetros. O K-Means foi selecionado por seu uso ser menos intensivo em termos de computação, tornando mais rápido, e são adequados para conjuntos de dados muito grandes e ser próprio para agrupamento por distância entre pontos.

Com isso, foram aplicadas as principais etapas de mineração de textos, as técnicas de pré-processamento (tokenização, *stopwords*, *stemming*, BoW, TF-IDF, *n-grams*), a técnica de mineração de textos para agrupar conjuntos de dados (*k-means*) e a técnica de pós-processamento para apresentar os

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389

resultados (*word clouds*). Após essas estapas, este estudo passou a analisar os resultados obtidos nos dois experimentos.

Resultados

Forma obtidos resultados dos dois experimentos. Esses resultados puderam ser comparados para se verificar quais descobertas surgiram que pudessem contribuir para os objetivos deste estudo.

Resultado do primeiro experimento

O vocabulário gerado no primeiro experimento resultou em 8.274 termos (ou tokens). Com o ranqueamento dos termos mais frequentes do *dataset*, foram encontrados 1.049 termos que apareceram pelo menos 10 vezes nos documentos.

Com o BoW foi gerada uma matriz com esses termos mais frequentes, resultando em 250 linhas por 1.049 colunas. Cada linha representa um documento e as colunas são os termos mais frequentes do vocabulário. A *Figura 4* apresenta a distribuição da frequência desses termos em cada documento.

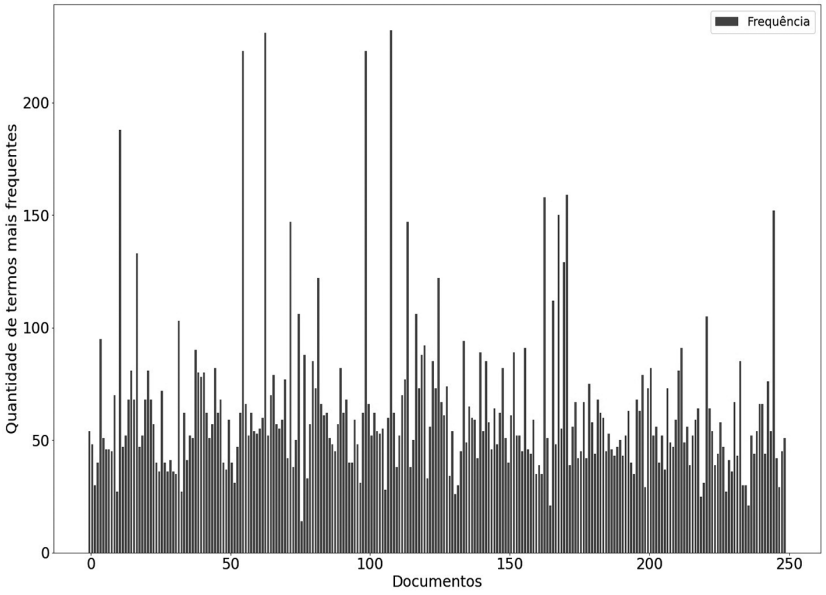

Figura 4. Distribuição da frequência dos termos do vocabulário nos documentos

Os pesos do TF-IDF foram calculados, sendo gerada uma matriz de dispersão com 66.052 colunas (contendo unigramas, bigramas e trigramas) por 250 linhas que representam a quantidade de documentos. Uma amostra dos resultados n-gramas dessa matriz está na *Figura 5*.

dfNgramas - DataFrame

Index	unigrams	bigrams	trigrams
0	romaria	eunara cunha	rosiclesia eunara cunha
1	rocilmara	humberta orfila	eunara cunha cordeiro
2	linecker	rosiclesia eunara	jasone rani almeida
3	auster	cunha cordeiro	romaria rosiclesia eunara
4	cilmara	jasone rani	auster sionice orgelia
5	relacao	rani almeida	sionice orgelia rodrigues
6	orfila	romaria rosiclesia	cilmara tacila beiro
7	mercadorias	relacao item	ronieles aldeciane schneider
8	onibus	cilmara tacila	humberta orfila vaudenir
9	humberta	sionice orgelia	orfila vaudenir tacila
10	rosiclesia	auster sionice	rocilmara auster sionice
11	tacila	orgelia rodrigues	greisi dilerman rachadel
12	nota	tacila beiro	fostina greisi dilerman
13	dilerman	ronieles aldeciane	neilon orfila tacila

Format Resize ☑ Background color ☑ Column min/max Save and Close Close

Figura 5. Dataset com *n-grams* extraído do TF-IDF

Essa matriz é composta com 8.274 unigramas, 27.265 brigramas e 30.513 trigramas. Este primeiro experimento será comparado com o resultado do segundo experimento.

Resultado do segundo experimento

O segundo experimento utilizou *stemming* para reduzir a dimensionalidade do vocabulário. Nesse experimento, o vocabulário gerado contém 5.407 termos. O ranqueamento dos termos mais frequentes do vocabulário, com frequência de pelo menos 10 ocorrências, resultou em 1.078 termos.

Neste caso, a matriz gerada com o BoW resultou em 250 linhas por 1.078 colunas. Nas linhas estão dispostos os documentos e nas colunas os termos mais frequentes do vocabulário. A distribuição da frequência desses termos está na *Figura 6*.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389

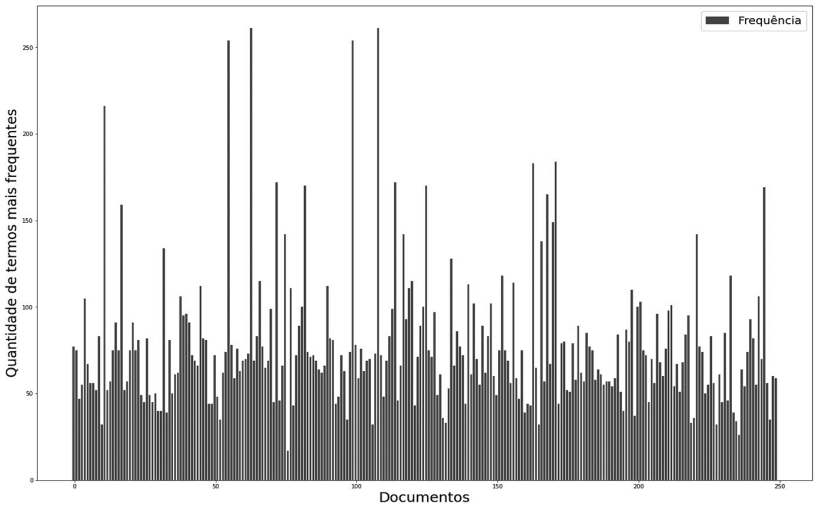

Figura 6. Distribuição da frequência dos termos do vocabulário nos documentos

A matriz de dispersão gerada para o cálculo dos pesos do TF-IDF resultou na forma de 62.064 x 250. A *Figura 7* apresenta uma amostra desse conjunto de dados com o n-gramas.

dfNgramas - DataFrame

Index	unigrams	bigrams	trigrams
0	rom	eun cunh	rosicles eun cunh
1	rocilm	rosicles eun	eun cunh cord
2	lineck	humbert orfil	jason ran alme
3	pag	cunh cord	rom rosicles eun
4	mercad	jason ran	aust sion orgel
5	aust	ran alme	sion orgel rodrig
6	receb	relaca it	cilm tacil beir
7	not	rom rosicles	humbert orfil vauden
8	encontr	cilm tacil	orfil vauden tacil
9	trabalh	not fiscal	roniel aldecian schneid
10	cilm	sion orgel	rocilm aust sion
11	relaca	aust sion	fostin greis dilerman
12	argentin	orgel rodrig	greis dilerman rachadel

Format | Resize | Background color | Column min/max | Save and Close | Close

Figura 7. Dataset com *n-grams* extraído do TF-IDF

Essa matriz é composta com 5.407 unigramas, 26.313 brigramas e 30.344 trigramas. A comparação deste resultado com o primeiro experimento está descrita no próximo item.

Comparação dos experimentos

Os vocabulários gerados nos dois experimentos têm uma diferença de tamanho, sendo o segundo experimento cerca de 35% menor, pois com o *stemming* houve uma redução da dimensionalidade dos termos do *corpus*. Algumas diferenças estão apresentadas no *Quadro 1*.

		Primeiro experimento	Segundo experimento	Diferença
Vocabulário		8.274	5.407	2.867
Termos com 10 ou mais ocorrências		1.049	1.078	-29
Conjunto de n-gramas	Bigramas	27.265	26.313	952
	Trigramas	30.513	30.344	169

Quadro 1. Relação de termos frequentes com o peso TF

Pode-se observar que dos termos extraídos com 10 ou mais frequências, o resultado do segundo experimento foi de uma quantidade maior de termos frequentes, apesar de ter um vocabulário menor. A quantidade de n-gramas extraídos do TF-IDF teve comportamento semelhante.

Tomando por base os trigramas extraídos do cálculo do TF-IDF, o ranqueamento dos pesos obtidos nos dois experimento foi comparado a fim de verificar se há diferença acentuada entre eles. Essa comparação determinou o resultado do experimento escolhido para utilização na etapa seguinte de processamento.

Os resultados vistos no ranqueamento foram semelhantes nos dois experimentos. Isso pode ter ocorrido em razão do tamanho do *dataset* utilizado. O *Quadro 2* apresenta uma amostra da comparação desses resultados.

Trigramas extraídos com TF-IDF			
Primeiro Experimento		Segundo Experimento	
rosiclesia eunara cunha	1.258411481	rosicles eun cunh	1.276647444
eunara cunha cordeiro	1.189530036	eun cunh cord	1.20725675
jasone rani almeida	0.991411358	jason ran alme	1.004510033

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389

romaria rosiclesia eunara	0.976917018	rom rosicles eun	0.988558846
auster sionice orgelia	0.951230086	aust sion orgel	0.964718881
sionice orgelia rodrigues	0.939244637	sion orgel rodrig	0.953204469
cilmara tacila beiro	0.937303188	cilm tacil beir	0.948106967
ronieles aldeciane schneider	0.891948811	humbert orfil vauden	0.898771493
humberta orfila vaudenir	0.885349259	orfil vauden tacil	0.898771493
orfila vaudenir tacila	0.885349259	roniel aldecian schneid	0.897228426
rocilmara auster sionice	0.742616682	rocilm aust sion	0.753424509
fostina greisi dilerman	0.722217532	fostin greis dilerman	0.72873275
greisi dilerman rachadel	0.722217532	greis dilerman rachadel	0.72873275
neilon orfila tacila	0.697492064	neilon orfil tacil	0.706798274
orfila tacila nunes	0.697492064	orfil tacil nun	0.706798274
canrobert sidete machado	0.642012072	canrobert sidet mach	0.652673946

Quadro 2. Comparação dos trigramas extraídos nos dois experimentos

Sabendo que a *stemming* reduz a dimensionalidade dos termos no vocabulário, a sua aplicação neste estudo não foi suficientemente expressiva. Como não houve diferença significante, a etapa da mineração de textos foi realizada com o resultado do primeiro experimento.

Visualização com word cloud

Ao gerar a visualização com nuvem de palavras apresentou os termos mais frequentes no corpus, deixando-os em evidência na "nuvem". A técnica facilitou a visualização desses termos, considerando o volume do vocabulário de trigramas.

O resultado apresentado nessa técnica contou com o ajuste da biblioteca utilizada para considerar o dicionário de trigramas. A nuvem de palavras está representada na *Figura 8*.

O resultado da *word cloud* indica a evidente valorização de nomes pelo TF-IDF. Embora não seja objetivo deste estudo o reconhecimento de entidades nomeadas (NER), essa descoberta evidenciou esse comportamento ao utilizar a ponderação da técnica TF-IDF para valorizar nomes. Com essa técnica, os termos representativos conjugam nomes e fatos, sugerindo a relação entre eles. Além de nomes, outra descoberta neste estudo é a possibilidade de estabelecer conexões com o uso do TF-IDF.

Foi visualizado grande volume de nomes presentes na nuvem de palavras. A *Figura 9* apresenta outra nuvem de palavras gerada sem os nomes próprios no mesmo dicionário e considerando bigramas e trigramas.

Figura 8. Termos mais representativos do dicionário de trigramas

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389

Figura 9. Termos mais representativos dos *n-grams* sem nomes próprios

Com a retirada dos nomes próprios os temas abordados no contexto do dicionário surgem em destaque. Esses termos representam mais claramente o resultado do TF-IDF. A próxima etapa tentou agrupar os resultados para identificar as categorias representativas.

Resultado da clusterização

Como se viu no resultado do TF-IDF, esse método valoriza nomes. Contudo, para esta etapa foram eliminados os nomes da lista de atributos a serem utilizados na clusterização.

Para a clusterização, o conjunto de dados utilizado para treino representou 60% do dataset gerado no primeiro experimento e o conjunto para teste representou 40% do dataset. Após testar um número variado de clusters, com valor de K entre 2 e 19. A cada interação desse intervalo os textos foram classificados em clusters, com base nas *features* mais relevantes.

Com o resultado do *Elbow Method* e do *Silhouette*, utilizado na análise da métrica estatística da distância de cada grupo, o número de clusters foi definido em k=8. Os valores da variável *Inertia_* (do algoritmo do método do cotovelo) e a pontuação média do indicador *Silhouette* estão no *Quadro 3*.

Interação de clusters	Valores de ***Inertia_***	Valores de score de ***Silhouette***
2	26.407766542630895	0.21447120803450873
3	23.40613742895484	0.4543829253092717
4	21.5162601542226	0.533569334507583
5	20.19718766312582	0.614872536807049
6	19.119907149891077	0.6472925266298454
7	17.34682388257048	0.7082602221754799
8	14.755351548672904	0.7916165788142853
9	14.170636026494282	0.8182662988835868
10	13.449997605349845	0.8418286611777918
11	12.860182294729992	0.8798681665203688
12	12.5578304435155	0.8874530403988156
13	11.99133960052284	0.9080805754423239
14	11.65154548866381	0.9202136520380176
15	11.176742884157882	0.9367101164812598
16	10.942296599429019	0.9284410700703736
17	10.684578800515458	0.9199451324245363
18	10.497505698669887	0.9074891070016042

Quadro 3. Cálculo da distância de cada grupo

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389

O resultado do *Elbow Method* sugere uma leve diminuição na soma dos erros quadráticos das instâncias de cada cluster quando ocorre a interação com 8 clusters. O gráfico do método do cotovelo identifica visualmente o declínio desse atributo *Inertia_*, como na *Figura 10*.

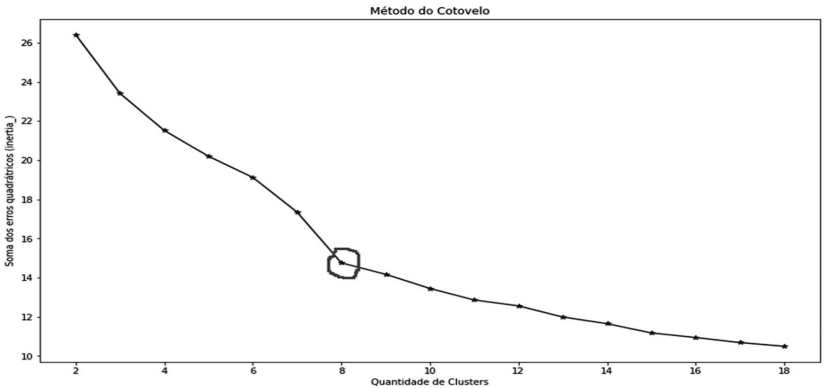

Figura 10. Gráfico do método do cotovelo

O valor de *Silhouette* entre 0 e 1 mostra a sobreposição. Caso o valor do *Silhouette* fosse próximo de 1, seria uma indicação de classificação mais resolutiva e se o valor fosse mais próximo do 0 significa superposição dos cluster. Tendo o método do cotovelo indicado o número de 8 clusters, o valor de pontuação média do *Silhouette* nessa interação é uma variação razoavelmente próxima de 1, o suficiente para definir o valor de k=8. A *Figura 11* apresenta o gráfico do *Silhouette*.

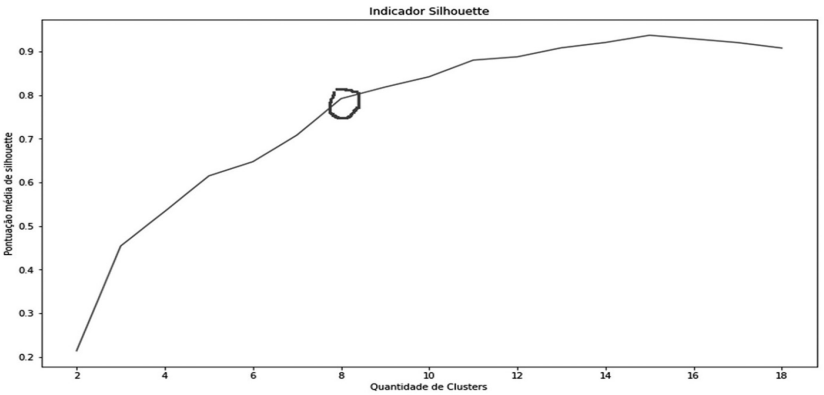

Figura 11. Gráfico do indicador *Silhouette*

Com esses resultados do método do cotovelo e do indicador *Silhouette*, foi utilizado o agrupamento com 8 clusters. Com os clusters gerados e a análise dos termos mais representativos em cada cluster foi possível identificar categorias no agrupamento. Esses termos estão representados no *Quadro 4*.

Cluster	Termos mais representativos nos clusters
0	notas fiscais – mil reais – apreensao efetuada – credito rural – caixa economica – exercendo funcao – operacoes credito rural
1	nota falsa – prisao flagrante – quais recebeu – recebeu nota – condicoes identificar – nota falsa reconhecedora – nota falsa pagamento – entrega nota falsa
2	alarmes ocorrências – alarmes instalados – vigilante – verifica arrombamento – alarme disparado – entregar interceptado
3	apreensao material – vigilancia – prestacao servicos seguranca – seguranca firmados seguranca – servico seguranca
4	quantidade maconha – cor vermelha – cor verde – maconha comprimido cor – comprimido cor – comprimido cor vermelha
5	grupo rapazes – grupo argentinos – dolares americanos – cem dolares – rapazes argentinos – pagamento deveria
6	cinquenta reais – cedula cinquenta – cedula cinquenta reais – contenda detentor – numero serie – recebido cedula
7	programa realizado – pagamento programa –moeda pagamento – pedra maconha – pedra maconha tamanho – maconha tamanho – cem reais – reais moeda

Quadro 4. Termos mais representativos em cada cluster

Os termos mais relevantes dos clusters estão relacionados com áreas de atuação da Polícia Federal brasileira. Eles estão indicados juntamente com sua área respectiva, conforme o *Quadro 5*.

Cluster	Área de atuação
0	Corrupção
1	Fazendária
2	Fiscalização de Segurança Privada
3	Defesa Institucional
4	Tráfico de Entorpecentes
5	Controle Migratório
6	Financeiro
7	Tráfico Internacional de Entorpecentes

Quadro 5. Termos relacionados à área de atuação

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389

Como os algoritmos de aprendizado não supervisionado não utilizam categorias, essas dependem de análise. O *Quadro 5* apresenta as categorias identificadas na análise de cada cluster, com base dos termos mais relevantes.

Considerações finais

Este estudo teve o objetivo de descobrir conhecimento não trivial em documentos textuais da investigação policial brasileira. Para isso, utilizou técnicas de mineração de textos, buscando identificar os termos mais representativos do *corpus*.

Os resultados desta pesquisa constituíram as bases para compreensão de como as técnicas de mineração de textos em documentos textuais do inquérito policial podem auxiliar à investigação policial brasileira. O emprego das técnicas e métodos utilizados permitiram descobrir características dos documentos da investigação policial, tais como os termos mais relevantes de um conjunto de documentos.

A análise de diversos documentos textuais de inquéritos policiais diferentes permitiu o estudo comparativo de documentos que tratam temas diversos. O uso de técnicas para descoberta de conhecimento resultou na identificação das categorias desses documentos. Na avaliação desses resultados obtidos, essas categorias foram identificadas como temas de atuação da Polícia Federal brasileira.

Além das técnicas de descoberta de conhecimento indicarem categorias de atuação da polícia, verificando os termos mais representativos em documentos, elas também apresentaram bons resultados para identificar nomes e conexões com fatos. Os termos mais representativos servirão de análise sobre quais áreas concentram mais volume, de acordo com o escopo do universo da amostra.

Este estudo recomenda a aplicação das técnicas nele utilizadas em amostra com maior volume documentos textuais e maior amplitude de casos. Novos estudos podem aproveitar as técnicas e as bibliotecas utilizadas nesta pesquisa e também fazer uso de outras técnicas de mineração de textos.

Referências

Agência de Notícias - Polícia Federal. 2016. "Polícia Federal lança sistema de inquérito eletrônico". Acao, Divisão de Comunicação Social, 24 de outubro. Acessado 14 de dezembro de 2020.
http://www.pf.gov.br/agencia/noticias/2016/10/policia-federal-lanca-sistema-de-inquerito-eletronico

Al-Saif, Hissah, e Hmood Al-Dossari. 2018. "Detecting and classifying crimes from arabic twitter posts using text mining techniques". *International Journal of Advanced Computer Science and Applications* 9 (10): 377-387. https://doi.org/10.14569/IJACSA.2018.091046

Aranha, Christian, e Emmanuel Passos. 2006. "A Tecnologia de Mineração de Textos". *Revista Eletrônica de Sistemas de Informação* 5 (2): 1-8. https://doi.org/10.21529/RESI.2006.0502001

Baeza-Yates, Ricardo, e Berthier Ribeiro-Neto. 2013. *Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca*. 2a. ed. Traduzido por Leandro Krug Wives e Viviane Pereira Moreira. Brasil: Bookman.

Berry, Michael W., org. 2004. *Survey of Text Mining*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-4305-0

Borko, Harold. 1968. "Ciência da informação: o que é isto?". *American Documentation:* 1-6. Tradução livre. Acessado 14 de dezembro de 2020. https://edisciplinas.usp.br/pluginfile.php/2532327/mod_resource/content/1/Oque%C3%A9CI.pdf

Cardoza, Clinton, e Rupali Wagh. 2017. "Text analysis framework for understanding cyber-crimes". *International Journal of Advanced and Applied Sciences* 4 (10): 58-63. https://doi.org/10.21833/ijaas.2017.010.010

Costantino, Gianpiero, Antonio La Marra, Fabio Martinelli, Andrea Saracino, e Mina Sheikhalishahi. 2017. "Privacy-preserving text mining as a service". *IEEE Symposium on Computers and Communications* (ISCC), 890-897. https://doi.org/10.1109/ISCC.2017.8024639

Fayyad, Usama. 2001. "Knowledge Discovery in Databases: An Overview". In *Relational Data Mining*, organizado por Sašo Džeroski e Nada Lavrač, 28-47. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-04599-2_2

Fayyad, Usama, Gregory Piatetsky-Shapiro, e Padhraic Smyth. 1996. "From Data Mining to Knowledge Discovery in Databases". *AI Magazine* 17 (3): 37-54. https://doi.org/10.1609/aimag.v17i3.1230

Kuang, Da, P. Jeffrey Brantingham, e Andrea L. Bertozzi. 2017. "Crime topic modeling". *Crime Science* 6 (1). https://doi.org/10.1186/s40163-017-0074-0

Manning, Christopher D., Prabhakar Raghavan, e Hinrich Schütze. 2009. *Introduction to Information Retrieval.* UK: Cambridge University Press.

Meena, K., e Raj Lawrance. 2019. "An automatic text document classification using modified weight and semantic method". *International Journal of Innovative Technology and Exploring Engineering* 8 (12): 2608-2622. https://doi.org/10.35940/ijitee.K2123.1081219

Orengo, Viviane Moreira, e Christian Huyck. 2001. "A Stemming Algorithm for the Portuguese Language". *Proceedings Eighth Symposium on String Processing and Information Retrieval*, 186-193. Laguna de San Rafael, Chile: IEEE. https://doi.org/10.1109/SPIRE.2001.989755

Ramsden, Andrew, e Andrew Bate. 2008. "Using Word Clouds in Teaching and Learning", agosto. https://researchportal.bath.ac.uk/en/publications/using-word-clouds-in-teaching-and-learning

Salton, Gerard, e Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. USA: McGraw-Hill, Inc.

Sampaio, Rosana Ferreira, e Marisa Cotta Mancini. 2007. "Estudos de revisão sistemática: um guia para síntese criteriosa da evidência científica". *Revista Brasileira de Fisioterapia* 11 (1): 83-89. https://doi.org/10.1590/S1413-35552007000100013

Silva, Leandro, Sarajane Peres, e Clodis Boscarioli. 2017. *Introdução a Mineração de Dados com aplicações em R*. Rio de Janeiro: Elsevier.

Silva, Marcio Ponciano da. 2019. "Focos de inovação na Polícia Federal para combater a corrupção e o crime organizado". In *Carreiras típicas de Estado: desafios e avanços na prevenção e no combate à corrupção*, 275–284. Belo Horizonte: Fórum.

Weiss, Sholom Menachem, Nitin Indurkhya, Tong Zhang, e Fred Damerau. 2010. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer.

Yang, Yongpeng, Monisha Manoharan, e Kathleen Suzanne Barber. 2014. "Modelling and Analysis of Identity Threat Behaviors through Text Mining of Identity Theft Stories". IEEE Joint Intelligence and Security Informatics Conference, 50-63. https://doi.org/10.1109/JISIC.2014.35

Zhai, ChengXiang, e Sean Massung. 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. USA: Association for Computing Machinery and Morgan & Claypool.

Zhou, Xuezhong, Yonghong Peng, e Baoyan Liu. 2010. "Text mining for traditional Chinese medical knowledge discovery: A survey". *Journal of Biomedical Informatics* 43 (4): 650-660. https://doi.org/10.1016/j.jbi.2010.01.002

Para citar este texto:

Silva, Marcio Ponciano da e Angel Freddy Godoy Viera. 2021. "Descoberta de conhecimento com uso de técnicas de mineração de textos aplicadas em documentos textuais da investigação policial brasileira". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 35 (88): 161-183. http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.88.58389