Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Avaliação de métodos de Similaridade Textual no contexto de Investigação Policial.

Antonio Ricardo Marques Junior

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Recuperação da Informação, NLP

João Arthur Brunet Monteiro
(Orientador)

Campina Grande, Paraíba, Brasil

AVALIAÇÃO DE MÉTODOS DE SIMILARIDADE TEXTUAL NO CONTEXTO DE INVESTIGAÇÃO POLICIAL.

ANTONIO RICARDO MARQUES JUNIOR

DISSERTAÇÃO APROVADA EM 03/02/2020

JOÃO ARTHUR BRUNET MONTEIRO, Dr., UFCG
Orientador(a)

TIAGO LIMA MASSONI, Dr., UFCG
Examinador(a)

NAZARENO FERREIRA DE ANDRADE, Dr., UFCG
Examinador(a)

FLAVIO VINICIUS DINIZ DE FIGUEIREDO, Dr., UFMG
Examinador(a)

CAMPINA GRANDE - PB

M357a Marques Junior, Antonio Ricardo.
Avaliação de métodos de similaridade textual no contexto de investigação policial / Antonio Ricardo Marques Junior. - Campina Grande, 2020.
63 f. : il.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2020.
"Orientação: Prof. Dr. João Arthur Brunet Monteiro.
Referências.

1. Recuperação da Informação. 2. Processamento de Linguagem Natural. 3. Aprendizagem de Máquina. 4. Investigação Policial. 5. Similaridade Textual. I. Monteiro, João Arthur Brunet. do. II. Título.

CDU 004.22(043)

FICHA CATALOGRÁFICA ELABORADA PELO BIBLIOTECÁRIO GUSTAVO DINIZ DO NASCIMENTO CRB-15/515

Resumo

A Polícia Federal (PF) atua, dentre suas diversas atribuições, na apuração de inquéritos através de delegados e agentes federais em seus respectivos núcleos de investigação. Uma das tarefas mais recorrentes realizada pelos investigadores ocorre no processo de instauração de inquéritos, onde o responsável deve verificar se já existe um procedimento de investigação criminal para o fato em questão. Entretanto, por se tratar de uma atividade subjetiva e que depende do indivíduo que a realiza, existe a possibilidade da instauração de mais de um inquérito apurando o mesmo fato, dificultando o processo de investigação.

Este estudo compara modelos clássicos e do estado da arte em Recuperação da Informação como Distância de Cosseno, Similaridade de Jaccard, Doc2Vec e WMD, na busca por inquéritos relevantes a partir de informações estruturadas e não-estruturadas (documentos textuais), visando identificar duplicidade de inquéritos, casos similares que auxiliem em tomadas de decisão em investigações ou para treinamento de novos delegados e crimes que possam estar relacionados.

Para a construção dos modelos foram utilizados dados de inquéritos não-sigilosos do ePol, plataforma *web* que gerencia atividades policiais e interliga as unidades da PF. Os modelos construídos retornam o *top* 4 inquéritos similares a um inquérito passado como entrada. Dado que o problema trata de dados não-supervisionados, a avaliação foi realizada por meio de especialistas no contexto, representados por delegados e escrivães da PF, onde estes responderam a formulários submetidos diariamente com inquéritos a serem comparados.

Os resultados mostram que métodos clássicos como similaridade de jaccard e distância de cosseno atingem bons resultados para detecção de inquéritos semelhantes, com NDCGs iguais a 0.8812 e 0.8371 respectivamente. O modelo WMD ainda apresenta um NDCG próximo aos já citados (0.8037) e o doc2vec atinge o pior resultado (0.6743).

O estudo sugere que o desempenho dos modelos baseados em redes neurais estão abaixo dos demais devido a base de treinamento não ser considerada grande o suficiente para um modelo de rede neural profunda, o que pode dificultar a tarefa de aprendizado para este tipo de abordagem. Para detecção de duplicidade e relação entre inquéritos os resultados não foram satisfatórios de acordo com a métrica utilizada. Entretanto, vale salientar que,

ao contrário da semelhança entre inquéritos, duplicidade e relação entre inquéritos não são eventos comuns de ocorrerem neste contexto.

Os modelos sugeridos no estudo podem ser utilizados junto a plataforma ePol, auxiliando na identificação de duplicidade e assim otimizando o trabalho da PF ao reduzir o desperdício de recursos da corporação, além de sugerir inquéritos semelhantes para, por exemplo, auxiliar no treinamento de novos delegados sobre como e quais ações devem ser tomadas na condução de um inquérito policial.

Abstract

The Brazilian Federal Police (PF) operates, among its diverse duties, in the investigation of cases through federal agents in their respective departments. One of the most recurrent tasks carried out by investigators occurs in the process of open investigations, where the person in charge must verify if there is already a criminal investigation procedure for the fact in question. However, because it is a subjective activity and it depends who performs it, there is the possibility of setting up more than one investigation ascertaining the same fact, making the investigation process difficult.

This study compares classic and and state-of-art models in information retrieval such as Cosine Distance, Jaccard Similarity, Doc2Vec, and WMD, in search of relevant inquiries from structured and unstructured data (textual documents), aiming to detect document inquiries duplicity, similar cases that assist decision-making in investigations or to train new delegates through similar crimes.

To build the IR models, we used non-confidential data from ePol, the web platform which manages investigations' activities and interconnects the Federal Police Stations of Brazil. Each model returns the 4 most similar inquires to a previous inquiry selected as input. 55 inquiries were used as queries for each model and their responses were submitted to an evaluation. Given the problem deals with unsupervised data, the evaluation was fulfilled by contextual experts, represented by PF delegates and clerks, where they answered surveys daily regarding comparisons between inquires.

The results show classical methods such as jaccard similarity and cosine distance achieve good results for similar inquiries' detection, with NDCGs equal to 0.8812 and 0.8371 respectively. The WMD model still has an NDCG close to those already mentioned (0.8037) and doc2vec achieves the worst result (0.6743).

The study suggests the performance of models based on neural networks are below the others because the training base is not considered large enough for a deep neural network model, which can make the learning task for this type of approach more difficult. For detection of duplicity and relationship between inquiries, the results were not satisfactory according to NDCG metric. However, it should be noted that, unlike the similarity between

inquiries, duplicity and relationship between inquiries are not common events to occur in this context.

The models suggested in this study can be used as a feature of the ePol platform, identifying duplicity between inquiries and thereby optimizing PF's work by reducing the waste of corporate resources, suggesting similar inquiries to new delegates and helping them regarding what actions should be taken in a police investigation.

Agradecimentos

Agradeço à minha família, em especial à minha mãe Lídia, por todo o esforço e dedicação na construção da minha formação tanto acadêmica quanto pessoal, ao meu pai Ricardo e minha irmã Ana Alice, por todo o apoio, a todos os colegas do curso de Computação que tive a oportunidade de conhecer, conviver e aprender constantemente.

À Coordenação da Pós-graduação em Computação da UFCG (COPIN) por todo o esforço em auxiliar sobre questões relativas a documentações, normas, procedimentos e processos internos.

Ao Laboratório SPLab e a todos do Projeto ePol, em especial ao professor e orientador João Arthur que viabilizou a execução deste trabalho, além de acompanhar de perto todo o processo. Aos líderes técnicos e gerentes do ePol, em especial a Júlio e Bruno por tantas dúvidas sanadas. Ao time de Analytics com quem tive o prazer de conviver por pouco mais de 1 ano: Aline Costa, Antunes Dantas, Arthur Costa, Arthur Lustosa, Davi Laerte, Ítalo Medeiros, Martha Michelly e Tatiana Saturno.

Ao professor Leandro Balby pela contribuição com o conhecimento específico em Aprendizagem de Máquina e Recuperação de Informação, além de dicas na linha da pesquisa.

A todos e todas do laboratório Analytics, em especial ao professor Nazareno por todo o conhecimento passado, por apresentar a área de Ciência de Dados e ter mostrado que Computação pode ser legal.

Conteúdo

Lista de Símbolos

PF - *Polícia Federal*

IPL - *Inquérito Policial*

DPF - *Delegado de Polícia Federal*

EPF - *Escrivão de Polícia Federal*

RI - *Recuperação de Informação*

HD - *Hipótese Distributiva*

NLP - *Natural Language Processing*

word2vec - *Word to Vector*

doc2vec - *Document to Vector*

WMD - *Word Mover's Distance*

NDCG - *Normalized Discounted Cumulative Gain*

Capítulo 1

Introdução

1.1 Contextualização

A Polícia Federal (PF) atua, dentre suas diversas atribuições, na apuração de inquéritos através de delegados e agentes federais em seus respectivos núcleos de investigação. Uma das tarefas mais recorrentes realizada pelos investigadores ocorre no processo de instauração de inquéritos, onde o responsável deve verificar se já existe um procedimento de investigação criminal para o fato em questão a fim de evitar duplicação. Além disso, existe uma dificuldade em lidar com o grande volume de inquéritos, onde os servidores da PF não fazem uso de forma sistemática da base de dados históricos de suas investigações.

A duplicidade de procedimentos de investigação criminal que apuram os mesmos fatos é considerado um tema relevante pelas autoridades. No entanto, a doutrina e a jurisprudência já tinham o entendimento no sentido de que a instauração simultânea de dois inquéritos policiais para a apuração dos mesmos fatos, além de configurar constrangimento ilegal, consistira numa violação clara ao princípio do "ne bis in idem" ou princípio da vedação a dupla incriminação, sendo este um dos princípios fundamentais do direito penal nacional e internacional que proíbe que uma pessoa seja processada, julgada e condenada mais de uma vez pela mesma conduta. [FONTENELE, 2015].

Inicialmente, cumpre destacar que a duplicidade de procedimentos investigatórios idênticos é uma situação extremamente desfavorável para o andamento das investigações criminais, pois além de gerar um gasto público desnecessário, pode levantar alegações de nulidade das investigações e operações policiais.

De acordo com as narrativas de delegados e escrivães da PF expostas em reuniões que tivemos com os mesmos, a atual maneira de evitar duplicidade de inquéritos é lenta e custosa, onde o responsável, no momento da instauração de um novo inquérito, verifica se já existe um procedimento de investigação criminal para o fato em questão. Segundo a PF, a verificação é feita a partir de consultas montadas no sistema com critérios definidos pelo investigador, processo este que é passível de falha humana, podendo assim não impedir a criação de procedimentos investigatórios duplicados para o mesmo fato. Caso existam duplicações, um dos inquéritos deve ser trancado e apensado ao outro para eventual aproveitamento do que for útil, e não simplesmente anexado ao primeiro, permanecendo em aberto na distribuição [dup, 2016].

Uma alternativa que pode auxiliar no processo de identificação de inquéritos duplicados é a recuperação de documentos relevantes a partir da similaridade de itens. Atualmente a tarefa de identificar documentos similares é bastante explorada, principalmente considerando a similaridade semântica. Ferrero et al. [Ferrero et al., 2017] propõem a utilização de representação distribuída de palavras (*word embeddings*) na detecção de plágio em textos de vários idiomas, apresentando novos métodos de detecção de similaridade. Já Kenter & De Rijke [Kenter and De Rijke, 2015] propõem um modelo de medir a similaridade semântica em textos curtos.

Entende-se recuperar documentos relevantes como sendo encontrar material (geralmente documentos) de uma natureza não estruturada (geralmente texto) que satisfaz uma necessidade de informação dentro de grandes coleções [Manning et al., 2009]. Já a pesquisa de similaridade é o termo mais geral utilizado para uma variedade de mecanismos que compartilham o princípio de busca (tipicamente, muito grande) em espaços de objetos onde o único comparador disponível é a semelhança entre qualquer par de objetos [Wikipedia, 2017a].

Com isso, considerando os conceitos envolvidos, este trabalho avalia modelos de recuperação de inquéritos baseado em itens similares, visando auxiliar o especialista na identificação de duplicidade dos processos investigatórios. Além do auxílio na identificação de duplicidade, a recuperação de casos relevantes pode ajudar na celeridade das investigações, uma vez que itens similares podem fornecer informações que auxiliem em tomadas de decisão de uma determinada investigação.

A aplicação de técnicas de extração de tópicos em documentos textuais por tokenização

em n-gramas e medidas estatísticas como tf-idf, aliadas ao cálculo de similaridade como a de cossenos e distância de jaccard são abordagens comumente utilizadas no contexto de Processamento de Linguagem Natural para identificar semelhanças entre documentos. Sendo assim, aplicamos técnicas de tokenização para criação de unigramas, bigramas e trigramas, aliadas a um conhecimento de domínio coletado através de especialistas da PF, referente a regras de associação representadas por expressões regulares para a construção de modelos baseados em sintaxe como a distância de cosseno e similaridade de jaccard, além do treinamento de modelos baseados em redes neurais como o doc2vec e WMD, com o intuito de: ter uma representação vetorial dos inquéritos; comparar modelos clássicos e do estado da arte em Recuperação da Informação e Semântica Distributiva que calculam distâncias entre documentos a partir de uma representação baseada em *tokens*; avaliar os modelos propostos com especialistas (delegados e escrivães).

1.2 Objetivo do Trabalho

O objetivo geral deste trabalho é avaliar modelos de recuperação de documentos visando detectar comportamentos atípicos no processo de instauração de Inquéritos Policiais. Para tanto, tais comportamentos foram definidos em três casos de uso onde serão contemplados como objetivos específicos. São eles:

1.2.1 Detecção de Inquéritos Duplicados

A Duplicação de Inquéritos é o ato de instaurar mais de uma IPL para o mesmo fato em questão. A principal consequência da duplicação diz respeito ao desperdício de recurso público, onde equipes distintas da PF acabam sendo alocadas para investigar o mesmo fato. Neste sentido, detectar duplicações tornaria a PF mais produtiva, pois menos duplicações implicaria em mais IPLs sendo conduzidas.

1.2.2 Detecção de Inquéritos Semelhantes

Inquéritos semelhantes são entendidos como documentos sintaticamente parecidos, onde estes possuem uma fração considerável de palavras em comum, ou também podem ser seman-

ticamente parecidos, onde apenas as ideias contidas nos documentos podem ser as mesmas, mas escritas com outras palavras. Uma forma de aplicação da recuperação de inquéritos semelhantes refere-se ao auxílio no treinamento de novos delegados e escrivães, uma vez que inquéritos semelhantes podem sugerir ações a serem executadas na condução de um novo IPL.

1.2.3 Detecção de Inquéritos Relacionados

Os Inquéritos Relacionados dizem respeito a IPLs instaurados para investigar fatos distintos, mas que de alguma maneira podem ter alguma relação. Como exemplo podemos citar um crime de homicídio e outro de roubo, onde estes possuem o mesmo suspeito. Nesse caso, ter ciência da relação entre os IPLs pode ajudar a solucionar ambos os casos. Vale lembrar que a relação entre IPLs possui um conceito mais subjetivo, onde os IPLs podem ser de naturezas bem diferentes e ainda sim ter algum tipo de relação.

1.3 Metodologia

Para alcançar os objetivos relativos a detectar inquéritos duplicados, semelhantes e relacionados, nós conduzimos um estudo em parceria com a PF e graduandos do curso de Ciência da Computação da Universidade Federal de Campina Grande (UFCG), compondo um time parte do projeto ePol, uma plataforma *web* que gerencia inquéritos policiais e interliga as unidades da PF, agilizando as investigações e prisões, dentre outras atividades realizadas pela polícia. No estudo foram utilizados dados de inquéritos policiais, técnicas de construção de modelos em RI e avaliações com o auxílio de especialistas para sugerir o uso ou não de modelos de RI que auxiliem a PF nas tarefas de detecção de duplicação, semelhança e relação de inquéritos.

Os dados utilizados na pesquisa são referentes a inquéritos não-sigilosos investigados pela Polícia Federal, registrados em 51 unidades, no período entre junho de 2016 e outubro de 2017, totalizando 1541 observações. Além dos dados estruturados, foram utilizadas 11818 Peças produzidas durante as investigações, sendo estas anexadas aos seus respectivos inquéritos. Na fase inicial, os documentos foram pré-processados para serem representados a partir de *tokens*. Nesta etapa foram utilizados unigramas, bigramas e trigramas, além de re-

gras de associação representadas por expressões regulares (formatos de CPF, nomes próprios em caixa-alta, etc) para a construção de cinco modelos: Aleatório, Similaridade de Cossenos, Distância de Jaccard, Doc2Vec e WMD. Em seguida, foram escolhidos 55 documentos do corpus para serem submetidos como entrada de cada modelo, onde estes retornaram seus 4 documentos mais relevantes, referentes a cada documento de entrada.

A partir das saídas para cada documento fornecido aos modelos como entrada, foram construídos questionários onde especialistas compararam se dois inquéritos eram duplicados, semelhantes ou relacionados. Foram utilizados 55 documentos como entrada em cada um dos 5 modelos, onde cada modelo retornou 4 documentos, totalizando 1100 comparações. Por conta do grande número de documentos a serem avaliados, os questionários foram aplicados diariamente contendo entre 8 a 12 comparações, levando cerca de 10 a 15 minutos para serem respondidos por especialistas. 1 delegado de 2 escrivães responderam os questionários.

Após a resposta dos questionários para construção do *ground truth*, foi possível realizar uma avaliação dos modelos, onde foi escolhido como métrica o NDCG (Normalized Discounted Cumulative Gain). Os resultados mostraram que: para a detecção de duplicação, o modelo WMD (0.18) superou os modelo clássico de distância de jaccard (0.16) e se equiparou ao de similaridade de cossenos (0.1786135312); para a detecção de inquéritos semelhantes, a distância de jaccard supera todos os outros métodos (0.88); para a detecção de inquéritos relacionados, a similaridade de cossenos (0.5161150486) se mostrou superior diante dos demais modelos. Por fim, neste contexto em que foram avaliados estes modelos, as abordagens clássicas se sobressaíram quando comparadas aos modelos de semântica distributiva. Entretanto, vale salientar que modelos semânticos prometem bons resultados quando se trabalha com uma imensa quantidade de dados.

1.4 Relevância

A relevância deste trabalho gira em torno de que a própria Polícia Federal reportou sobre dificuldades referentes a detecção de inquéritos duplicados, que gera um desperdício de recurso público mensurado em cerca de 30 mil reais em média por inquérito duplicado, além de atrasar o andamento de outras investigações. Outra tarefa que torna este trabalho relevante

neste contexto diz respeito ao auxílio na investigação de inquéritos conduzidos por novos delegados, onde estes podem aprender com ações tomadas em inquéritos semelhantes.

Este estudo também contribui para a área de Recuperação da Informação, onde é exposto um trabalho comparativo entre técnicas clássicas de do estado da arte, tratando de um problema não-supervisionado e com uma avaliação realizada por especialistas.

O restante desta dissertação está dividido da seguinte maneira: no próximo capítulo será apresentada a fundamentação teórica, composta pela definição de conceitos que envolvem o contexto policial, além de técnicas, métrica e algoritmos utilizados no estudo. Em seguida, um capítulo sobre trabalhos relacionados traz resumos de trabalhos recentes e importantes para o contexto. O capítulo subsequente descreve o estudo comparativo que realizamos, citando as questões de pesquisa, metodologia, descrição dos dados utilizados, préprocesamento dos dados, avaliação, resultados e conclusão. Um capítulo para Discussão dá continuidade ao documento levantando os principais pontos referentes ao trabalho, além a da interpretação dos resultados obtidos. Em seguida um capítulo de conclusão reune observações e lições aprendidas. Por fim, o capítulo sobre trabalhos futuros elencando os próximos passos da pesquisa finaliza o documento.

Capítulo 2

Fundamentação Teórica

2.1 Inquérito Policial

O Inquérito Policial (IPL) é o método policial administrativo fundamental no procedimento investigativo da Polícia Judiciária Brasileira. Ele apura certo crime e precede a ação penal, sendo usualmente considerado como pré-processual, apesar de estabelecer atividade em unidade com o processo penal. O IPL é composto de provas de autoria e materialidade de crime, que, comumente são produzidas por investigadores de polícia e peritos criminais, é organizado e numerado pelo Escrivão de Polícia Federal (EPF), e presidido pelo Delegado de Polícia Federal (DPF).

O IPL vem a ser o procedimento administrativo, preliminar, presidido pelo DPF, no intuito de identificar o autor do ilícito e os elementos que atestem a sua materialidade (existência), contribuindo para a formação da opinião delitiva do titular da ação penal, ou seja, fornecendo elementos para convencer o titular da ação penal se o processo deve ou não ser deflagrado. Pontue-se que a Lei nº 12.830/2013, ao dispor sobre a investigação criminal conduzida pelo delegado de polícia, deixa consignado que a apuração investigativa preliminar tem como objetivo apuração de circunstâncias, materialidade e autoria das infrações penais (art. 2º, § 1º) [ALENCAR and Távora, 2016].

O objetivo do IPL é a reunião de provas de ocorrência de infração penal, para permitir que a ação penal seja proposta. Cumpre frisar que o inquérito policial não acusa e não encerra um juízo de formação de culpa. Sua finalidade é assim meramente investigativa [FONTENELE, 2015].

Compreender o conceito de Inquérito, seus detalhes e peculiaridades ajuda a entender o objeto de estudo, além esclarecer que tipo de informação será considerada.

2.1.1 Peça

Documentos padronizados que são produzidos no processo investigatório, onde cada tipo de Peça possui uma finalidade na investigação. Alguns exemplos de Peças:

- Despacho - Peça produzida pelo Delegado de Polícia Federal (DPF) para instruir o Inquérito policial, utilizadas para solicitações e comunicação a outros setores da Polícia ou de outro órgão.
- Portaria - Peça produzida pelo DPF que dá início ao inquérito policial.
- Termo de Remessa - Peça produzida pelo Escrivão de Polícia Federal (EPF) e que encaminha o inquérito policial (IPL) para outro órgão.
- Certidão Tradicional - Peça produzida pelo EPF que relata qualquer fato que o mesmo tenha realizado, com a finalidade de comunicar algo que tenha sido efetivamente cumprido e que foi solicitado pelo Despacho ou algum fato importante para o IPL.
- Ofício - Peça produzida por qualquer policial (EPF ou DPF) encaminhada para outro órgão e que solicita ou comunica algum fato, com a finalidade de solicitar informações, cópias de documentos, apresentação de servidores a outros órgãos, etc.

  Por se tratar de elementos que compõem um Inquérito Policial, suas definições, formas de utilização e conteúdo podem ajudar na tarefa de identificar inquéritos relevantes.

2.1.2 Apensamento

Ação de anexar um IPL a outro por se tratarem do mesmo fato a ser investigado. Um apensamento geralmente ocorre quando se identifica inquéritos duplicados no sistema de gerenciamento da PF. Para a pesquisa o apensamento pode ser entendido como rótulos que definem as relações existentes entre os inquéritos na base da dados considerada.

2.1.3 ePol

O ePol é uma ferramenta *web* desenvolvida na Universidade Federal de Campina Grande (UFCG) em parceria com a Polícia Federal que gerencia inquéritos policiais e interliga as unidades da PF, agilizando as investigações e prisões, dentre outras atividades realizadas pela polícia. Por meio do sistema, os policiais fazem cruzamento de dados e acompanham em tempo real o andamento de inquéritos instaurados, com exceção dos que correm em segredo de Justiça [epo, 2013]. O ePol representa o contexto em que a pesquisa é realizada, pois é a partir desta ferramenta que são registrados os inquéritos utilizados no estudo.

2.2 Recuperação de Informação

A Recuperação de Informação (RI) se preocupa com a representação, busca e manipulação de grandes coleções de texto eletrônico e outros dados de linguagem humana. Os sistemas e serviços agora são generalizados, com milhões de pessoas dependendo diariamente deles para facilitar negócios, educação e entretenimento.

Os motores de busca da Web - Google, Bing e outros - são, de longe, os serviços de RI mais utilizados e altamente utilizados, fornecendo acesso a informações técnicas atualizadas, localizando pessoas e organizações, resumindo notícias e eventos e simplificando a comparação de compras. Os sistemas de bibliotecas digitais ajudam os pesquisadores médicos e acadêmicos a aprender sobre novos artigos de revistas e apresentações de conferências relacionadas às suas áreas de pesquisa. Os consumidores se voltam para os serviços de busca locais para encontrar revendedores que ofereçam produtos e serviços desejados. Em grandes empresas, os sistemas de pesquisa corporativa atuam como repositórios de *e-mails*, memorandos, relatórios técnicos e outros documentos comerciais, oferecendo memória corporativa preservando esses documentos e permitindo o acesso aos conhecimentos contidos neles. Os sistemas de pesquisa no desktop permitem aos usuários pesquisar seus *e-mails* pessoais, documentos e arquivos [Büttcher et al., 2016].

O conceito de recuperar documentos é importante para o estudo por representar a tarefa de recuperar inquéritos a partir de uma certa entrada.

2.3 Semântica Distributiva

Uma questão importante para o estudo do significado da palavra é planejar precisamente critérios de identidade para o conteúdo semântico das palavras. Na verdade, seguindo o bem conhecido preceito de Quine ”Nenhuma entidade sem identidade”(Quine 1969: 23), não podemos esperar investigar profundamente o significado lexical a menos que possamos especificar em que condições duas palavras podem ser dito ter o mesmo significado ou - se considerarmos a noção de sinonímia muito forte - ser semanticamente semelhante [Andrews et al., 2009].

Ou explicitamente ou implicitamente, a semelhança semântica tem um papel crucial em qualquer linguística e investigação psicológica sobre o significado. Ricas provas empíricas foram acumuladas sobre como o grau de semelhança semântica entre as palavras afetam a forma de como são processadas ou armazenadas no léxico mental [Baroni et al., 2007]. Além disso, quando baseamos nossas generalizações linguísticas na semântica de classes paradigmáticas de expressões, a classe de verbos de movimento ou a classe de substantivos abstratos, confiamos na semelhança semântica para identificar as expressões pertencentes à mesma classe.

A marca de qualquer modelo de semântica distributiva é o pressuposto de que a noção de semelhança semântica, juntamente com as outras generalizações baseadas nela, podem ser definidas em termos de distribuições linguísticas. Isto passou a ser conhecido como o Hipótese Distributiva (HD), que pode ser indicada no seguinte: O grau de semelhança semântica entre duas expressões linguísticas A e B é uma função da similaridade dos contextos linguísticos em que A e B podem aparecer [Lenci, 2008].

Este conceito se relaciona com o estudo por se preocupar em identificar semelhanças semânticas entre palavras, frases e documentos. Sendo assim, aplicações de métodos de similaridade baseados em semântica distributiva podem ajudar a identificar inquéritos similares a partir de seus documentos.

2.4 Word Embedding

Word Embedding é o termo utilizado para um conjunto de modelagem de linguagem e técnicas de aprendizado de características em processamento de linguagem natural (NLP) onde palavras ou frases do vocabulário são mapeadas para vetores de números reais. Conceitualmente, envolve uma incorporação matemática de um espaço com uma dimensão por palavra para um espaço vetorial contínuo com uma dimensão muito menor. Os métodos para gerar esse mapeamento incluem redes neurais, redução de dimensionalidade na matriz de co-ocorrência da palavra, modelos probabilísticos e representação explícita em termos do contexto em que as palavras aparecem. Os vetores de pensamento (ou *thought vectors*) são uma extensão de *word embeddings* para frases inteiras ou mesmo documentos. Alguns pesquisadores esperam que estes possam melhorar a qualidade da tradução automática [Wikipedia, 2017b].

Este é um conceito que vem sendo bastante explorado nos últimos anos e que facilita a aplicação de medidas de similaridade entre termos, frases ou documentos. Alguns trabalhos citados na seção a seguir utilizam desta definição para representar e medir a similaridade semântica entre itens. Espera-se fazer uso deste conceito para representar os inquéritos a partir de seus documentos e com isso medir a similaridade entre eles.

2.5 Modelos

Por se tratar de um estudo comparativo, foram utilizados modelos conhecidos em similaridade textual visando medir o quão próximos os documentos estão uns dos outros. Com essa medida, foi possível definir quais documentos estão próximos de uma dada entrada e assim, compor um sistema de recuperação de documentos baseado nos modelos.

A escolha dos modelos se deu pela busca tanto por abordagens sintáticas quanto semânticas. Os modelos que focam apenas em sintaxe medem o quão similares dois documentos podem ser baseando-se na quantidade de *tokens* em comum. Já modelos semânticos levam em consideração o contexto em que os *tokens* aparece.

2.5.1 Similaridade de Cosseno

Similaridade de Cosseno é uma medida entre dois vetores não-nulos entre um espaço de produto interno, sendo calculadada através do cosseno do ângulo formado entre os vetores. O cosseno de 0 grau é igual a 1, e menor que 1 para qualquer ângulo num intervalo (0, pi] radianos. Dois vetores com a mesma orientação possuem a similaridade de cossenos igual a 1. Para vetores orientados a 90 graus um do outro terão similaridade igual a zero. Já dois vetores orientados diametralmente possuem similaridade igual a -1, independente de suas magnitudes. A semelhança de cosseno é particularmente usada no espaço positivo, onde o resultado é nitidamente limitado em [0, 1]. O nome deriva do termo ”direção do cosseno”, onde vetores unitários são maximamente ”similares”se forem paralelos e maximamente ”dissimilares”se forem ortogonais (perpendiculares). Isso é análogo ao cosseno, que terá valor 1 (valor máximo) quando os segmentos subtendem um ângulo zero, e valor zero (não correlacionados) quando os segmentos são perpendiculares.

Esses limites se aplicam a qualquer número de dimensões, e a similaridade de cosseno é comumente utilizada em espaços positivos de alta dimensão. Na recuperação de informação e mineração de texto, cada termo é atribuído a uma dimensão diferente e um documento é caracterizado por um vetor em que o valor em cada dimensão corresponde ao número de vezes que o termo aparece no documento. A semelhança do cosseno fornece uma medida útil de como dois documentos são passíveis de semelhança a partir de seu conteúdo [Singhal et al., 2001]. Esta técnica também é utilizada para medir a coesão dentro de *clusters* no campo da mineração de dados [Tan et al., 2005].

O termo ”distância de cosseno”é frequentemente utilizado para o complemento no espaço positivo, ou seja:

$$Dc(A, B) = 1 - Sc(A, B), onde : \tag{2.1}$$

- A - vetor A
- B - vetor B
- Dc - Distância de Cosseno
- Sc - Similaridade de Cosseno

No entanto, é importante notar que esta não é uma métrica de distância adequada, pois não tem a propriedade de desigualdade triangular - ou, mais formalmente, a desigualdade de Schwarz - e viola o axioma da coincidência; para reparar a propriedade de desigualdade do triângulo, mantendo a mesma ordem, é necessário converter a distância angular.

O cosseno de dois vetores diferentes de zero pode ser derivado usando a fórmula de produto escalar:

$$A.B = \|A|\ \|B\| \cos\theta \tag{2.2}$$

Dado dois vetores de atributos, A e B, a semelhança de cosseno é representada usando um produto escalar e magnitude como sendo:

$$similaridade = \cos\theta = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, onde: \tag{2.3}$$

- A_i e B_i são componentes dos vetores A e B.
- i - refere-se a um componente presente nos vetores. Também pode ser entendido como um *token* pertencente ao corpus, que agora assume um papel de uma dimensão em um espaço vetorial.
- n - número de componentes no espaço vetorial. Neste contexto podemos dizer que será o número de *tokens* presente em todo o corpus.

A similaridade resultante varia de -1, significando exatamente o oposto, para 1, significando exatamente o mesmo, com 0 indicando ortogonalidade ou descorrelação, enquanto entre tais valores indica uma similaridade intermediária ou dissimilaridade.

Para correspondência entre documentos, os vetores de atributo A e B são geralmente os vetores de termo dos documentos. A semelhança do cosseno pode ser vista como um método de normalizar o comprimento do documento durante a comparação.

No caso da recuperação da informação, a similaridade de dois documentos varia de 0 a 1, já que a frequência dos termos não pode ser negativa. O ângulo entre dois vetores de documentos não pode ser maior que 90 graus.

2.5.2 Distância de Jaccard

O índice de Jaccard, também conhecido como Intersecção sobre União, e o coeficiente de similaridade de Jaccard (originalmente dado o nome francês coeficiente de communauté por Paul Jaccard), é uma estatística usada para aferir a similaridade e diversidade de conjuntos de amostras. O coeficiente de Jaccard mede a similaridade entre conjuntos de amostras finitas e é definido como o tamanho da interseção dividido pelo tamanho da união dos conjuntos de amostra:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{2.4}$$

(Se A e B estiverem vazios, definimos $J(A,B) = 1$.)

$$0 <= J(A,B) <= 1. \tag{2.5}$$

A distância Jaccard, que mede a dissimilaridade entre conjuntos de amostras, é o complementar ao coeficiente de Jaccard e é obtida subtraindo-se o coeficiente de Jaccard de 1, ou, equivalentemente, dividindo-se a diferença dos tamanhos da união e a interseção de dois conjuntos pelo tamanho da união:

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

2.5.3 Doc2Vec

O Doc2Vec é um modelo proposto como eficiente abordagem em redes neurais para o aprendizado de *embeddings* para representar conjuntos de palavras. Ele foi criado a partir de uma variação de outro modelo conhecido como Word2Vec, onde ambos possuem a mesma base de construção. A principal diferença entre eles é que o word2vec representa uma palavra a partir de um vetor, enquanto que o Doc2Vec representa o documento (formado por um conjunto de palavras) por um vetor. Em todos os modelos, a proposta é utilizar a matriz de pesos criada entre a camada escondida e a entrada para representar os documentos. O modelo criado não é utilizado após o treinamento.

Existem duas abordagens de construção para o Doc2Vec: dbow e dmpv. O dbow funciona da mesma forma que o skip-gram, exceto que a entrada é substituída por um token especial representando o documento. Nesta arquitetura, a ordem das palavras no documento é ignorada, caracterizando assim o nome *bag of words*. Já o dmpv funciona de maneira semelhante ao cbow. Para uma dada entrada, o dmpv introduz um documento adicional *token*, além de várias palavras de destino. Ao contrário do cbow, no entanto, esses vetores não são somados, mas concatenados. O objetivo é novamente prever o contexto de uma palavra dado o documento concatenado e um vetor de palavras.

A princípio, é definida uma tarefa para que uma rede neural seja treinada. Neste caso queremos predizer uma palavra dado seu contexto. Exemplo:

O Serviço Público Federal em Campina Grande.

Na frase acima, definimos a tarefa de predizer cada palavra nesta frase a partir do seu contexto. Dada uma palavra específica no meio de uma frase, temos como entrada da rede o conjunto de palavras que está ao redor desta palavra, e como saída a própria palavra. A rede irá informar a probabilidade de um conjunto de palavras (documento) ser a "palavra mais próxima" que escolhemos. Neste exemplo, vamos definir contexto como sendo as palavras que aparecem o mais próximas da palavra alvo. Com isso, temos:

id	**t1**	**t2**	**target**
1234	O	Público	Serviço
8765	Serviço	Federal	Público
0989	Público	em	Federal
0910	Federal	Campina	em
0911	em	Grande	Campina

Tabela 2.1: Exemplo de representação de um contextos de palavras e suas palavras centrais como target

A rede aprenderá as estatísticas do número de vezes que cada pareamento aparece. Dessa forma, a rede provavelmente obterá muito mais amostras de treinamento ("Serviço", "Federal") com alvo em "Público"do que de ("Serviço", "Federal") com alvo em "O". Após o treinamento do modelo, a predição para o documento ("Serviço", "Federal") como entrada

produzirá uma probabilidade muito maior para a palavra “Público” do que para “O”.

Para a construção do modelo, foi criado um vocabulário de palavras a partir dos documentos utilizados na pesquisa. Os 1542 inquéritos utilizados são compostos um vocabulário de 37 mil palavras. Cada documento foi representado como um vetor binário, utilizando codificação *one-hot*, e cada componente do vetor refere-se a uma palavra definida no vocabulário do *corpus*, onde possuirá o valor 1 na posição correspondente à palavra presente no documento e 0s em todas as outras posições, além de 37 mil componentes.

A saída da rede é um único vetor (também com 37 mil componentes) contendo, para cada palavra do vocabulário, a probabilidade de que uma palavra próxima selecionada aleatoriamente seja aquela palavra do vocabulário.

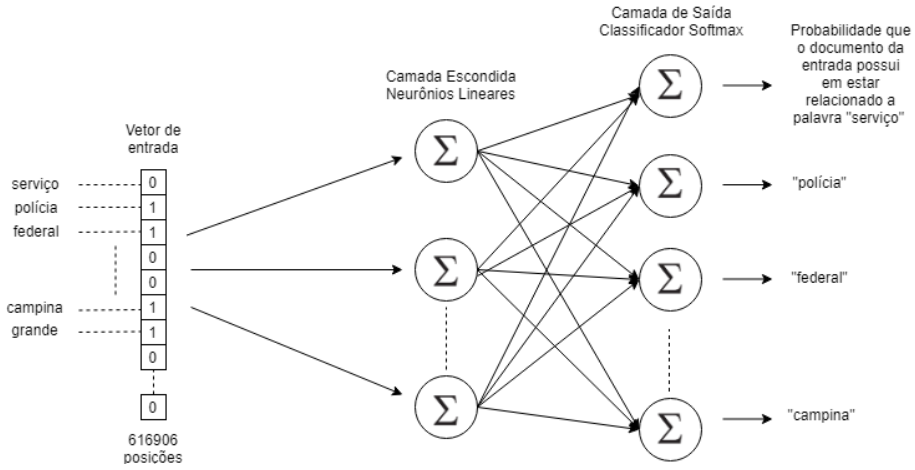


Figura 2.1: arquitetura de modelo doc2vec

Outro detalhe importante sobre a arquitetura diz respeito aos neurônios da camada escondida, onde estes não possuem função de ativação. Já os neurônios da camada de saída utilizam a função softmax.

Neste experimento, foram aprendidos vetores de documentos com 300 *features*. Esta tarefa é realizada através da camada escondida, onde esta será representada por uma matriz de pesos com 1542 observações e 300 colunas. Cada observação nesta matriz representará um documento do estudo. O número de colunas é definido a partir de um hiperparâmetro do modelo, podendo ser um valor arbitrário e sem necessariamente ter uma relação direta com os dados. O valor padrão escolhido é de 300, tendo em vista que este é número utilizado por um trabalho de sucesso publicado pelo Google, utilizando dados de notícias [Mikolov et al.,

2013].

Após o treinamento do modelo, e a representação dos documentos pela matriz de pesos aprendida, foi possível calcular a similaridade entre os vetores a partir do cálculo da distância euclidiana. Com isso, um *ranking* de documentos similares foi construído para cada documento.

2.5.4 Distância do Motor de Palavras

O modelo Distância do Motor de Palavras (WMD) é um método que nos permite avaliar a distância entre dois documentos de uma forma significativa, mesmo quando eles não têm palavras em comum. Ele usa o modelo word2vec para criação *embeddings* de vetores de palavras, e mostrou-se ter melhores desempenhos que muitos dos métodos avançados na classificação de k-vizinhos mais próximos [Kusner et al., 2015].

O WMD é ilustrado abaixo por duas frases muito semelhantes (ilustração tirada do blog de Vlad Niculae). As sentenças não têm palavras em comum, mas ao combinar as palavras relevantes, o WMD é capaz de medir com precisão a (des) similaridade entre as duas frases. O método também usa a representação em forma de *bag of words* dos documentos, anotada como d na figura abaixo. A intenção por trás do método é que encontramos a ”distância de deslocamento”mínima entre os documentos, ou seja, a maneira mais eficiente de ”mover”a distribuição do documento 1 para a distribuição do documento 2.

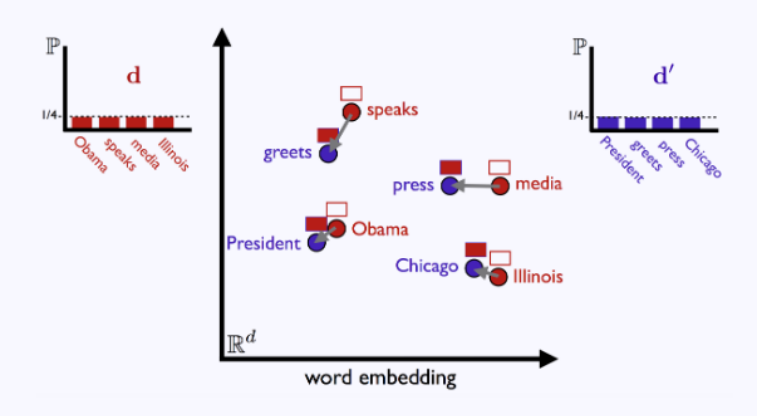


Figura 2.2: ilustração tirada do blog de Vlad Niculae

2.5.5 Modelo Aleatório

O modelo aleatório retorna um conjunto de documentos de forma aleatória para uma dada entrada qualquer. A intenção deste modelo é de ter apenas uma base comparativa com os outros modelos propostos. Caso um dado modelo X tenha desempenho similar ou pior ao aleatório, será um indicativo de que o modelo X não possui utilidade no problema proposto.

2.6 Métrica de Avaliação: NDCG

2.6.1 Discounted Cumulative Gain (DCG)

O DCG é uma medida de qualidade para um *ranking*. Na Recuperação de Informação é frequentemente utilizada para medir a eficácia dos algoritmos de pesquisa na *Web* ou aplicativos relacionados. Usando uma escala de relevância gradativa de documentos em um conjunto de resultados fornecidos por um modelo, o DCG mede a utilidade ou o ganho de um documento com base em sua posição na lista de resultados. O ganho é acumulado do topo da lista de resultados até o final, com o ganho de cada resultado descontado em classificações mais baixas [Järvelin and Kekäläinen, 2002].

Duas hipóteses são feitas no uso do DCG e suas medidas relacionadas:

- Documentos altamente relevantes são mais úteis quando aparecem mais cedo em uma lista de resultados de mecanismos de pesquisa (possuem classificações mais altas).
- Documentos altamente relevantes são mais úteis do que documentos marginalmente relevantes, que por sua vez são mais úteis do que documentos não relevantes.

A premissa do DCG é que os documentos altamente relevantes que aparecem mais abaixo em uma lista de resultados de pesquisa devem ser penalizados, pois o valor de relevância gradual é reduzido logaritmicamente proporcional à posição do resultado.

A fórmula tradicional do DCG acumulada em uma determinada posição de classificação p é definida como:

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2 (i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 (i+1)} \tag{2.6}$$

Anteriormente, não havia justificativa teoricamente sólida para utilizar um fator de redução logarítmico, além do fato de que este produz uma redução suave. Mas Wang et al. (2013) [Wu and Wang, 2017] dão garantia teórica para o uso do fator de redução logarítmica através do NDCG. Os autores mostram que para cada par de funções de classificação substancialmente diferentes, o NDCG pode decidir qual é o melhor de uma maneira consistente. Com isso, uma formulação alternativa ao DCG enfatiza mais fortemente a recuperação de documentos relevantes:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2 (i + 1)} \tag{2.7}$$

A última fórmula é comumente utilizada na indústria, incluindo grandes empresas de busca na *web* e plataformas de competição de ciência de dados, como a Kaggle [kag, 2010].

Essas duas formulações de DCG são as mesmas quando os valores de relevância dos documentos são binários: $rel_i \in \{0, 1\}$.

2.6.2 Normalized Discounted Cumulative Gain (NDCG)

As listas de resultados da pesquisa variam em tamanho, dependendo da consulta. Comparar o desempenho de um mecanismo de pesquisa de uma consulta para a próxima não pode ser alcançado consistentemente utilizando apenas o DCG, portanto, o ganho cumulativo em cada posição para um valor de p escolhido deve ser normalizado nas consultas. Isso é feito classificando todos os documentos relevantes no *corpus* por sua relevância relativa, produzindo o máximo possível de DCG por meio da posição p, também chamada Ideal DCG (IDCG) através dessa posição. Para uma consulta, o ganho cumulativo descontado normalizado, ou nDCG, é calculado como:

$$nDCG_p = \frac{DCG_p}{IDCG_p}, \tag{2.8}$$

onde o $IDCG_p$ é o valor de DCG ideal,

$$IDCG_p = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i} - 1}{\log_2 (i + 1)}, \tag{2.9}$$

e o REL_p representa a lista de documentos relevantes, ordenados por sua relevância, no *corpus* até a posição p.

Os valores de nDCG para todas as consultas podem ser calculados para obter uma medida do desempenho médio do algoritmo de classificação de um mecanismo de pesquisa. Em um algoritmo de classificação perfeita, o DCG_p será o mesmo que o $IDCG_p$, produzindo um nDCG de 1.0. Todos os cálculos de nDCG são valores relativos no intervalo de 0,0 a 1,0 e, portanto, são comparáveis entre consultas.

A principal dificuldade encontrada no uso de nDCG é a indisponibilidade de uma ordenação ideal de resultados quando apenas o feedback de relevância parcial está disponível.

Capítulo 3

Estudo Comparativo

3.1 Questões de Pesquisa

Uma das tarefas mais recorrentes realizada pelos investigadores ocorre no processo de instauração de inquéritos, onde o responsável deve verificar se já existe um procedimento de investigação criminal para o fato em questão a fim de evitar duplicação. Além disso, existe uma dificuldade em lidar com o grande volume de inquéritos, onde os servidores da PF não fazem uso de forma sistemática da base de dados históricos de suas investigações. Partindo destes pontos, foram discutidos com a PF possíveis casos de uso para serem aplicados aos dados, que derivaram a seguinte questão de pesquisa:

Dado um Inquérito, é possível encontrar inquéritos duplicados, semelhantes ou relacionados?

A Duplicação de Inquéritos é o ato de instaurar mais de uma IPL para o mesmo fato em questão. A principal consequência da duplicação diz respeito ao desperdício de recurso público, onde equipes distintas da PF acabam sendo alocadas para investigar o mesmo fato. Neste sentido, detectar duplicações tornaria a PF mais produtiva, pois menos duplicações implicaria em mais IPLs sendo conduzidas.

Inquéritos semelhantes são entendidos como documentos sintaticamente parecidos, onde estes possuem uma fração considerável de palavras em comum, ou também podem ser semanticamente parecidos, onde apenas as ideias contidas nos documentos podem ser as mes-

mas, mas escritas com outras palavras. Uma forma de aplicação da recuperação de inquéritos semelhantes refere-se ao auxílio no treinamento de novos delegados e escrivães, uma vez que inquéritos semelhantes podem sugerir ações a serem executadas na condução de um novo IPL.

Os Inquéritos Relacionados dizem respeito a IPLs instaurados para investigar fatos distintos, mas que de alguma maneira podem ter alguma relação. Como exemplo podemos citar um crime de homicídio e outro de roubo, onde estes possuem o mesmo suspeito. Nesse caso, ter ciência da relação entre os IPLs pode ajudar a solucionar ambos os casos. Vale lembrar que a relação entre IPLs possui um conceito mais subjetivo, onde os IPLs podem ser de naturezas bem diferentes e ainda sim ter algum tipo de relação.

3.2 Metodologia

Para realização deste estudo, foram utilizados dados referentes a inquéritos não-sigilosos investigados pela Polícia Federal, registrados em 51 unidades, no período entre junho de 2016 e outubro de 2017, totalizando 1541 observações. Além dos dados estruturados, foram utilizadas 11818 Peças produzidas durante as investigações, sendo estas anexadas aos seus respectivos inquéritos. Na fase inicial, os documentos foram pré-processados para serem representados a partir de *tokens*. Nesta etapa foram utilizados unigramas, bigramas e trigramas, além de regras de associação representadas por expressões regulares (formatos de CPF, nomes próprios em caixa-alta, etc) para a construção de cinco modelos: Similaridade de Cosseno, Distância de Jaccard, Doc2Vec e WMD e um modelo Aleatório. Em seguida, foram escolhidos 55 documentos do *corpus* para serem submetidos como entrada de cada modelo, onde estes retornaram seus 4 documentos mais relevantes, referentes a cada documento de entrada.

A partir das saídas para cada documento fornecido aos modelos como entrada, foram construídos questionários onde especialistas compararam se dois inquéritos eram duplicados, semelhantes ou relacionados. Foram utilizados 55 documentos como entrada em cada um dos 5 modelos, onde cada modelo retornou 4 documentos, totalizando 1100 comparações.

Por conta do grande número de documentos a serem avaliados, os questionários foram aplicados diariamente contendo entre 8 a 12 comparações, levando cerca de 10 a 15 minutos

para serem respondidos por especialistas. 1 delegado de 2 escrivães responderam os questionários. Como métrica de avaliação foi utilizado nDCG, sendo esta comum na área de RI para avaliar resultados de consultas a documentos e que considera a ordem em que estes foram retornados.

3.3 Sobre os Dados

As principais informações sobre um inquérito são armazenadas em uma Tabela chamada Casos. O conteúdo utilizado desta Tabela é referente a um atributo textual com o resumo do Caso, além de um atributo identificador.

Tabela 3.1: Campos referentes aos Dados de Inquéritos utilizados no estudo.

Campo	Descrição
DS-RESUMO	Texto referente ao resumo do Caso
CD-CASO	Código do Caso que a Peça pertence

Outra Tabela utilizada refere-se a Peças que compõem um inquérito.

Tabela 3.2: Campos referentes aos Dados de Peças utilizados no estudo.

Campo	Descrição
CD-CASO	Código do Caso que a Peça pertence
ID-PECA	Código identificador da Peça
MM-CONTEUDO-HTML	Conteúdo da Peça gerado em HTML
NO-PECA	Nome da Peça

Após o acesso aos documentos referentes a Casos e Peças, foi feito a união entre o conteúdo textual dos casos e suas respectivas peças, visando unificar todo texto referente a um inquérito em um único documento. Isso facilitou a manipulação dos documentos, tendo em vista que um inquérito fica representado por uma única *string*. A figura C.1 encontrada no apêndice C exemplifica este processo.

3.3.1 Pré-processamento

Inicialmente, os documentos foram pré-processados para serem representados a partir de um processo conhecido como Tokenização. Dada uma sequência de caracteres e uma unidade de documento definida, a tokenização é a tarefa de ”cortá-la em pedaços”, chamados de *tokens*, podendo ser excluídos alguns caracteres irrelevantes ao contexto, a exemplo de pontuações.

Documento	**Lista de Tokens**
SERVIÇO PÚBLICO FEDERAL MJ - DEPARTAMENTO DE POLÍCIA FEDERAL DPF/CGE/PB - DELEGACIA DE POLÍCIA FEDERAL EM CAMPINA GRANDE	[”SERVIÇO”, ”PÚBLICO”, ”FEDERAL”, ”MJ”, ”DEPARTAMENTO”, ”DE”, ”POLÍCIA”, ”FEDERAL”, ”DPF”, ”CGE”, ”PB”, ”DELEGACIA”, ”DE”, ”POLÍCIA”, ”FEDERAL”, ”EM”, ”CAMPINA”, ”GRANDE”]

Tabela 3.3: Exemplo de tokenização de um documento

Os *tokens* também podem possuir mais de uma palavra, caracterizando assim os n-gramas. Um n-grama é uma sequência contínua de n itens de uma dada amostra de texto ou fala. Os itens podem ser fonemas, sílabas, letras, palavras ou pares de bases de acordo com a aplicação. Os n-gramas são tipicamente coletados de um texto ou *corpus* de fala. Usando prefixos numéricos latinos, um n-grama de tamanho 1 é referido como um ”unigrama”; tamanho 2 é um ”bigrama”(ou, menos comumente, um ”digrama”); tamanho 3 é um ”trigrama”. Neste trabalho os documentos foram representados em parte por unigramas, bigramas e trigramas.

Documento	**n-gramas**
SERVIÇO PÚBLICO FEDERAL MJ -	[”SERVIÇO”, ”PÚBLICO”, ”FEDERAL”, ”MJ”, ”SERVIÇO PÚBLICO”, ”PÚBLICO FEDERAL”, ”FEDERAL MJ”, ”SERVIÇO PÚBLICO FEDERAL”, ”PÚBLICO FEDERAL MJ”]

Tabela 3.4: unigramas, bigramas e trigramas de um documento

Outra forma de representação dos documentos utilizada diz respeito a padrões bem conhecidos no texto, como formatos de telefones, cpfs, cnpjs, datas, padrões de códigos utilizados pela PF e formatos de escrita de nomes próprios. As informações que seguiam esses formatos foram capturadas a partir de expressões regulares. Uma expressão regular, regex ou regexp é uma sequência de caracteres que define um padrão de pesquisa em texto. A Tabela a seguir mostra as expressões regulares utilizadas para extração de *tokens*:

Dado	**regex**
CPF 1	`[0-9]{3}.[0-9]{3}.[0-9]{3}-[0-9]{2}`
CPF 2	`[0-9]{3}.[0-9]{3}.[0-9]{3}.[0-9]{2}`
CPF 3	`[0-9]{11}`
CPF 4	`[0-9]{9}-[0-9]{2}`
CPF 5	`[0-9]{9}.[0-9]{2}`
IPL	`[0-9]{4}.[0-9]{7}-[A-Z]{3}/[A-Z]{3}/[A-Z]{2}`
telefone	`[0-9]{4}-[0-9]{4}`
data 1	`[0-9]{2}/[0-9]{2}/[0-9]{4}`
data 2	`[0-9]{2}/[0-9]{2}/[0-9]{2}`

Tabela 3.5: Expressões Regulares aplicadas nos documentos

Após a extração dos *tokens* de cada documento, foi criado um espaço vetorial formado pelo conjunto de todos os *tokens* presentes no *corpus*. Nele foram representados todos os documentos já pré-processados, onde o valor presente em cada dimensão corresponde a frequência do *token* no documento em questão.

id	**SERVICO**	**PUBLICO**	**JOSE DA SILVA**	**ASSALTO A CAIXA**	**PEDRO CAVAL-CANTE**	**CAMPINA GRANDE**
1234	1	1	1	1	0	1
8765	1	1	0	1	1	1
0989	1	1	0	0	0	0

Tabela 3.6: Exemplo de representação vetorial dos documentos

3.4 Modelos

Após a etapa de pré-processamento, os dados foram submetidos para construção de quatro modelos: Similaridade de Jaccard, Distância de Cosseno, Doc2Vec e WMD. Para os modelos similaridade de jaccard e distância de cosseno, os *scores* de similaridade entre os inquéritos foram calculados a partir de suas respectivas fórmulas descritas no capítulo 2 referente a Fundamentação Teórica. Já para os modelos doc2vec e WMD, os inquéritos foram passados em um formato de *string*, sem pré-processamento para que os modelos fossem treinados de acordo com as etapas também descritas na Fundamentação Teórica. A não utilização de pré-processamento se deve ao fato de que estes tipos de modelos fazem suas próprias manipulações textuais e necessitam dos documentos completos para aprender a relação semântica entre as palavras. Por fim, os *scores* de similaridade para estes modelos foram calculados a partir da distância euclidiana entre as representações vetorias dos inquéritos.

Escolhemos 55 inquéritos para realizar consultas nos 5 modelos e consideramos como resposta de cada consulta os 4 inquéritos com maior *score* de similaridade. Além dos 4 modelos descritos, um modelo aleatório foi utilizados para base de comparação, onde o esperado

é que os modelos propostos obtivessem resultados melhores que o modelo aleatório.

A seguir um exemplo de respostas da mesma consulta realizada em cada um dos modelos. O CDCASO refere-se ao código do inquérito passado como consulta. As demais colunas representam os modelos considerados no estudo. Cada modelo retorna uma lista contendo 4 códigos de inquéritos para cada consulta realizada.

CDCASO	**aleatorio**	**doc2vec**	**jaccard**	**wmd**	**cosine**
15852	[10721, 2461, 63, 1863]	[15845, 15590, 15846, 15848]	[15851, 15854, 15826, 15864]	[15851, 15826, 5142, 1802]	[15851, 15846, 15854, 15847]

Tabela 3.7: Exemplo de respostas de cada modelo para uma dada consulta. Os modelos retornam uma lista de códigos de inquéritos para um código de inquério dado como entrada

Com as respostas dos modelos para as 55 consultas realizadas, foram construídos formulários que comparavam a entrada de uma consulta com as saídas

3.5 Avaliação

3.5.1 Formulário Estruturado

Na avaliação foi utilizado um formulário estruturado como instrumento de pesquisa. Este formulário foi composto por 12 comparações de Inquéritos, onde especialistas responderam de forma objetiva sobre três quesitos em cada comparação. É possível observar no Apêndice A um exemplo de quesito contido no questionário.

Foram utilizados 55 inquéritos como entrada para cada modelo citado, onde estes retornaram seu *top* 4 documentos para cada consulta, ordenados por relevância. 1100 avaliações foram realizadas por 3 especialistas, sendo estes 1 delegado e 2 escrivães. As avaliações ocorreram diariamente, geralmente pela manhã, levando cerca de 10 minutos para serem respondidas. Cada especialista respondeu 1 formulário por dia. Um exemplo de comparação contida nos formulários se encontra no Apêncice A.

A pergunta referente a semelhança dos inquéritos, por existir a noção de níveis de semelhança (muito ou pouco semelhante), foi extraída a partir de 4 níveis: ”muito semelhante”, ”semelhante”, ”pouco semelhante”e ”não é semelhante”. Com isso tornou-se possível uma avaliação dos itens considerando a ordem de importância em que uma consulta retornou seus documentos. Esta ordem de importância foi definida mapeando os níveis em valores discretos de 0 a 3, onde 0 significa ”não é semelhante”e 3 ”muito semelhante”.

Para os quesitos referentes a duplicação e relação de inquéritos, as opções de resposta possíveis foram binárias, tendo em vista que não existe o conceito de níveis de duplicação ou relação entre inquéritos (”pouco duplicado”, ”muito duplicado”, ”pouco relacionado”ou ”muito relacionado”) no contexto definido.

Após a coleta das respostas de cada formulário, foi construído um *dataset* contendo cada comparação realizada no formulário estruturado. É possível observar um exemplo do *dataset* nas Tabelas B.1 e B.2 do Apêndice B. Em seguida foi feita uma formatação deste agrupando as repostas nas colunas ”semelhante”, ”duplicado”e ”relacionado”por modelo e queryid, para representar as respostas de cada consulta como listas com 4 valores. Desse modo, as observações foram filtradas por modelo e para ser calculada métrica de avalição para as colunas ”semelhante”, ”duplicado”e ”relacionado”.

3.5.2 Métrica de Avaliação: NDCG

O NDCG foi escolhido como métrica de avaliação por calcular um *score* para um grupo de respostas a consultas onde são consideradas a ordem de importância dos itens retornados. A descrição mais detalhada do NDCG pode ser encontrada no capítulo 2.

De acordo com o exemplo de *dataset* contido na Tabela B.3 do Apêndice B, o NDCG foi calculado para cada modelo e caso de uso, onde foram utilizadas as consultas representadas pelos valores da coluna ”queryid”, seus resultados nas colunas ”semelhante”, ”duplicado”e ”relacionado”, além do *ranking* em que cada item aparece como retorno da consulta, onde este último ficou representado pelo índice do valores em suas listas.

3.6 Resultados

Os resultados a seguir estão divididos para cada caso de uso definido anteriormente. Os dados utilizados lidam com um problema não-supervisionado. Com isso, não foi possível utilizar métricas como Precisão, Revocação e derivados, dado que não são conhecidos todos os IPLs duplicados, semelhantes ou relacionados. A avaliação ficou por conta de especialistas que julgaram o retorno das consultas realizadas em cada modelo, possibilitando assim o cálculo da métrica nDCG.

3.6.1 Semelhança de IPLs

A detecção de IPLs semelhantes visa descobrir inquéritos instaurados que possuem mesma natureza, seja no fato ou também na sua forma de conduzir uma investigação. Os resultados para cada modelo são exibidos a seguir:

modelo	**nDCG**
similaridade de jaccard	0.8812
distância de cosseno	0.8371
wmd	0.8037
doc2vec	0.6743

Tabela 3.8: Cálculo do nDCG para as consultas em cada modelo, em relação aos casos semelhantes

Na Tabela acima é possível perceber o bom desempenho de quase todos os modelos, excetuando o doc2vec com um *score* um pouco distante dos demais. Um fato interessante diz respeito aos melhores modelos serem de natureza sintática (similaridade de jaccard e distância de cosseno), onde estes consideram apenas o quão grande é a intersecção dos *tokens* entre os documentos e o quão próximos os documentos estão dentro de um espaço construído com *tokens* presentes no *corpus*. A interpretação destes resultados é compreendida ao perceber que o *corpus* possui muitas IPLs que tratam de crimes de mesma natureza como assaltos a caixa eletrônico, ou crimes contra o INSS. Além disso, os documentos possuem uma forma bem definida em sua escrita, onde esta quase sempre é obedecida.

A Tabela a seguir mostra resultados de consultas contendo IPLs avaliados como **muito semelhantes** aos seus respectivos IPLs fornecidos como entrada nas consultas, e que foram apontados pelos três modelos com melhor desempenho. A coluna IPLConsulta refere-se ao identificador do IPL utilizado como consulta. A coluna IPLsSemelhantes mostra os inquéritos que foram avaliados como muito semelhantes em relação a sua entrada. A coluna Modelos mostra quais modelos detectaram a duplicação.

IPLConsulta	**IPLsSemelhantes**	**Modelos**
303	302	cosseno, jaccard, wmd
1526	3662	cosseno, jaccard, wmd
2561	2581	cosseno, jaccard, wmd
2741	3142	cosseno, jaccard, wmd
3662	1526	cosseno, jaccard, wmd
3701	3721	cosseno, jaccard, wmd
8082	2841, 8084, 8101	cosseno, jaccard, wmd
8084	2841, 8082, 8101	cosseno, jaccard, wmd
8101	2841, 8082, 8084	cosseno, jaccard, wmd
15826	15852	cosseno, jaccard, wmd
15851	15852	cosseno, jaccard, wmd
15852	15851	cosseno, jaccard, wmd
16986	16987	cosseno, jaccard, wmd
16987	16986	cosseno, jaccard, wmd
17308	18964	cosseno, jaccard, wmd
18964	17308	cosseno, jaccard, wmd

Tabela 3.9: IPLs avaliados como muito semelhantes em retornos de consultas por modelos

Na Tabela acima, é possível observar a consulta com o IPL 303 como entrada e saída 302, onde este resultado também foi classificado como Duplicação de IPLs, mostrando que o IPL classificado como muito semelhante (302) trata do mesmo fato abordado no documento de entrada (303), deixando claro que ambos são escritos de forma bastante parecida. As

consultas realizadas com os inquéritos 2561, 8082, 8084, 8101 e 17308 são mais exemplos de IPLs semelhantes que também foram classificados como Duplicação, o que já mostra pela avaliação que um IPL considerado Duplicado deve ser muito semelhante a sua duplicata.

A consulta 1526 teve como resposta outro IPL (3662) que trata de um crime bastante semelhante, onde ambos se referem a crimes contra a ordem tributária, através de omissão de receita. Entretanto, é possível observar que os IPLs tratam de Fatos distintos, que ocorreram em períodos diferentes e com envolvidos distintos. Este é um típico exemplo de IPLs semelhantes, onde tratam de Fatos bem parecidos, mas que não são os mesmos. Outro ponto interessante é que o IPL 3662 quando passado como entrada da consulta, tem o IPL 1526 em seu *top* 4 como resultado.

A consulta com o Inquérito 2741 teve em seu *top* 4 o IPL 3142. Ambos tratam de ”Arrombamento à Agência dos Correios”, que ocorreram em cidades distintas. Nos dois fatos ocorridos os responsáveis das agências também realizaram ações parecidas, que se resumem a uma breve vistoria para confirmar o arrombamento.

Os IPLs 3701, 15826, 15851 e 15852 tiveram em suas respostas inquéritos semelhantes, onde todos tratam de ”suposto crime de estelionato, devido recebimento pós-óbito de benefícios previdenciários”. A diferença desses IPLs está na titularidade do beneficiário e o número do benefício.

Os IPLs 16986 e 16987 possuem resumos bem parecidos, ambos tratando de um ”Desmembramento de IPL, subsistindo a investigação em relação à apuração de possíveis crimes previstos na Lei 8.666/93 e desvios de verbas públicas por parte de funcionários públicos municipais ou por Prefeitos”. As diferenças entre os IPLs estão nas prefeituras investigadas, nos tipos de obras realizadas, no valor da obra e na data em que foram realizadas.

3.6.2 Relação entre IPLs

A detecção de IPLs relacionados busca por inquéritos instaurados que possuem algum tipo de ligação. Essa ligação não possui uma definição fechada, podendo ser representada por um envolvido que aparece em vários IPLs, ou a mesma vítima em diversos Fatos. Os resultados para cada modelo são exibidos a seguir:

modelo	**nDCG**
distância de cosseno	0.5161
similaridade de jaccard	0.4880
wmd	0.4190
doc2vec	0.1828

Tabela 3.10: Cálculo do nDCG para as consultas em cada modelo, em relação aos casos relacionados

Os resultados acima mostram novamente um melhor desempenho para modelos sintáticos tradicionais (distância de cosseno e similaridade de jaccard), onde novamente o doc2vec fica como pior modelo. Por a métrica nDCG possuir uma escala que varia entre 0 e 1, os valores apresentados acima não parecem satisfatórios quando comparados aos resultados para o caso de uso anterior. Este resultado tem duas possíveis interpretações: Os modelos podem não ser bons o suficiente para a tarefa em questão, ou o *corpus* não possui um número considerável de IPLs relacionados aos inquéritos utilizados nas consultas.

A Tabela a seguir mostra resultados de consultas contendo IPLs avaliados como **relacionados** aos seus respectivos IPLs fornecidos como entrada nas consultas, e que foram apontados pelos três modelos com melhor desempenho. A coluna IPLConsulta refere-se ao identificador do documento utilizado como consulta. A coluna IPLsRelacionados mostra os documentos que foram avaliados como relacionados. A coluna Modelos mostra quais modelos detectaram a relação.

IPLConsulta	**IPLsRelacionados**	**Modelos**
303	302	cosseno, jaccard, wmd
2561	2581	cosseno, jaccard, wmd
2581	2561	cosseno, jaccard, wmd
3662	1526	cosseno, jaccard, wmd
3701	3721	cosseno, jaccard, wmd
8082	2841, 8084, 8101	cosseno, jaccard, wmd
8084	2841, 8082, 8101	cosseno, jaccard, wmd
8101	2841, 8082, 8084	cosseno, jaccard, wmd
15851	15852	cosseno, jaccard, wmd
15852	15851	cosseno, jaccard, wmd
17308	18964	cosseno, jaccard, wmd
18964	17308	cosseno, jaccard, wmd

Tabela 3.11: IPLs avaliados como relacionados em retornos de consultas por modelos

É possível observar que as consultas por IPLs citadas anteriormente (2561, 8082, 8084, 8101 e 17308) e que foram classificadas como muito semelhantes e duplicadas, também aparecem como relacionadas. Outro fato importante é que toda comparação entre IPLs avaliada relacionada, também é apontada como semelhante. Como o conceito de relação de IPLs pode ser mais abrangente e subjetivo, a avaliação nos mostra que a relação entre IPLs pode ser entendida como um evento que ocorre quando há semelhança entre IPLs.

3.6.3 Duplicação de IPLs

A Duplicação de IPLs visa detectar inquéritos instaurados para investigar um mesmo fato. Os resultados para cada modelo são exibidos a seguir:

modelo	**nDCG**
wmd	0.18
distância de cosseno	0.1786
similaridade de jaccard	0.16
doc2vec	0

Tabela 3.12: Cálculo do nDCG para as consultas em cada modelo, em relação aos casos duplicados

Como pode ser visto, o WMD possui o melhor desempenho, seguido da Distância de Cossenos e Similaridade de Jaccard, onde os três possuem valores bem próximos. O doc2vec não retorna duplicações, se assemelhando ao modelo aleatório, e para a detecção de IPLs duplicadas poderia ser descartado de acordo com este experimento.

De acordo com os valores obtidos pela métrica nDCG, é possível interpretar os resultados como abaixo do esperado, pois estes não possuem valores próximos a um. Entretanto, vale lembrar que cada modelo retorna o *top* 4 documentos para cada consulta, onde espera-se que se houver duplicação no *corpus*, esta deve estar contida neste resultado. Além disso, é sabido que a Duplicação é um fenômeno que acontece pouco no *corpus*, o que leva as consultas a naturalmente retornarem poucos casos de duplicação.

A Tabela a seguir mostra IPLs identificados como duplicados na avaliação:

IDConsulta	**IDsDuplicação**	**Modelos**
302	781	cosseno
303	302	cosseno, jaccard, wmd
2561	2581	cosseno, jaccard, wmd
3662	1526	cosseno, jaccard, wmd
8084	2841, 8082, 8101	cosseno, jaccard, wmd
8084	8083	cosseno, jaccard
10002	13403	cosseno
15464	16530	wmd
15851	15852	cosseno, jaccard, wmd
15852	15851	cosseno, jaccard, wmd
17308	18964	cosseno, jaccard, wmd
18964	17308	cosseno, jaccard, wmd

Tabela 3.13: Duplicações identificadas em retornos de consultas por modelos

A partir da Tabela acima, algumas considerações podem ser feitas:

Os par de IPLs (15851,15852) apresentou um comportamento esperado em 3 dos modelos (cosseno, jaccard e wmd), onde o 15851 retornou o 15852 e vice-versa. Isso demonstra uma certa consistência no resultado, uma vez que dado um documento duplicado como entrada, pode-se esperar o documento original no retorno. O mesmo ocorreu com o par (17308, 18964).

Outro caso interessante diz respeito aos pares (302, 781) e (303, 302). Neste exemplo, era de se esperar que a consulta com o documento 302 retornaria o documento 303. No entanto, apenas no modelo de Distância de Cossenos, um novo documento foi avaliado como duplicado (781). A imagem a seguir mostra um trecho da avaliação realizada pelos especialistas, onde é possível observar alguns dados dos IPLs 302 e 781. Os resumos de ambos os IPLs não parecem tratar do mesmo fato.

A partir da figura acima, é possível considerar que as avaliações realizadas pelos especi-

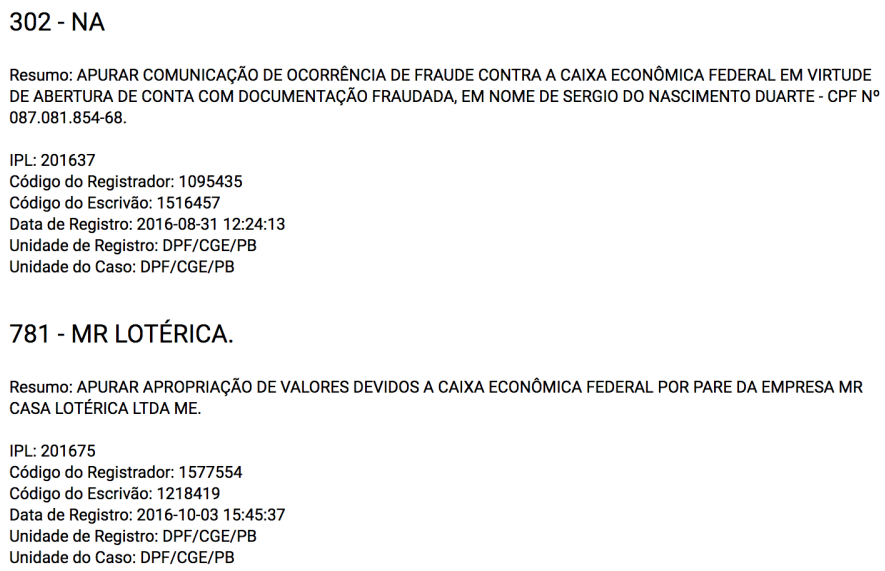


Figura 3.1: Trecho de avaliação onde são comparados os IPLs 302 e 781

alistas podem conter falhas como erros de interpretação, respostas inconsistentes ou falta de entendimento nos quesitos apresentados. Entretanto, vale ressaltar que o especialista refere-se ao indivíduo melhor qualificado para julgar o caso de uso em questão, representando a opinião mais confiável neste contexto.

A consulta referente ao IPL 8084 foi avaliada possuindo quatro IPLs duplicadas (2841, 8082, 8101 e 8083). Todos os 5 IPLs em questão referem-se ao mesmo envolvido, com os resumos indicando o mesmo tipo de Fato (”Apurar irregularidades na aplicação de recursos públicos federais, mais especificamente com relação ao Convênio...”), com diferenças em datas de instauração e números de Convênios. Apesar destes IPLs tratarem de um mesmo envolvido e com mesmo tipo de crime, os alvos do crime parecem ser diferentes, onde cada IPL trata de um convênio diferente. Nesse sentido, a avaliação destes IPLs como duplicados nos permite entender que esses eventos poderiam compor um mesmo inquérito, pois mesmo que as irregularidades ocorram em diferentes convênios, o envolvido demonstra estar agindo de forma sistemática sob a aplicação de recursos públicos federais.

3.7 Conclusão

No estudo comparativo buscamos responder se é possível recuperar inquéritos duplicados, semelhantes ou relacionados a partir de um dado inquérito. A relevância deste estudo está em buscar soluções para a duplicação de inquéritos, onde esta consiste em uma violação ao princípio da dupla incriminação, além de configurar desperdício de recursos públicos. No caso de inquéritos semelhantes, o estudo possibilita a recuperação destes visando o auxílio no treinamento de novos delegados e escrivães, uma vez que itens semelhantes podem sugerir ações a serem executadas na condução de um novo IPL. Em relação a recuperação inquéritos relacionados, que refere-se a detecção de IPLs que possuem algum tipo de ligação, deve auxiliar na solução destes, haja vista que podem haver informações complementares entre os inquéritos, ou que podem configurar em um encadeamento de crimes investigados por diferentes núcleos de investigação.

Para realização do estudo, foram utilizados dados de inquéritos não-sigilosos investigados pela Polícia Federal, além de Peças produzidas durante as investigações, onde estas encontram-se anexadas aos seus respectivos inquéritos. Houve uma etapa de pré-processamento, onde os inquéritos foram representados por *tokens* derivados de cada documento. Dentre estes *tokens* estão unigramas, bigramas, trigramas e expressões regulares que extraíram CPFs, CNPJs, nomes próprios, datas, números de telefones, etc. Com os dados pré-processados, foram construídos 5 modelos para serem avaliados: distância de cosseno, similaridade de jaccard, Doc2Vec, WMD e um modelo aleatório. Em seguida, foram escolhidos 55 inquéritos para serem utilizados como consultas aos modelos. Todos os modelos retornaram seu top4 inquéritos mais similares com a entrada da consulta.

Na etapa de avaliação, foram construídos formulários estruturados onde estes comparavam pares de inquéritos e especialistas responderam os quesitos avaliando se os inquéritos eram duplicados, semelhantes ou relacionados. Um exemplo de comparação encontra-se no Apêndice A. Após as respostas coletadas, foi utilizada a métrica de avaliação NDCG para medir assertividade dos modelos comparados.

Os resultados foram divididos por casos de uso, onde a recuperação de inquéritos semelhantes obteve os melhores resultados, destancando-se os modelos de similaridade de jaccard e distância de cosseno (0.8812 e 0.8371, respectivamente). Acredita-se que este resultado se

deve ao fato de que o *corpus* possui muitas IPLs que tratam de crimes de mesma natureza, além de possuir um vocabulário bem específico do contexto.

Em relação a recuperação de inquéritos relacionados, os modelos obtiveram um resultado mediano, com destaque novamente para distância de cosseno e similaridade de jaccard (0.5161 e 0.4880 respectivamente). Este resultado remete aos modelos não serem bons o suficiente para a tarefa em questão, ou que o *corpus* não possui um número considerável de inquéritos relacionados aos utilizados nas consultas.

Já para a recuperação de inquéritos duplicados, os modelos tiveram os piores resultados de acordo com a métrica NDCG. Entretanto, uma possível interpretação destes resultados deve considerar que o problema de duplicação é algo que, em geral, acontece pouco em qualquer contexto, o que não é diferente no contexto aqui estudado. Outro ponto interessante de ressaltar é que todos os modelos retornam seus top4 inquéritos mais similares a um dado inquérito passado como entrada, mesmo que o maior *score* de similaridade encontrado seja baixo. Sendo assim, seria necessário um estudo mais aprofundado dos *scores* de similaridade para buscar um limiar que determinasse a inclusão de um inquérito como resposta a uma consulta.

Capítulo 4

Discussão

O objetivo deste trabalho foi avaliar métodos sintáticos e semânticos para representação de documentos em um contexto de investigação policial e, a partir de tais representações, recuperar documentos visando cobrir casos de uso como duplicação, semelhança e relação entre inquéritos. Foram avaliados modelos clássicos como distância de cosseno e similaridade de jaccard, além de modelos construídos a partir de redes neurais, como o doc2vec e o word mover's distance. Os modelos mais clássicos partem de uma abordagem puramente sintática, onde a representação vetorial dos documentos se aproxima a medida que cresce o número de tokens comuns aos documentos. Já as abordagens semânticas consideram o contexto em que os *tokens* ocorrem para aproximar ou distanciar ainda mais a representação vetorial dos documentos.

O estudo mostrou que os modelos avaliados apresentam um bom desempenho para recuperar inquéritos semelhantes, onde a similaridade de jaccard e distância de cosseno apresentaram os melhores resultados de acordo com a métrica NDCG (0.8812 e 0.8371, respectivamente). O doc2vec apresenta resultados abaixo do esperado, corroborando com a ideia de que bases de dados pequenas podem não possuir informação suficiente para uma rede neural aprender padrões desconhecidos.

A respeito do bom desempenho para a recuperação de inquéritos semelhantes com modelos baseados em similaridade sintática, levantamos a hipótese de que a escrita dos documentos possui uma peculiaridade de seu contexto, onde a partir de uma análise manual em amostras de descrições dos inquéritos, percebemos um padrão no vocabulário utilizado pela PF na escrita destes documentos. Além disso, um fator que pode ser determinante para o

bom desempenho em recuperar inquéritos semelhantes é a pouca variedade de tipos de crime registrados na base de dados utilizada no estudo, onde isso aumentaria as chances dos modelos de retornar inquéritos semelhantes. A partir da base de inquéritos utilizada, observamos uma coluna denominada CDORGAOLESADO, onde esta representa o código de qual órgão público foi vítima do crime registrado em um dado inquérito. Vale salientar que a Polícia Federal atua no âmbito de interesses da União e tem como objetivo a apuração de crimes e infrações penais cometidas contra a União e também suas empresas públicas.

A Tabela a seguir mostra os 5 órgãos mais lesados registrados na base de inquéritos utilizada:

CDORGAOLESADO	**count**
207	313
2	303
152	240
24	145
41	102

Tabela 4.1: Número de inquéritos em que os órgãos lesados estão registrados

Dos 1541 inquéritos utilizados neste trabalho, 71% possuem um dos 5 órgãos listados acima registrados como lesados, onde este pode ser um fator que facilitaria a busca dos modelos por inquéritos semelhantes, haja vista que crimes que ocorrem com frequência a um órgão público podem possuir um maior grau de semelhança na visão dos especialistas. Como exemplo, podemos citar crimes contra a Previdência Social por possível recebimento de benefício pós óbito.

Em relação ao mal desempenho na deteção de inquéritos duplicados, algumas considerações devem ser feitas. O problema referente a duplicação de documentos é conhecido por ser algo que acontece pouco nos dados, independente de contexto. No contexto aqui estudado, a base de dados é definida em um problema não supervisionado, onde não é possível saber a priori quais documentos possuem duplicações. Entretanto, de acordo com os especialistas na área e pelas análises manuais realizadas em amostras dos dados, foi possível

observar que o número de duplicações encontradas na base de dados é muito baixo. Outro ponto que contribui para estes resultados na recuperação de duplicações diz respeito a todos os modelos retornarem os seus 4 documentos mais similares a um dado documento passado como entrada. Sendo assim, mesmo que a similaridade entre a entrada e os retornos sejam baixas, estes documentos ainda serão retornados. Como solução para corrigir este problema, poderia ser feito um estudo sobre a definição de um limiar de similaridade, onde os modelos só retornariam inquéritos considerados duplicados quando o valor de similaridade atingisse ou ultrapassasse tal limiar.

Para a detecção de inquéritos relacionados, ainda existe um nível de subjetividade onde o avaliador especialista pode considerar inquéritos semelhantes como relacionados, sendo este um cenário que necessita de uma definição mais clara e fechada sobre a relação entre inquéritos.

Capítulo 5

Trabalhos Futuros

A partir dos resultados obtidos com os modelos, é possível pensar em formas de aplicação destes para auxiliar DPFs e EPFs em suas atividades relacionadas a investigações de IPLs. Dado que a ferramenta atual para a condução e gerenciamento de inquéritos é o ePol, e que esta possui uma série de módulos e microserviços implementados para a realização de tarefas específicas dentro do sistema, uma forma de aplicação dos artefatos propostos neste trabalho seria a criação de um módulo ou microserviço no sistema que atuaria retornando IPLs para uma dada consulta. Esta consulta seria feita a partir de um IPL dado como entrada e o retorno seria uma lista de IPLs ordenados pela sua similaridade, em ordem decrescente e de acordo com o caso de uso escolhido.

Os cenários de utilização do novo recurso proposto para o sistema estariam ligados diretamente com os casos de uso avaliados neste trabalho. Para a detecção de IPLs duplicados, o usuário do sistema poderia realizar a consulta por duplicação para um dado IPL e, ao encontrar uma possível duplicação, poderia executar a ação de apensamento, anexando os IPLs um ao outro. A ideia de anexar os IPLs, ao contrário da exclusão de um deles, é válida pois como são IPLS que podem ter sido instauradas por usuários diferentes, elas podem ter informações complementares para a investigação do fato.

Outro cenário possível diz respeito ao caso de uso sobre inquéritos relacionados. O investigador poderia buscar por IPLs relacionados ao IPL atual, visando descobrir a relação entre os crimes, que podem possuir o mesmo envolvido ou a mesma vítima, por exemplo. Neste cenário, inquéritos relacionados também poderiam possuir informações complementares e que dariam mais suporte ao investigador para solucionar as investigações.

Por fim, um cenário possível é de novos DPFs iniciando seus trabalhos na PF e que, por falta de experiência, podem haver dúvidas sobre como conduzir novos inquéritos. Sendo assim, o sistema poderia sugerir IPLs considerados semelhantes ao IPL atual, onde o DPF teria a possibilidade de entender como é feita a condução de um IPL, quais ações são tomadas em uma situação parecida, quanto tempo leva para solucionar um IPL daquele tipo, etc.

Com a incorporação de um módulo proposto nos moldes mencionados acima, cria-se a possibilidade do investigador assumir o papel de avaliador enquanto realiza seu trabalho, podendo assim dar *feedbacks* constantes sobre as consultas retornadas em cada cenário descrito, e assim tornar mais precisa e confiável a assertividade dos modelos utilizados. Além disso, novos *feedbacks* sobre novas consultas contribuem para o aumento de rótulos na base de inquéritos, tornando possível a utilização de abordagens supervisionadas em aprendizagem de máquina para criação de novos modelos e assim dar continuidade a pesquisa com trabalhos futuros.

Outro possível trabalho futuro, diz respeito ao estudo de limiares de similaridade para os modelos, onde o modelo só retornaria os inquéritos com *scores* de similaridade iguais ou maiores que o limiar definido. Entretanto, para este estudo seria necessário um maior número de avaliações e possivelmente avaliadores, visando construir uma base de dados com mais rótulos e assim realizar um processo iterativo envolvendo criação de novos modelos e testes com diferentes limiares de similaridade.

Capítulo 6

Trabalhos Relacionados

No contexto de perícia digital, Mesquita et al. [Mesquita et al.,] propõem uma arquitetura que utiliza de conhecimentos adquiridos em perícias similares anteriores para a solução de perícias atuais, onde foram apresentados bons resultados, sugerindo procedimentos periciais adicionais em cerca de 76% parte dos casos.

Chandra et al. [Chandra et al., 2008] propõem um modelo que visa encontrar tendências de crime semelhantes entre várias séries de crime de diferentes locais e, posteriormente, usar essa informação para futuras previsões de novos crimes. A análise mostra que a técnica proposta geralmente supera as técnicas existentes em agrupamento em dados de séries temporais multivariadas.

Em um contexto de engenharia de *software*, Ye et al. [Ye et al., 2016] propõem um modelo baseado em *word embeddings* para estimar a similaridade semântica entre linguagem natural (expressa em texto) e código, visando facilitar tarefas de busca em engenharia de *software*, a exemplo de localização de *bugs* e *features*, respostas de questões em fóruns e comunidades, ou a comunicação entre uma equipe técnica e *stakeholders* em um projeto de *software*. Avaliações empíricas mostraram que os termos aprendidos do espaço vetorial levam a melhorar - em uma tarefa de localização de *bugs* e uma tarefa de ligar documentos de uma API a perguntas de programação.

Kenter & De Rijke [Kenter and De Rijke, 2015] propõem um modelo de medir a similaridade semântica em textos curtos combinando *insights* de métodos baseados em fontes externas de conhecimento semântico e *word embeddings* como forma de representar a informação. Através de um conjunto de avaliação acessível ao público comumente usado para a tarefa de

similaridade semântica, o estudo mostra que o método proposto supera métodos base que funcionam nas mesmas condições.

Duplessis et al. [Duplessis et al., 2017] propõem um processo *corpus-based*, não-supervisionado e independente de linguagem que visa representar e indexar enunciados de diálogo de domínio aberto para um sistema de recuperação de documentos que possa ser incorporado em um agente de conversação. O estudo indica que o modelo proposto executa objetivamente bem se comparado a outros modelos de recuperação em uma tarefa de seleção de exemplos de diálogo derivados de um grande corpus de diálogos escritos. Dentre os modelos utilizados para comparação, estão o TF-IDF, doc2vec, um modelo baseado em trigramas e um aleatório.

Wu et al. [Wu and Wang, 2017] propõem um novo método para identificar os principais tópicos em grandes quantidades de artigos. A abordagem se concentra em construir vetores de texto e melhorar a eficiência e precisão do agrupamento de documentos com base no modelo Word2Vec, combinando o coeficiente de similaridade de Jaccard e a frequência de dimensão inversa para calcular o grau de importância entre cada dimensão no vetor de texto e o documento correspondente. Na avaliação o método proposto é comparado a outro método que é eficiente na classificação de arquivos e que utiliza os valores do TF-IDF para ponderar os vetores de palavras criados a partir do Word2Vec. Os resultados mostram uma acurácia superior do método proposto (0.75) ao método base de comparação (0.55).

Ben et al. [Ben-Lhachemi and Nfaoui, 2018] apresentam uma abordagem para recomendação de *hashtags* no Twitter com base em embeddings de *tweets* e agrupamento destes utilizando o algoritmo k-means, onde as *hashtags* recomendadas a um dado *tweet* serão provenientes de *tweets* contidos nos *clusters* mais similares à entrada. Para base de comparação, o estudo utiliza de três formas para gerar os *embeddings*: média de vetores word2vec, média de vetores word2vec utilizando o tf-idf como peso e vetores doc2vec. Como métricas de avaliação, são utilizadas precisão, revocação e F1. Os resultados mostram que as técnicas utilizadas para calcular os *embeddings* de *tweets* influenciam no conjunto final das *hashtags* recomendadas.

Na temática sobre Desambiguidade de Entidades, Zhang et al. [Zhang et al., 2018] propõem um método de desambiguação baseado na semelhança semântica de palavras ambíguas. Primeiramente, de acordo com as entidades no contexto da palavra ambígua,

um classificador é construído para prever a classificação da palavra ambígua e uma lista de entidades candidatas é obtida de acordo com a classificação. Em seguida, os triplos contextuais do *Resource Description Framework* (RDF) relacionados a palavras ambíguas num Grafo de Conhecimento são mapeados para o mesmo espaço vetorial com os triplos RDF relacionados às entidades na tabela de entidades candidatas. Finalmente, a similaridade semântica é obtida de acordo com a similaridade do cosseno, e a similaridade top-k é selecionada. A validade desse método para desambiguação de entidades é avaliada por um conjuntos de dados presente na literatura em [Berant et al., 2013], conhecido como *The Web Queries dataset*. Os resultados mostraram que o método proposto foi superior à linha de base de similaridade de contexto existente. Além disso, o método aprimorado é adequado para a maior parte do gráfico de conhecimento.

Na área de recomendação de itens a usuários em sistemas de *e-commerce*, Phi et al. [Phi et al., 2016] propõem um sistema de recomendação focando em recomendadores item a item e usuário a item, que são as duas funções mais usadas nos serviços on-line para apresentar itens relevantes a um item ou a um usuário em particular. Este trabalho aplica o método de representação distribuída em dados de *e-commerce* para desenvolver dois tipos de sistemas de recomendação, tratando os itens como palavras e as sessões dos usuários como sentenças para, em seguida, treinar os modelos word2vec e doc2vec com base nesses itens e nas informações do usuário. Os vetores de itens resultantes do modelo Word2vec são utilizados para calcular a semelhança de cosseno entre itens e encontrar os itens semelhantes em um item. Da mesma forma, o modelo doc2vec ajuda os usuários a encontrar itens relevantes que possam interessá-los usando similaridade entre itens e vetores. Por fim, os vetores de itens de ambos os modelos de incorporação são utilizados para criar uma recomendação adicional de usuário para item. Os experimentos mostraram que o melhor sistema alcançou uma taxa de acerto de 0.24 por recomendar itens aos usuários nos dados de teste, o que superou as abordagens convencionais em uma extensão significativa.

Li et al. [Li et al., 2019] investigam um método baseado em analogia de Inspeção de Casos Legais em dados de processos legais jurídicos. No trabalho são utilizados os vetores de documentos gerados a partir do doc2vec (recurso de caso baseado em semântica, SCF) e o recurso definido pelo modelo de julgamento de caso (recurso de caso baseado em modelo, MCF) como duas maneiras de encontrar casos semelhantes. Os métodos de medição de

similaridade entre dois casos e o desvio de julgamento de caso também são definidos, onde foram selecionados a semelhança de cosseno para calcular a semelhança entre os casos com SCF. Resultados experimentais em um conjunto de dados do mundo real mostraram a eficácia do nosso método, onde a taxa de recuperação de casos irracionais ao usar o MCF é maior que a do SCF.

Yang et al. [Yang et al., 2019] avaliam diferentes métodos de vetorização de documentos, incluindo TFIDF, LDA, LSA, word2Vec e doc2Vec, visando otimizar o processo de revisão sistemática dos resultados em pesquisas médicas publicados na Wikipédia. Para cada documento, sua similaridade com cada outro documento no conjunto foi calculada usando métricas de similaridade vetorial estabelecidas, como similaridade de cosseno e divergência de KL. Os modelos foram avaliados comparando os resultados com dois padrões: (1) Revisões Cochrane atualmente citadas em artigos da Wikipedia e (2) um conjunto de dados fornecido por um especialista médico que indica quais Revisões Cochrane poderiam ser consideradas para artigos específicos da Wikipédia. O modelo com o melhor desempenho utilizou a representação de documentos TFIDF e a semelhança de cossenos.

Zhou et al. [Zhou et al., 2019] propõe um método de medida de similaridade de texto denominado descentralização por distância do vetor de palavras (WVDD), que pode lidar com relações semânticas complexas, incluindo componentes de sentenças, ordem das palavras e pesos para o idioma chinês. Em seguida, a análise de agrupamento é realizada para os resultados de similaridade obtidos. O algoritmo K-means baseado na arquitetura Spark para computação paralela é adotado para acelerar a velocidade de processamento do *cluster*. Na verificação experimental, os conjuntos de testes são um número significativo de comentários de clientes publicados no site da Jingdong, que é um abrangente *shopping online*. A medida F é usada para avaliar a precisão dos resultados obtidos pelo método proposto. A superioridade do método proposto é verificada e comparada com o modelo de vetor de sentenças doc2vec e o modelo de *bag of words*. O método proposto pode ser aplicado para analisar o idiomas, como comentários dos clientes *on-line* e dados de bate-papo na *web*.

Os trabalhos citados representam grande contribuição para o estado da arte nas áreas de recuperação de informação e semântica distributiva, auxiliando a recuperação de itens baseado em similaridade e propondo maneiras de medir a similaridade semântica entre textos, além de realizar comparativos entre as técnicas. Entretanto, não foram encontradas

evidências do uso de similaridade em seu estado da arte aplicadas ao contexto de recuperação de inquéritos policiais ou de trabalhos que rejeitem seu uso.

Capítulo 7

Conclusão

Neste capítulo estão as conclusões finais do trabalho, com algumas observações e lições aprendidas.

Este trabalho avaliou técnicas clássicas e do estado da arte em Recuperação da Informação em um contexto de investigação policial, onde estas foram representadas a partir de modelos de RI que retornaram inquéritos policiais para um dado inquérito como entrada. A base de dados utilizada foi fornecida pela Polícia Federal, onde esta foi coletada da plataforma ePol, sistema este utilizado pela PF na condução de inquéritos policiais. Os modelos clássicos utilizados foram a distância de cosseno e similaridade de jaccard, assim como modelos baseados em redes neurais Doc2Vec e WMD. Além destes modelos, um modelo aleatório também foi avaliado sendo considerado como *baseline*, onde para haver alguma relevância nos demais modelos, estes teriam que ter um resultado melhor do que o *baseline*.

De acordo com a métrica utilizada (NDCG), os modelos avaliados apresentaram um bom desempenho para detectar Inquéritos Semelhantes, onde a similaridade de Jaccard e distância de Cosseno apresentam os melhores resultados (0.8812 e 0.8371 respectivamente). O doc2vec apresentou resultados abaixo do esperado (0.6743), corroborando com a ideia de que bases de dados pequenas podem não possuir informação suficiente para uma rede neural aprender o comportamento esperado. O modelo aleatório obteve um *score* igual a zero para todos os casos de uso.

Em relação aos inquéritos semelhantes, foi discutido no capítulo anterior como este cenário pode ter atingido um bom desempenho, onde a pouca variedade nos tipos de crime podem afetar positivamente nos resultados, além do próprio vocabulário ser bem específico

do contexto.

Em relação a recuperação de inquéritos relacionados, os modelos obtiveram um resultado mediano, com destaque novamente para distância de cosseno e similaridade de jaccard (0.5161 e 0.4880 respectivamente). Este resultado remete aos modelos não serem bons o suficiente para a tarefa em questão, ou que o *corpus* não possui um número considerável de inquéritos relacionados aos utilizados nas consultas.

Já para o mal desempenho na recuperação de inquéritos duplicados, a possível causa deve-se a combinação de fatores como o evento de duplicação ser algo raro nos dados e toda consulta retornar exatos 4 documentos, independente do valor da similaridade com a entrada ser alto ou baixo. Sendo assim, seria necessário um estudo mais aprofundado dos *scores* de similaridade para buscar um limiar que determinasse a inclusão de um inquérito como resposta a uma consulta.

Outro ponto importante diz respeito ao trabalho de avaliação com especialistas, pois por se tratar de uma atividade manual e que depende de indivíduos, existe a possibilidade de ocorrer falhas na avaliação, como está exposto no Capítulo 4 referente ao Estudo Comparativo, subseção 4.6.3 e Figura 4.1, onde é possível ter havido algum erro de interpretação, resposta inconsistente ou falta de entendimento nos quesitos apresentados.

Por fim, concluímos este trabalho ressaltando a importância deste para o contexto de investigação policial, pois o auxílio na busca por inquéritos agiliza o trabalho dos investigadores de um modo geral. Além disso, este estudo acresce a área de RI com um experimento comparativo utilizando dados reais, não-supervisionados, realizando uma avaliação com especialistas no contexto e registrando observações importantes para serem levadas em consideração nos estudos futuros da área.

Bibliografia

[kag, 2010] (2010). Kaggle - online community of data scientists and machine learning practitioners. http://www.kaggle.com.

[epo, 2013] (2013). Sistema desenvolvido na ufcg vai interligar informações da pf no país. http://g1.globo.com/pb/paraiba/noticia/2013/11/sistema-desenvolvido-na-ufcg-vai-interligar-informacoes-da-pf-no-pais.html. Accessed: 2017-11-27.

[dup, 2016] (2016). Breves considerações acerca da duplicidade de procedimentos investigatórios criminais idênticos conduzidos pela polícia e ministério público. https://jus.com.br/artigos/51038/breves-consideracoes-acerca-da-duplicidade-de-procedimentos-investigatorios-criminais-identicos-conduzidos-pela-policia-e-ministerio-publico. Accessed: 2017-11-27.

[ALENCAR and Távora, 2016] ALENCAR, R. R. and Távora, N. (2016). Curso de direito processual penal.

[Andrews et al., 2009] Andrews, M., Vigliocco, G., and Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3):463.

[Baroni et al., 2007] Baroni, M., Lenci, A., and Onnis, L. (2007). Isa meets lara: An incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 49–56. Association for Computational Linguistics.

[Ben-Lhachemi and Nfaoui, 2018] Ben-Lhachemi, N. and Nfaoui, E. H. (2018). Hashtag recommendation using word sequences' embeddings. In *International Conference on Big Data, Cloud and Applications*, pages 131–143. Springer.

[Berant et al., 2013] Berant, J., Chou, A., Frostig, R., and Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544.

[Büttcher et al., 2016] Büttcher, S., Clarke, C. L., and Cormack, G. V. (2016). *Information retrieval: Implementing and evaluating search engines*. Mit Press.

[Chandra et al., 2008] Chandra, B., Gupta, M., and Gupta, M. (2008). A multivariate time series clustering approach for crime trends prediction. In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, pages 892–896. IEEE.

[Duplessis et al., 2017] Duplessis, G. D., Charras, F., Letard, V., Ligozat, A.-L., and Rosset, S. (2017). Utterance retrieval based on recurrent surface text patterns. In *European Conference on Information Retrieval*, pages 199–211. Springer.

[Ferrero et al., 2017] Ferrero, J., Agnes, F., Besacier, L., and Schwab, D. (2017). Using word embedding for cross-language plagiarism detection. *arXiv preprint arXiv:1702.03082*.

[FONTENELE, 2015] FONTENELE, A. K. P. C. e. B. C. (2015). *Manual do Delegado de Polícia Federal*. JUSPODIVM.

[Järvelin and Kekäläinen, 2002] Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

[Kenter and De Rijke, 2015] Kenter, T. and De Rijke, M. (2015). Short text similarity with word embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1411–1420. ACM.

[Kusner et al., 2015] Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

[Lenci, 2008] Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

[Li et al., 2019] Li, S., Guo, B., Cai, Y., Ye, L., Zhang, H., and Fang, B. (2019). Legal case inspection: An analogy-based approach to judgment evaluation. In *International Conference on Artificial Intelligence and Security*, pages 148–158. Springer.

[Manning et al., 2009] Manning, C. D., Raghavan, P., and Schütze, H. (2009). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[Mesquita et al.,] Mesquita, F. I., Hoelz, B. W. P., and Ralha, C. G. Raciocínio baseado em casos aplicado em análise live.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Phi et al., 2016] Phi, V.-T., Chen, L., and Hirate, Y. (2016). Distributed representation based recommender systems in e-commerce. In *DEIM Forum*.

[Singhal et al., 2001] Singhal, A. et al. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

[Tan et al., 2005] Tan, P., Steinbach, M., and Kumar, V. (2005). Introduction to data mining. ed. *Addison-Wesley Longman Publishing Co., Inc.*

[Wikipedia, 2017a] Wikipedia (2017a). Similarity search — wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Similarity$_s earcholdid = 800167647. [Online; accessed 28 - November - 2017]$.

[Wikipedia, 2017b] Wikipedia (2017b). Word embedding — wikipedia, the free encyclopedia. [Online; accessed 28-November-2017].

[Wu and Wang, 2017] Wu, C. and Wang, B. (2017). Extracting topics based on word2vec and improved jaccard similarity coefficient. In *2017 IEEE Second International Conference on Data Science in Cyberspace (DSC)*, pages 389–397. IEEE.

[Yang et al., 2019] Yang, J., Ward, J., Gharavi, E., Dawson, J., and Alvarado, R. (2019). Bi-directional relevance matching between medical corpora. In *2019 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–6. IEEE.

[Ye et al., 2016] Ye, X., Shen, H., Ma, X., Bunescu, R., and Liu, C. (2016). From word embeddings to document similarities for improved information retrieval in software engineering. In *Proceedings of the 38th International Conference on Software Engineering*, pages 404–415. ACM.

[Zhang et al., 2018] Zhang, K., Zhu, Y., Gao, W., Xing, Y., and Zhou, J. (2018). An approach for named entity disambiguation with knowledge graph. In *2018 International Conference on Audio, Language and Image Processing (ICALIP)*, pages 138–143. IEEE.

[Zhou et al., 2019] Zhou, S., Xu, X., Liu, Y., Chang, R., and Xiao, Y. (2019). Text similarity measurement of semantic cognition based on word vector distance decentralization with clustering analysis. *IEEE Access*, 7:107247–107258.

Apêndice A

Formulário de Avaliação entre IPLs

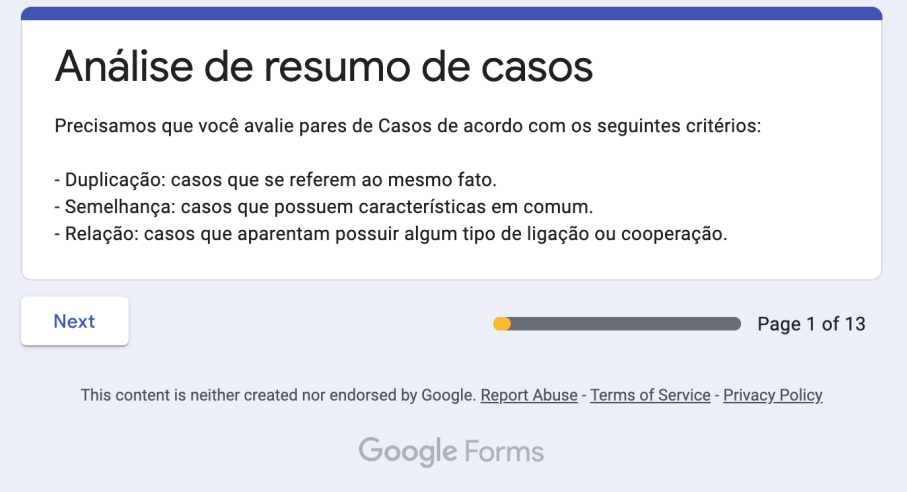
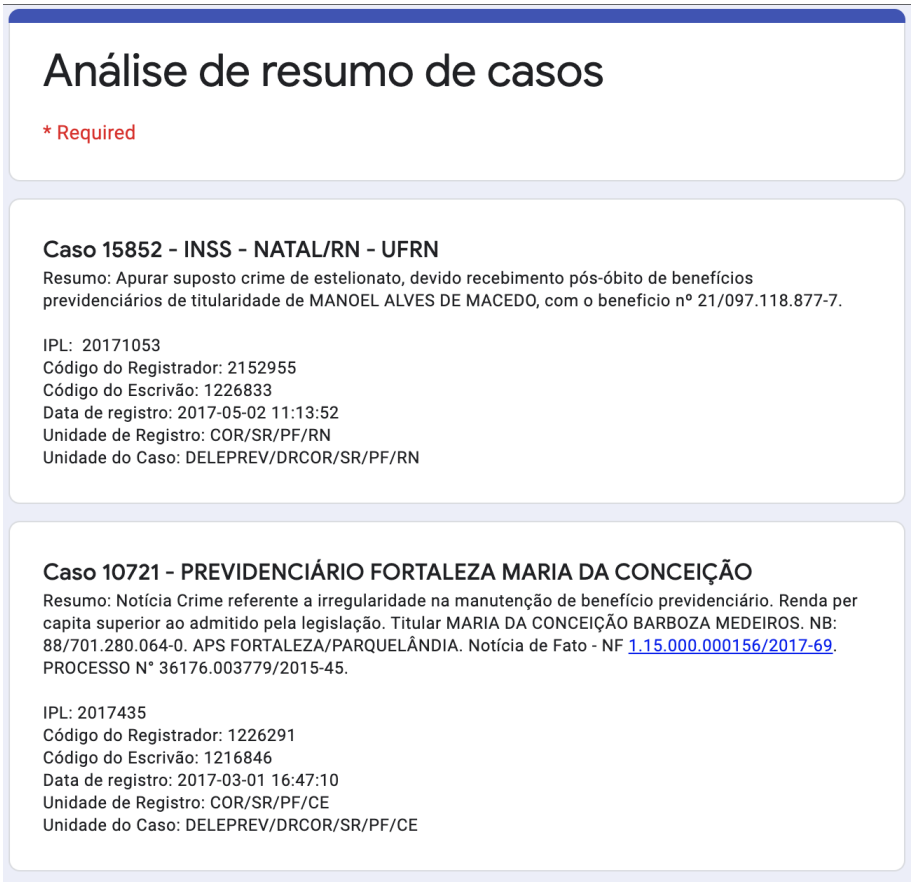


Figura A.1: Formulário de Avaliação entre IPL - Parte 1



Análise de resumo de casos

* Required

Caso 15852 - INSS - NATAL/RN - UFRN

Resumo: Apurar suposto crime de estelionato, devido recebimento pós-óbito de benefícios previdenciários de titularidade de MANOEL ALVES DE MACEDO, com o beneficio nº 21/097.118.877-7.

IPL: 20171053
Código do Registrador: 2152955
Código do Escrivão: 1226833
Data de registro: 2017-05-02 11:13:52
Unidade de Registro: COR/SR/PF/RN
Unidade do Caso: DELEPREV/DRCOR/SR/PF/RN

Caso 10721 - PREVIDENCIÁRIO FORTALEZA MARIA DA CONCEIÇÃO

Resumo: Notícia Crime referente a irregularidade na manutenção de benefício previdenciário. Renda per capita superior ao admitido pela legislação. Titular MARIA DA CONCEIÇÃO BARBOZA MEDEIROS. NB: 88/701.280.064-0. APS FORTALEZA/PARQUELÂNDIA. Notícia de Fato - NF 1.15.000.000156/2017-69. PROCESSO N° 36176.003779/2015-45.

IPL: 2017435
Código do Registrador: 1226291
Código do Escrivão: 1216846
Data de registro: 2017-03-01 16:47:10
Unidade de Registro: COR/SR/PF/CE
Unidade do Caso: DELEPREV/DRCOR/SR/PF/CE

Figura A.2: Formulário de Avaliação entre IPL - Parte 2

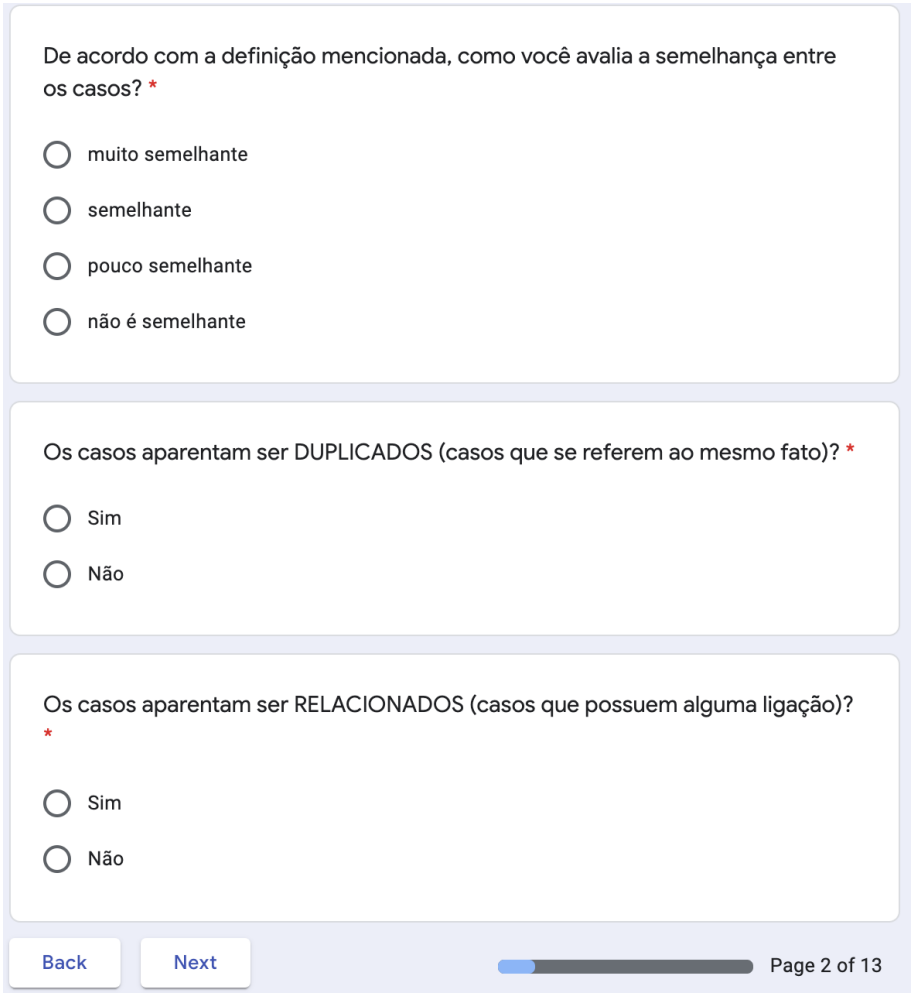


Figura A.3: Formulário de Avaliação entre IPL - Parte 3

Apêndice B

Representação de Dataset das respostas coletadas

formid	queryid	resp	semelhante	duplicado	relacionado	modelo	ranking
1	15852	10721	2	0	0	random	1
1	15852	2461	0	0	0	random	2
1	15852	63	0	0	0	random	3
1	15852	1863	1	0	0	random	4
1	15852	15845	3	0	1	doc2vec	1
1	15852	15590	3	0	1	doc2vec	2
1	15852	15846	3	0	1	doc2vec	3
1	15852	15848	3	0	1	doc2vec	4
1	15852	15851	3	1	1	jaccard	1
1	15852	15854	3	0	1	jaccard	2
1	15852	15826	3	0	1	jaccard	3
1	15852	15864	3	0	1	jaccard	4

Tabela B.1: Exemplo de respostas de um formulário para uma dada consulta. Os modelos retornam uma lista de códigos de inquéritos para um código de inquério dado como entrada - parte 1

formid	queryid	resp	semelhante	duplicado	relacionado	modelo	ranking
1	15852	15851	3	1	1	wmd	1
1	15852	15826	3	0	1	wmd	2
1	15852	5142	3	0	0	wmd	3
1	15852	1802	2	0	0	wmd	4
1	15852	15851	3	1	1	cosine	1
1	15852	15846	3	0	1	cosine	2
1	15852	15854	3	0	1	cosine	3
1	15852	15847	3	0	0	cosine	4

Tabela B.2: Exemplo de respostas de um formulário para uma dada consulta. Os modelos retornam uma lista de códigos de inquéritos para um código de inquério dado como entrada - parte 2

Descrição das colunas da Tabela B.2:

- formid - identificador do formulário.
- queryid - identificador de inquérito utilizado para consulta.
- resp - identificador de inquérito retornado numa lista de inquéritos para uma dada consulta.
- semelhante - valor que representa nível de semelhança entre os itens em query e resp. Este é um valor discreto que varia entre 0 e 3, onde 0 indica que os inquéritos não são semelhantes e 3 são muito semelhantes.
- duplicado - indica se os inquéritos são duplicados ou não (0 não é duplicado, 1 é duplicado).
- relacionado - indica se os inquéritos são relacionados ou não (0 não é relacionado, 1 é relacionado).

- modelo - indica qual modelo retornou o inquérito da coluna ”resp”para a consulta referente ao inquérito na coluna ”query”.

- ranking - ordem em que o inquérito da coluna ”resp”aparece na lista de inquéritos retornados para a consulta ao inquérito contido na coluna ”query”.

modelo	**queryid**	**semelhante**	**duplicado**	**relacionado**
cosine	15852	[3, 3, 3, 3]	[1, 0, 0, 0]	[1, 1, 1, 0]
cosine	16530	[3, 3, 3, 3]	[0, 0, 0, 0]	[1, 1, 1, 1]
...	...	...	...	...
doc2vec	15852	[3, 3, 3, 3]	[0, 0, 0, 0]	[1, 1, 1, 1]
doc2vec	16530	[3, 0, 0, 3]	[0, 0, 0, 0]	[1, 0, 0, 1]
...	...	...	...	...
jaccard	15852	[3, 3, 3, 3]	[1, 0, 0, 0]	[1, 1, 1, 1]
jaccard	16530	[3, 3, 3, 3]	[0, 0, 0, 0]	[1, 1, 1, 1]
...	...	...	...	...
random	15852	[2, 0, 0, 1]	[0, 0, 0, 0]	[0, 0, 0, 0]
random	16530	[0, 0, 0, 0]	[0, 0, 0, 0]	[0, 0, 0, 0]
...	...	...	...	...
wmd	15852	[3, 3, 3, 2]	[1, 0, 0, 0]	[1, 1, 0, 0]
wmd	16530	[3, 0, 0, 0]	[0, 0, 0, 0]	[1, 0, 0, 0]
...	...	...	...	...

Tabela B.3: Exemplo de respostas dos formulários para as consultas realizadas nos modelos.

Descrição das colunas da Tabela B.3:

- modelo - indica a qual modelo pertencem a as respostas para a consulta realizada com o identificador em ”queryid”.
- queryid - identificador de inquérito utilizado para consulta.
- semelhante - lista de valores que representam o nível de semelhança entre os itens comparados em uma consulta.
- duplicado - lista de valores que representam se os inquéritos são duplicados ou não (0 não é duplicado, 1 é duplicado).
- relacionado - lista de valores que representam se os inquéritos são relacionados ou não (0 não é relacionado, 1 é relacionado).

Apêndice C

Diagramas sobre Pré-Processamento dos Dados

Com o objetivo de complementar a explicação das etapas referentes aos Dados e seu respectivo Pré-processamento descritos no Capítulo 4, a figura descreve o processo de representação de um documento no contexto trabalhado.

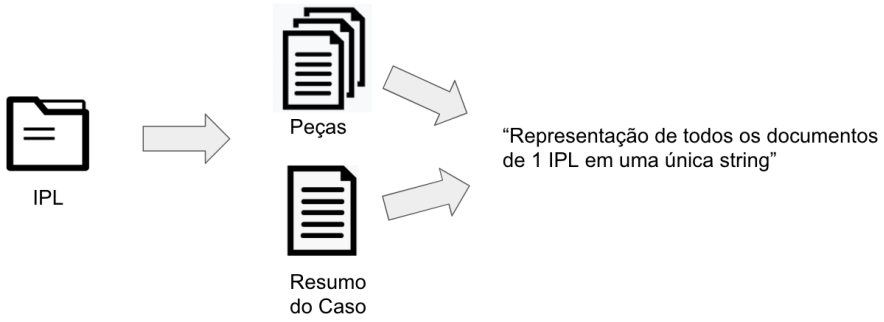


Figura C.1: Concatenação de todos os documentos contidos em uma IPL por uma única string a ser processada