ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS APLICADA A POLÍTICAS PÚBLICAS

Relatório de Pesquisa: Aplicação de técnicas de Recuperação de Informações na seleção de editais de licitações para auditoria.

Discente: Ronaldo Gonçalves Borges

Orientador: Gustavo Cordeiro Galvão Van Erven

Brasília-DF, setembro de 2022

Sumário

1. Introdução: Problema abordado, justificativa e objetivos

A Controladoria-Geral da União (CGU), composta pela Secretaria de Transparência e Prevenção da Corrupção (STPC), Secretaria Federal de Controle Interno (SFC), Corregedoria-Geral da União (CRG), Secretaria de Combate à Corrupção (SCC) e Ouvidoria-Geral da União (OGU)[1], tem a SFC como o órgão central do Sistema de Controle Interno do Poder Executivo Federal, a qual foram conferidas atribuições por meio da Lei 10.180[2], de 6 de fevereiro de 2001.

Dentre as competências previstas no Art. 24 dessa Lei, nos incisos VI e VII estão dispostas responsabilidades que dão amparo à atuação da CGU no sentido de promover a melhoria da gestão e proteção da *res publica* e que estão aderentes à necessidade de fiscalização dos atos administrativos relativos às compras públicas, a saber:

> [...]
>
> VI - realizar auditoria sobre a gestão dos recursos públicos federais sob a responsabilidade de órgãos e entidades públicos e privados;
>
> VII - apurar os atos ou fatos inquinados de ilegais ou irregulares, praticados por agentes públicos ou privados, na utilização de recursos públicos federais e, quando for o caso, comunicar à unidade responsável pela contabilidade para as providências cabíveis;
>
> [...]

Adicionalmente, vale salientar que, de acordo com o Art. 169 da Lei 14.133[3], as contratações públicas deverão ser submetidas ao controle preventivo por parte do órgão central de Controle Interno da Administração Pública.

No que tange à gestão das compras públicas, existe o sistema ComprasNet que, segundo o Manual "Comprasnet : informações gerais / Secretaria de Logística e Tecnologia da Informação[4]", foi:

> [...] desenvolvido para atender a Administração Federal, os Fornecedores e a Sociedade, visando dotar a Administração Pública de um conjunto de ferramentas da tecnologia da informação, voltadas à gestão das compras e dos contratos firmados entre órgãos governamentais e fornecedores de bens e serviços, aumentando para os fornecedores a oportunidade de participação em processos licitatórios, em função de maior difusão, desburocratização e redução nos custos das licitações, promovendo total transparência e controle pela sociedade das ações e decisões, no âmbito das Compras Públicas, garantido, assim, o princípio básico da publicidade, que norteia os procedimentos licitatórios consagrados na Lei n.º 8.666, de 21 de junho de 1993.

Nesse contexto, verifica-se que o referido sistema é crítico para o processo de execução das compras públicas e, portanto, uma das principais fontes de dados e informações dos atos administrativos relativos às aquisições públicas que serão objetos de avaliação por parte do Controle Interno do Poder Executivo Federal.

Assim, ressalta-se que o monitoramento e avaliação periódica dos editais de licitação publicados nesse sistema são ações preventivas que buscam melhorar a eficiência, economicidade e a efetividade das aquisições de bens e serviços por parte da Administração, zelando pela qualidade do gasto e pelo patrimônio público.

---

[1] As competências das unidades constituintes da CGU estão apresentadas em https://www.gov.br/cgu/pt-br/acesso-a-informacao/institucional/historico, acessado em maio de 2022.
[2] Disponível em https://www.planalto.gov.br/ccivil_03/LEIS/LEIS_2001/L10180.htm, acessado em maio de 2022.
[3] Disponível em https://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2021/Lei/L14133.htm, acessada em maio de 2022.
[4] Disponível em http://www.comprasnet.gov.br/ajuda/info.pdf, acessado em maio de 2022.

Relativamente à dinâmica das publicações de editais de licitações na Internet, diariamente são publicados centenas de editais no sítio do ComprasNet[5], com média diária de 300 editais, os quais apresentam características que, se identificadas, poderão ensejar a necessidade de uma verificação mais aprofundada por parte das equipes de auditoria da SFC com o intuito de se evitar contratações ou compras com incorreções ou vícios que possam trazer algum tipo de prejuízo à Administração Pública.

Diante desse número expressivo de publicações, o monitoramento e avaliação manuais desses editais torna-se ineficiente e apresenta fragilidades que potencializam o risco de perda de oportunidades relacionadas ao empreendimento de ações de controle preventivas e tempestivas que possam ter como resultado a recomendação de medidas corretivas ou suspensões de editais.

Visando sanar essa necessidade e mitigar o risco da realização de licitações com irregularidades ou erros, foi desenvolvido o sistema Analisador de Licitações, Contratos e Editais (ALICE) que atua de forma autônoma na realização de análises de dados, estruturados e não estruturados, publicados no ComprasNet.

A ALICE desempenha papel fundamental no apoio às ações de controle preventivas, em especial às auditorias preventivas em licitações que buscam, em até oito dias corridos (prazo mínimo para apresentação das propostas de pregão fixado no Art. 4, inciso V, da Lei 10.520), gerar recomendações acerca dos riscos identificados nas contratações e sobre eventuais irregularidades, evitando-se a materialização dos impactos desses riscos que poderão trazer prejuízos para a Administração.

A ferramenta ALICE implementa trilhas de auditoria[6] que analisam, os editais publicados e enviam alertas (via e-mail ou por meio de *Application Programming Interface* - API) quando determinados padrões textuais são detectados nos textos extraídos dos documentos que constituem os editais. Na figura 1, apresenta-se um esquema simplificado do funcionamento desse sistema, considerando a sua principal fonte de editais (o ComprasNet):

Figura 1 – Funcionamento da ferramenta ALICE

Fonte: CGU.

[5] http://comprasnet.gov.br/acesso.asp?url=/ConsultaLicitacoes/ConsLicitacaoDia.asp

[6] Algoritmos computacionais que realizam análises de artefatos de contratações (ex.: editais, termos de referência, estudo técnico preliminar) ou cruzamentos de bases de dados, implementados com base na legislação ou em problemas anteriormente identificados por órgãos de controle.

As trilhas em comento estão classificadas nas seguintes categorias:

- Trilhas Regex – empregam expressões regulares para a identificação de palavras ou expressões que indiquem potencial situação de irregularidade nos editais, por exemplo: exigência de certidão negativa de protesto ou previsão de retenção de pagamento em função da regularidade fiscal da empresa;
- Trilhas de Cruzamento de Dados – realizam consultas a banco de dados a fim de obter evidências em diversas bases institucionais de possíveis irregularidades ou de achados que possam aumentar o risco de insucesso na contratação, por exemplo: licitantes proibidos de contratar com a administração pública com base em consulta ao Cadastro Nacional de Empresas Inidôneas e Suspensas(CEIS) ou licitante com baixo capital social.

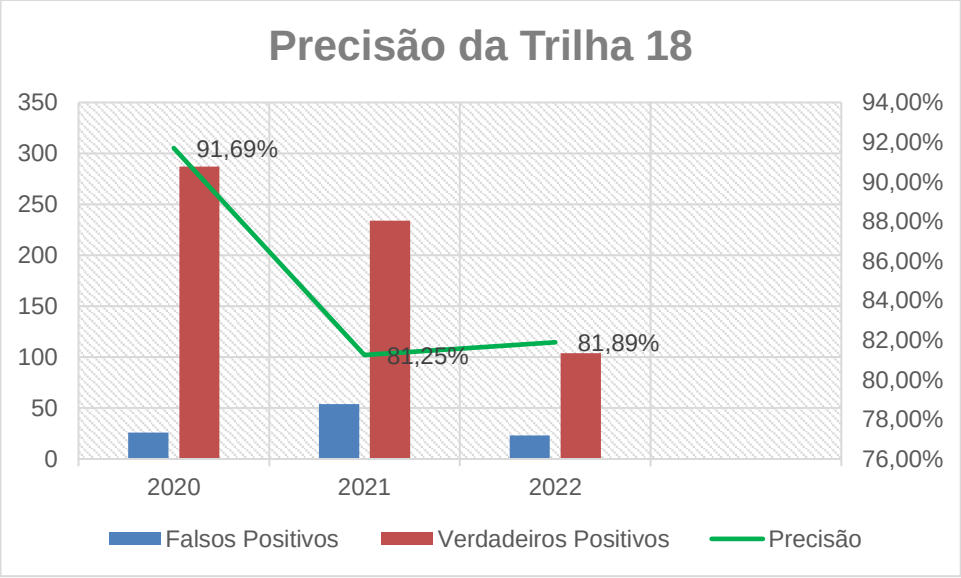
Por fim, ressalta-se que a possibilidade de aprimoramento do processo de criação das trilhas de auditoria e dos seus desempenhos poderá resultar em uma diminuição dos custos de desenvolvimento e manutenção e em um potencial aumento da efetividade do controle preventivo das compras públicas.

1.1. Problema Abordado

Para algumas trilhas que foram implementadas empregando expressões regulares para a identificação de padrões de irregularidade nos editais publicados (Trilhas Regex), observou-se, nos últimos três anos, uma elevação no número de alertas classificados como falsos positivos, o que poderá resultar em diminuição da confiabilidade no sistema.

Para fins de exemplificação do problema, apresenta-se, no Gráfico 1, a precisão e os quantitativos de falsos positivos e de verdadeiros positivos para a trilha “18 - Exigência de comprovação de quadro permanente sem permitir contrato de prestação de serviços” para o período de 2020 a 2022:

Gráfico 1 – Precisão da Trilha 18



Fonte: Autor.

Além disso, as Trilhas Regex exigem conhecimentos especializados em expressões regulares para a configuração de cenários editalícios que tipificam irregularidades relacionadas às contratações públicas de segmentos específicos da Administração, gerando um maior custo de desenvolvimento e manutenção dessas configurações.

1.2. Justificativa

Inicialmente cabe salientar a relevância dos benefícios auferidos com o uso da ferramenta ALICE na prevenção de erros ou fraudes nos processos licitatórios da Administração. Para ilustrar esse fato, foi evidenciado, no período de dezembro de 2018 a junho de 2021, um total de R$ 6 Bilhões[7] em pregões revogados, suspensos ou corrigidos em consequência das auditorias preventivas realizadas com base nos alertas gerados por essa ferramenta.

Benefícios dessa magnitude demonstram o potencial desse instrumental de controle preventivo e a necessidade do seu aprimoramento contínuo em busca de maior economicidade, eficiência e eficácia nas contratações públicas.

Dessa forma, quaisquer ações de melhoria dessa ferramenta, seja no sentido de alcançar melhor desempenho, precisão ou simplificação das configurações das trilhas com a participação direta das equipes de auditoria, trará benefícios para os controles internos das contratações públicas.

Com esses aprimoramentos, vislumbra-se uma redução no número de alertas que sejam classificados como falsos positivos e o aumento da confiabilidade no sistema.

Ademais, salienta-se que a adoção de medidas que possam trazer uma maior eficiência, eficácia e efetividade no combate à corrupção está em harmonia com o disposto no Plano Anticorrupção do Governo Federal, para o período de 2020 a 2025, em especial com o subtema Controle Interno, ação 47 que se refere ao “Desenvolvimento de funcionalidade do Sistema Alice para auxiliar na gestão”, e que visa:

> Ampliar as funcionalidades do Sistema Alice, para pesquisa e mineração de dados, inclusive com a utilização do instrumento da inteligência artificial, bem como implementação de solução para a utilização do sistema por gestores, permitindo que estes realizem as correções necessárias de acordo com as inconsistências apontadas pelo sistema, inclusive aquelas associadas a riscos de fraudes.[8]

Assim, o referido Plano constitui-se em um referencial estratégico com o qual o trabalho desenvolvido mantém alinhamento e coerência.

[7]Fonte: Controladoria-Geral da União - CGU.
[8] Disponível em https://www.gov.br/cgu/pt-br/anticorrupcao/plano-anticorrupcao.pdf.

1.3. Objetivos

O objetivo principal deste trabalho é desenvolver alternativas que possibilitem a melhoria do processo de identificação de editais de licitações com indícios de irregularidades ou erros por meio do aprimoramento do mecanismo de análise de editais, atualmente implementado por meio de expressões regulares.

Em suma, os objetivos principal e secundários são estes:
Principal:

- Simplificar o mecanismo de busca de editais utilizando técnicas de Recuperação de Informações baseadas no processamento de linguagem natural, sem a necessidade de conhecimentos técnicos em expressões regulares;

Secundários:

- Diminuir o custo de desenvolvimento e manutenção de trilhas que analisam documentos; e
- Melhorar a precisão dessas trilhas.

2. Fundamentação Teórica

Segundo (MANNING; RAGHAVAN; SCHÜTZE, 2008), a Recuperação de Informações (*Information Retrieval* – IR) pode ser definida como:

> [...] encontrar material (normalmente documentos) de natureza não estruturada (usualmente textos) que satisfaça uma determinada necessidade de informação, dentro de um grande conjunto de documentos normalmente armazenado em computadores.

Segundo (BAEZA-YATES; RIBEIRO-NETO, 2011), a IR trata de diversos tipos de itens de informação e possui diversos processos (representação, armazenamento, organização e acesso) que atuam sobre eles, conforme o salientado abaixo:

> Recuperação de Informação lida com a representação, armazenamento, organização e acesso a itens de informação tais como documentos, páginas da Web, catálogos online, registros estruturados e semiestruturados e objetos do tipo multimídia. A representação e organização dos itens de informação deverão prover aos usuários uma forma fácil de acesso às informações do seu interesse.

Portanto, verifica-se que o problema abordado está aderente às definições apresentadas e encaixa-se no cenário de recuperação de documentos, o qual emprega palavras ou expressões-chave em alto nível (texto livre) no processo de recuperação.

Dentre os modelos clássicos de IR referenciados em (BAEZA-YATES; RIBEIRO-NETO, 2011) e (MANNING; RAGHAVAN; SCHÜTZE, 2008), destaca-se o modelo vetorial que é empregado na representação de documentos como vetores em um espaço vetorial.

Tal representação vetorial constitui-se em entrada para a aplicação de algoritmos de verificação de similaridade entre esses vetores, tais como: distância euclidiana, similaridade de cossenos e *Triangle Area Similarity – Sector Area Similarity (TS-SS)* (HEIDARIAN; DINNEEN, 2016).

A obtenção da representação vetorial pode ser alcançada por meio de diversas técnicas que são referenciadas como técnicas utilizadas no processo de representação de textos em vetores, onde palavras com significados similares possuem representações vetoriais similares (*word embedding)* (LI; GONG, 2021). Dentre elas destacam-se:

- *Term Frequency Inverse Document Frequency* (TF-IDF), utiliza métricas estatísticas (frequência e o inverso da frequência dos termos ou palavras) para definir o grau de importância (peso) de palavras dentro de um documento ou coleção de documentos (RAMOS et al., 2003);

- *Best Match 25 (BM25)*, trata-se de uma melhoria do TF-IDF que classifica os documentos pelas probabilidades logarítmicas de sua relevância (ROBERTSON; ZARAGOZA, 2009) e (LIU et al., 2009);

- *Global Vectors for Word Representation* (GloVe), trata-se de um modelo de aprendizagem não-supervisionada para representação de palavras (PENNINGTON; SOCHER; MANNING, 2014);

- *Word2Vec*, utiliza um modelo de redes neurais para aprender acerca dos relacionamentos entre as palavras considerando aspectos semânticos (CHURCH, 2017);

- *Bidirectional Encoder Representations from Transformers* (BERT), utiliza técnicas de aprendizagem de máquina baseada em *transformer* (modelo de *deep learning*) para Processamento de Linguagem Natural (YANG; ZHANG; LIN, 2019) e (SUN et al., 2019).

Desse modo, verifica-se o potencial de viabilidade do emprego de tais algoritmos para a definição do nível de similaridade entre os documentos de configuração das trilhas que conterão as expressões ou palavras que determinam casos de irregularidades e os editais, com base em operações sobre os vetores que os representarão.

Por fim, apresentam-se alguns trabalhos que empregaram técnicas acima relacionadas de forma aplicada, inclusive no contexto de detecção de fraudes:

- *Measuring Document Similarity with Word Embeddings* que "detalha uma metodologia para estimar a semelhança textual entre dois documentos, levando em consideração a possibilidade de que duas palavras diferentes tenham um significado semelhante" (SEEGMILLER; PAPANIKOLAOU; SCHMIDT, 2022).

- *Financial Fraud Detection Using Text Mining* que apresenta um projeto que tem por objetivo "implementar uma aplicação de processamento de linguagem natural, baseada em BERT, para analisar informações de relação de dados de texto e ajudar a detectar fraudes financeiras." (ZHENG, 2022)

- *Simple applications of BERT for ad hoc document retrieval* apresentam estudo "seguindo os sucessos recentes na aplicação do BERT para responder a perguntas, exploramos aplicativos simples para recuperação de documentos ad hoc." (YANG; ZHANG; LIN, 2019).

- *A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms* realiza "um levantamento de última geração (estado da arte) no qual analisa algoritmos baseados em densidade para criação de clusters de documentos. Além disso, as medidas de similaridade e avaliação são investigadas com base nos algoritmos selecionados." (MOHAMMED; JACKSI; ZEEBAREE, 2021)

- *Applying BERT to document retrieval with birch* que "apresenta o *Birch*, um sistema que aplica o BERT à recuperação de documentos por meio da integração com o kit de ferramentas de recuperação de informações *Anserini* de código aberto para demonstrar uma pesquisa de ponta a ponta em grandes coleções de documentos" (YILMAZ et al., 2019).

3. Metodologia

A fim de atingir os objetivos propostos, foi adotada a estratégia de desenvolver ranques das licitações baseados no cálculo da similaridade por cossenos e no cálculo da distância euclidiana das representações vetoriais dos textos dos arquivos das licitações e do texto que representa o conjunto de palavras-chave que tipificam situações indesejadas.

Nesse cenário, o texto que representa o conjunto de palavras-chave é denominado de query.

Assim, os referidos ranques apresentam as licitações que possuem arquivos com conteúdos mais similares à query informada.

A seguir são apresentados os detalhes das etapas da metodologia adotada:

1. **Definir Fontes dos Dados**:

Foram utilizadas as seguintes fontes de dados:

- Base de dados do e-Aud[9] (banco de dados no MS SQL server) – Foram explorados os registros dos alertas enviados pela ferramenta ALICE;

- Base de dados da ALICE (banco de dados no MS SQL Server) – Utilizada para extração dos textos dos editais de licitação;

- Arquivo de configuração da Trilha Regex "18 - Exigência de comprovação de quadro permanente sem permitir contrato de prestação de serviços" – contém as palavras-chave que tipificam irregularidades e, também, aquelas que se presentes descartam a emissão de alertas (palavras-negativas).

[9] Sistema de gestão da atividade de auditoria interna governamental no âmbito da CGU.

2. **Preparar Ambiente de desenvolvimento**:

Nesta etapa houve a seleção das ferramentas de desenvolvimento que foram utilizadas no projeto. As ferramentas utilizadas foram estas:

- Jupyter Notebook com Python 3.9.7;
- Microsoft SQL Management Studio;
- Notepad++;
- Bibliotecas utilizadas:
  - ✓ Numpy;
  - ✓ Pandas;
  - ✓ Spacy;
  - ✓ Ntlk;
  - ✓ Sklearn.

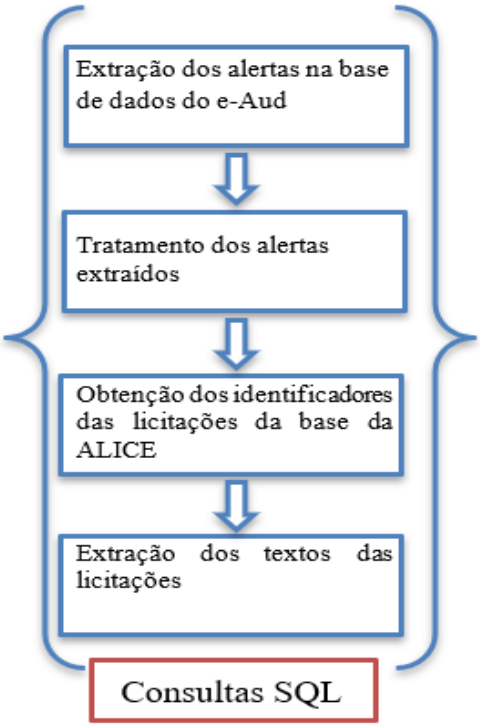
3. **Execução**:

I. Selecionar Trilha Regex:
Optou-se pela seleção de apenas uma Trilha Regex dentre um conjunto de trilhas com maiores reportes de falsos-positivos (alertas improcedentes).
A trilha selecionada foi a Trilha Regex "18 - Exigência de comprovação de quadro permanente sem permitir contrato de prestação de serviços". A escolha dessa trilha foi motivada pela perda de precisão conforme o demonstrado no Gráfico 1 acima.

II. Definir período de análise;
O período de análise foi de 2020 até 2022 (agosto).

III. Identificar os alertas gerados para a trilha em análise;

Para a extração dos alertas referentes à trilha selecionada, foram executados os seguintes passos descritos no diagrama abaixo:



1. No passso "Extração dos alertas na base de dados do e-Aud" foram realizadas consultas SQL[10] na base de dados do e-Aud para obtenção dos alertas enviados pela ALICE, sendo cada alerta detentor de um atributo que o define como improcedente (falso-positivo) ou não. Foram extraídos 730 alertas;

2. No passo "Tratamento dos alertas extraídos", para cada alerta recuperado foi gerada uma chave composta pelos atributos "número do processo", "identificador da UASG" e "identificador da modalidade da licitação" para que fosse possível a obtenção dos identificadores das licitações no banco de dados da ferramenta ALICE. Neste passo, houve a necessidade da criação de funções em *Transact SQL (T-SQL)* para a extração desses atributos a partir de campos textuais;

3. No passo "Obtenção dos identificadores das licitações da base da ALICE" foram utilizadas as chaves compostas criadas no passo anterior para obtenção da relação dos identificadores das licitações associadas aos alertas;

4. Finalmente, no passo "Extração dos textos das licitações", para os identificadores de licitações obtidos no passo anterior, foram extraídos os textos dos arquivos que constituem as licitações. Este passo resultou em um conjunto de 6449 textos para análise;

[10] Como as consultas SQL expõem os nomes e estruturas dos bancos de dados de sistemas críticos da CGU, elas não serão apresentadas neste trabalho.

IV. Desenvolver ensaio no *Jupyter Notebook* com interpretador *python*[11]:

Concluída a extração de dados realizada no item III, foram realizadas as cargas desses dados em dataframes do *pandas* e efetuada a eliminação de registros que não possuiam textos associados a uma determinada licitação.

Para a realização dos tratamentos dos textos das licitações e dos cálculos de similaridade e das distâncias euclidianas foram criadas quatro classes em *python* com responsabilidades bem definidas:

- Classe Preparador - responsável pelo tratamento do texto da *query* e dos textos dos arquivos das licitações. Os seguintes tratamentos foram implementados:
    a. *tokenização* (transformação dos textos em um conjunto de palavras);
    b. conversão dos textos em minúsculas;
    c. lematização;
    d. remoção de espaços e de tudo que não seja letra;
    e. remoção de *stopwords*.
- Classe Vetorizador – responsável pela vetorização dos textos utilizando uma determinada técnica de "word embedding". A operação implementada utiliza o vetorizador TF-IDF do *scikit learn*;
- Classe Query – responsável pela criação de uma *query* que será utilizada por um objeto Analisador;
- Classe Analisador - responsável pela realização do cálculo de similaridade por cossenos e da distância euclidiana entre uma *query* e os documentos de uma licitação.

Salienta-se que somente foi implementada a técnica de representação vetorial TF-IDF e que a *query* foi simplificada para conter apenas as palavras positivas, ou seja, palavras que representam prováveis situações de irregularidade em editais, o que difere dos arquivos de configuração das atuais Trilhas Regex que fazem uso também de palavras negativas que se presentes em um edital impedem o disparo de um alerta.

A *query* foi escrita em texto livre e se constitui em uma adaptação realizada a partir das expressões regulares presentes no arquivo de configuração da Trilha Regex escolhida.

Para a mensuração do desempenho das implementações em estudo, foi selecionada a métrica precisão que é definida como a relação entre o número de verdadeiros positivos (TP – *True positives*) e o somatório dos verdadeiros positivos com a quantidade de falsos positivos (FP – *False positives*):

$$\boldsymbol{Precisão} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

[11] *Link* para acesso ao código fonte está disponível no anexo I.

Para o cálculo da referida métrica foi utilizado um atributo, mantido na base do sistema e-Aud, que indica se um alerta foi procedente ou improcedente. Tal indicação é realizada por um auditor(a) responsável pela análise de procedência de um determinado alerta.

Uma vez geradas as representações vetoriais para os textos dos documentos e para a *query*, foram obtidas as similaridades por cosseno e as distâncias euclidianas, permitindo a geração dos ranques por similaridade e por distância.

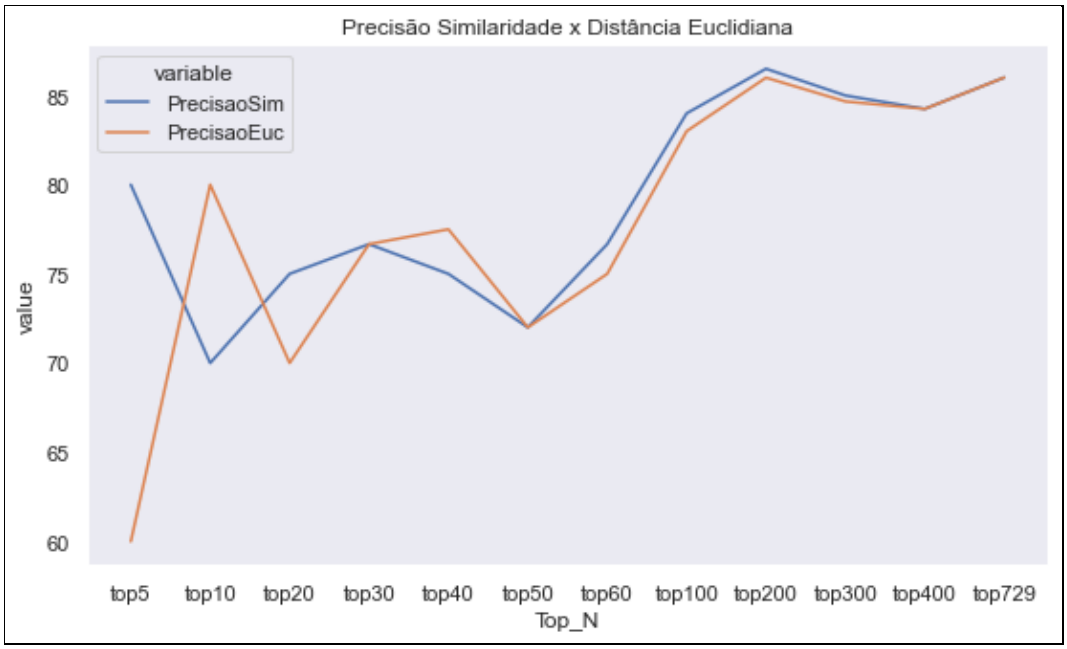
Os ranques em comento foram ordenados de forma decrescente pelas similaridades por cossenos e de forma crescente pelas distâncias euclidianas, possibilitando a manutenção no topo dos ranques dos textos dos documentos mais próximos da *query* em termos de similaridade e de distância euclidiana.

A partir desses ranques, com base no indicador de improcedência oriundo do sistema e-Aud, foram calculadas as precisões de ranques com tamanhos variados a fim de avaliar o desempenho da implementação, cujos resultados serão apresentados na próxima seção deste estudo.

4. Resultados

O Gráfico 2, abaixo, apresenta as precisões dos ranques constituídos pelos "*top n*" documentos mais similares com a *query* e de menor distância euclidiana da *query*. As referidas precisões foram calculadas com base no marcador de improcedência (falso positivo) definido no sistema e-Aud.

Gráfico 2 – Precisão Similaridade x Distância Euclidiana



Fonte: Autor.

Nesse gráfico, observa-se que a precisão para os ranques que representam as licitações com os documentos mais similares ou mais próximos da *query* aparentemente não foi satisfatória porque, intuitivamente, os mais similares/próximos deveriam ter indicador de improcedência negativado, o que pode ser explicado pela presença nos ranques iniciais ("top5" até o "top50") de uma maior proporção de licitações com indicador de improcedência ativo em relação ao número de procedentes, o que pode ser uma peculiaridade dessa amostra (vide Tabela 1 abaixo).

Tabela 1 – Proporção de Improcedentes e Precisão

Top_N	Proporção de Improcedentes (Similaridade)	Proporção de Improcedentes (Dist.Euclidiana)	Precisão (Similaridade)	Precisão (Dist.Euclidiana)
top5	0,2	0,4	0,8	0,6
top10	0,3	0,2	0,7	0,8
top20	0,25	0,3	0,75	0,7
top30	0,233333333	0,233333333	0,766666667	0,766666667
top40	0,25	0,225	0,75	0,775
top50	0,28	0,28	0,72	0,72
top60	0,233333333	0,25	0,766666667	0,75
top100	0,16	0,17	0,84	0,83
top200	0,135	0,14	0,865	0,86
top300	0,15	0,153333333	0,85	0,846666667
top400	0,1575	0,1575	0,8425	0,8425
top729	0,139917695	0,139917695	0,860082305	0,860082305

Fonte: Autor.

Contudo, também na Tabela 1, verifica-se que à medida que os tamanhos dos ranques aumentaram as proporções de improcedentes diminuiram, resultando em um aumento da precisão para os ranques finais.

Tal comportamento também pode ser reflexo da ausência de tratamento de palavras negativas[12] ou de alguma falha no conteúdo da *query*, tais como a perda de frases ou expressões no processo de tradução das expressões regulares para texto livre, cuja revisão poderia trazer maior precisão.

Por fim, salienta-se que a geração dos ranques permitirá, de imediato, a seleção de editais para a realização de análises mais pormenorizadas por parte das equipes de auditoria ou por algum mecanismo classificador para, por exemplo, a geração de alarmes.

[12] Palavras utilizadas para invalidar uma tipificação de irregularidade. Tais palavras também fazem parte dos arquivos de configuração de Trilhas Regex.

5. Conclusão e Trabalhos Futuros

5.1. Conclusão

Neste trabalho foi realizada a implementação de um mecanismo de recuperação de editais de compras públicas com base em um conjunto de expressões e palavras-chave que tipificam potenciais irregularidades, utilizando-se de técnicas de Recuperação de Informações.

O objetivo principal de simplificar o mecanismo de busca de editais utilizando técnicas de Recuperação de Informações baseadas no processamento de linguagem natural, sem a necessidade de conhecimentos técnicos em expressões regulares foi alcançado uma vez que neste ensaio a query foi criada e utilizada na identificação de documentos similares a ela sem o emprego das referidas expressões.

Além disso, considera-se também alcançado o primeiro objetivo secundário porque o emprego de texto livre na configuração das *queries* simplifica e diminui os custos de manutenção e desenvolvimento.

A implementação trouxe resultados positivos em relação à diminuição da complexidade de configuração de uma nova trilha, pois as configurações poderão ser efetuadas utilizando-se da linguagem natural ao invés de expressões regulares, o que permitirá que essas configurações possam ser executadas por pessoas sem conhecimentos de expressões regulares.

Outro aspecto benéfico foi a geração de ranques que permitirão a identificação das licitações que possuem documentos com maior similaridade à *query* configurada, o que possibilitará o desenvolvimento de classificadores para determinar se um alerta deverá ser emitido ou não com base nas representações vetoriais dos documentos dessas licitações.

Devido ao tempo exíguo, o estudo ficou centrado em uma trilha de auditoria do tipo “Trilha Regex”, porém isso não invalida o emprego das técnicas de recuperação de informações aqui utilizadas nas demais trilhas dessa categoria, pois os padrões de busca a serem atendidos estão configurados nos arquivos que representam as *queries*.

Sobre o código gerado, houve a preocupação de implementar as principais funcionalidades em Classes a fim de facilitar a inclusão de novos métodos de representação vetorial de textos e de algoritmos de cálculo de similaridade.

Assim, verifica-se que os resultados atingidos com este estudo têm potencial para inspirar projetos de melhoria ou serem refinados e aproveitados em futuras manutenções evolutivas da ferramenta ALICE.

5.2. Trabalhos futuros

- Com base em consultas a especialistas em compras públicas de determinadas áreas ou em consultas às unidades de auditoria de áreas específicas, revisar as palavras positivas e implementar o uso das palavras negativas nas configurações das *queries*;

- Implementar outras técnicas de representação vetorial, tais como *Best Match 25 (BM25), Word2Vec* ou modelagem de tópicos usando *Latent Dirichlet Allocation(LDA)*;

- Implementar outros algoritmos para o cálculo de similaridade ou distância entre dois vetores, tais como a *City Block (Manhattan) Distance;*

- A partir dos ranques de similaridade ou de distância euclidiana e das representações vetoriais dos documentos, construir um classificador para emissão de alertas;

6. Referências Bibliográficas

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Modern Information Retrieval**: the concepts and technology behind search. 2. ed. Essex, England: Pearson, 2011.

BRASIL. Ministério do Planejamento, Orçamento e Gestão. Secretaria de Logística e Tecnologia da Informação. **Comprasnet : informações gerais / Secretaria de Logística e Tecnologia da Informação.** Brasília: MP, 2005. 14p.

CHURCH, Kenneth Ward. **Word2Vec. Natural Language Engineering**, v. 23, n. 1, p. 155-162, 2017. Disponível em: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/B84AE4446BD47F48847B4904F0B36E0B/S1351324916000334a.pdf/div-class-title-word2vec-div.pdf . Acesso em: maio de 2022.

HEIDARIAN, Arash; DINNEEN, Michael. **A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering**. 2016. Disponível em: https://www.researchgate.net/publication/303513110_A_Hybrid_Geometric_Approach_for_Measuring_Similarity_Level_Among_Documents_and_Document_Clustering. Acesso em: maio de 2022.

LI, Saihan; GONG, Bing. **Word embedding and text classification based on deep learning methods**. In: MATEC Web of Conferences. EDP Sciences, 2021. p. 06022. Disponível em: https://www.matec-conferences.org/articles/matecconf/abs/2021/05/matecconf_cscns20_06022/matecconf_cscns20_06022.html . Acesso em: maio de 2022.

LIU, Tie-Yan et al. **Learning to rank for information retrieval. Foundations and Trends® in Information Retrieval**, v. 3, n. 3, p. 225-331, 2009. Disponível em: https://www.nowpublishers.com/article/DownloadSummary/INR-016 . Acesso em: maio de 2022.

MANNING, Christhofer D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to Information Retrieval**. New York: Cambridge University Press, 2008.

MOHAMMED, Shapol M.; JACKSI, Karwan; ZEEBAREE, S. **A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms**. Indonesian Journal of Electrical Engineering and Computer Science, v. 22, n. 1, p. 552-562, 2021. Disponível em: https://www.academia.edu/download/66181373/23698_47236_1_PB.pdf. Acesso em: maio de 2022.

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. **Glove: Global vectors for word representation**. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. 2014. p. 1532-1543. Disponível em: https://nlp.stanford.edu/pubs/glove.pdf. Acesso em: maio de 2022.

RAMOS, Juan et al. **Using tf-idf to determine word relevance in document queries**. In: **Proceedings of the first instructional conference on machine learning**. 2003. p. 29-48. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf . Acesso em: maio de 2022.

ROBERTSON, Stephen*;* ZARAGOZA, Hugo. **The Probabilistic Relevance Framework: BM25 and Beyond**. 2009. *NOW Publishers, Inc.* ISBN 978-1-60198-308-4. Disponível em: http://staff.city.ac.uk/~sb317/papers/foundations_bm25_review.pdf . Acesso em: maio de 2022.

SEEGMILLER, Bryan; PAPANIKOLAOU, Dimitris; SCHMIDT, Lawrence. **Measuring Document Similarity with Word Embeddings**. 2022. Disponível em: SSRN: https://ssrn.com/abstract=4088443 ou http://dx.doi.org/10.2139/ssrn.4088443. Acesso em: maio de 2022.

SUN, Chi et al. **How to fine-tune bert for text classification?**. In: China national conference on Chinese computational linguistics. Springer, Cham, 2019. p. 194-206. Disponível em: https://arxiv.org/pdf/1905.05583.pdf . Acesso em: maio de 2022.

YANG, Wei; ZHANG, Haotian; LIN, Jimmy. **Simple applications of BERT for ad hoc document retrieval**. 2019. Disponível em: https://arxiv.org/pdf/1903.10972.pdf. Acesso em: maio de 2022.

YILMAZ, Zeynep Akkalyoncu et al. **Applying BERT to document retrieval with birch.** In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. 2019. p. 19-24. Disponível em: https://aclanthology.org/D19-3004.pdf. Acesso em: maio de 2022.

ZHENG, Zhou. **Financial Fraud Detection Using Text Mining**. 2022. Disponível em: https://wp.cs.hku.hk/2021/fyp21078/wp-content/uploads/sites/242/2022/01/Intermediate-Report.pdf. Acesso em: maio de 2022.

7. Anexo I – Código Fonte

- O código fonte está disponível no seguinte endereço: https://drive.google.com/file/d/1daybv49XcFpinTugw5lqULP6kBqZtQCb/view?usp=sharing .