UNIVERSIDADE FEDERAL DO MARANHÃO
CAMPUS BACANGA

Thiago Gustavo Vieira de Paiva

EXPANSÃO DE QUERIES COM WORD EMBEDDINGS

São Luís
2021

THIAGO GUSTAVO VIEIRA DE PAIVA

EXPANSÃO DE QUERIES COM WORD EMBEDDINGS

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

São Luís-MA, 22 de Outubro de 2021

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Vieira de Paiva, Thiago Gustavo.
Expansão de Queries com Word Embeddings / Thiago Gustavo Vieira de Paiva. - 2021.
33 f.

Orientador(a): Tarcísio Souza Costa.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís, 2021.

1. Expansão de Queries. 2. Recuperação de Informação. 3. Word Embeddings. I. Souza Costa, Tarcísio. II. Título.

UNIVERSIDADE FEDERAL DO MARANHÃO

Thiago Gustavo Vieira de Paiva

Aprovada em 22 de outubro de 2021.

Banca examinadora:

Orientador: Prof. Msc. Tarcísio Souza Costa
Instituto Federal do Maranhão - IFMA

Prof. Dr. Geraldo Braz Júnior
Universidade Federal do Maranhão - UFMA

Prof. Dr. Ivo José da Cunha Serra
Universidade Federal do Maranhão - UFMA

São Luís, 22 de Outubro de 2021

Dedico este trabalho a todos aqueles que superaram limites e permaneceram acreditando em si.

Agradecimentos

Aos meus pais, Sílvia Teresa Vieira de Paiva e Arnóbio Cardoso de Paiva. Ao meu pai, pelo modelo de dedicação ao trabalho e superação de limites históricos. À minha mãe e professora, por sempre acreditar em mim, realizando inúmeros esforços e sacrifícios pela minha formação acadêmica, e, sobretudo, minha educação enquanto SER.

Aos meus irmãos, Ellen Caroline Vieira de Paiva e Matheus Vieira de Paiva, pelas vivências compartilhadas no processo educativo comum no qual se desenvolveu minha identidade.

À minha noiva, Lussandra Barbosa de Carvalho, pela cumplicidade e companheirismo diários; por sonhar comigo e ser parte dos meus sonhos.

À Profª Maria Lourdes Cardoso da Costa, pelo grande incentivo no meu processo de alfabetização ainda na Educação Infantil. Obrigado por me fazer crer que eu teria capacidade de terminar o caderno de caligrafia! É importante recordar seu trabalho agora em que concluo a escrita deste caderno de monografia.

Ao Prof. Dr. Anselmo Cardoso de Paiva, da Universidade Federal do Maranhão, pela inspiração, motivação e apoio que se fazem presentes desde o início desta caminhada.

Ao meu orientador, Prof. Msc. Tarcísio Souza Costa, do Instituto de Educação, Ciência e Tecnologia do Maranhão, amizade nascida logo na primeira semana de universidade e que se estendeu para além dela; sou-lhe grato pela participativa orientação acadêmica.

Ao Prof. Dr. Wolfgang Nejdl, da Leibniz Universität Hannover, e à Prof. Drª. Eirini Ntoutsi, da Freie Universität Berlin, pela oportunidade de aprendizado especializado no âmbito de pesquisa que culminou no presente trabalho.

Ao Prof. Dr. Geraldo Braz Júnior e ao Prof. Dr. Ivo José da Cunha Serra, pela gentileza e disponibilidade de compor a banca avaliadora deste trabalho.

Resumo

Neste trabalho, uma nova abordagem para expansão de *queries* é apresentada usando word embeddings. Esta abordagem incorpora recursos presentes em expressões para construir um modelo de expansão de *queries*, a fim de recuperar documentos relevantes considerando a similaridade entre os termos e a *query original*. Através do modelo *CBOW* do *Word2Vec*, os termos selecionados são semanticamente relacionados às expressões, além de *keywords* extraídas de URLs, títulos e dentro do próprio conteúdo de texto dos documentos. Os *scores* obtidos pelo *Word2Vec* para esses termos são usados para selecionar os melhores termos candidatos para compor as *queries expandidas*. Também realizamos o re-ranqueamento dos documentos baseado no local de contexto onde os termos foram encontrados em cada documento. O método proposto é avaliado num *dataset* de artigos de notícias, *The New York Times Annotated Corpus* e demonstra que os recursos propostos usados para expansão podem efetivamente aumentar a precisão dos documentos recuperados em comparação com os documentos recuperados pela *query original*.

Palavras-chave: Recuperação de Informação, word embeddings, expansão de queries.

Abstract

In this work we present a novel approach for query expansion. This approach incorporates word embedding features and expressions to build a query expansion model, in order to retrieve relevant documents considering the similarity of terms to the original query. By using *Word2Vec's CBOW* embedding approach, we select terms that are semantically related to the expressions and keywords from the document URLs, title and document content. The scores obtained by *Word2Vec* are used to select the best candidate terms for the expanded queries. We also re-rank documents based on where the relevant context was found in each document. The proposed method is evaluated using a *dataset* containing news articles, *The New York Times Annotated Corpus* which demonstrates the features used for the query expansion to increase the precision of retrieved documents in comparison to the retrieved documents from the original query.

Keywords:Information retrieval, word embeddings, query expansion.

Sumário

1 INTRODUÇÃO 10

2 OBJETIVOS 12
2.1 Gerais 12
2.2 Específicos 12

3 JUSTIFICATIVA 13

4 TRABALHOS RELACIONADOS 14
4.1 Recuperação de Informações 14
4.2 Expansão de Queries com Word Embeddings 14

5 METODOLOGIA 16
5.1 Corpus Anotado do New York Times 17
5.2 Construção do Modelo Online 17
5.2.1 Pré-processamento de URLs 17
5.2.2 Indexação de Documentos 18
5.3 Query 18
5.4 Algoritmo BM25 18
5.5 Expansão da Query 19
5.5.1 Pré-processamento das Notícias 20
5.5.2 Treinamento Word2vec 20
5.5.3 Query Expandida 20
5.5.4 Comparação de Documentos 21

6 RESULTADOS 22

7 CONCLUSÃO 31
7.1 Trabalhos Futuros 31

REFERÊNCIAS 33

1 Introdução

Query Expansion é um tema importante na área de Recuperação de Informação, do inglês *Information Retrieval* (IR), que se baseia na modificação de consultas (*queries*) automaticamente. Esta modificação é realizada através da adição de novos termos (expansão) de modo iterativo com o objetivo de recuperar novos documentos relevantes a cada execução da expansão. A pesquisa em *Query Expansion* na área de IR é vasta e, recentemente, muitos trabalhos têm sido direcionados para expansão dos termos através de similaridade baseada em *word embeddings* [Roy et al. 2016].

Satisfazer necessidades de informação, tais como recuperar documentos pertencentes a um dado tópico, é uma tarefa complexa.

Vários trabalhos abordaram o problema da recuperação de informação temporal [Li e Croft 2003] [Dakka, Gravano e Ipeirotis 2012] [Efron e Golovchinsky 2011], mas muitas das abordagens recentes em expansão de queries são focadas em documentos pequenos [Rao e Lin 2016] [Wang, Huang e Feng 2017], como *tweets* e faltam aplicações que solucionem esse problema nos arquivos de notícias. Além disso, o uso de word emdeddings na recuperação de informação é normalmente relacionada à recuperação adhoc.

Neste trabalho, o problema de encontrar documentos relevantes em coleções de notícias é abordado e consiste em recuperar os documentos que são topicamente relevantes.

Particularmente, o problema é tratado usando da expansão de queries com word embeddings. Abordagem esta inspirada pelo fato do uso de word embeddings ter demonstrado facilitar a previsão de contexto de vocabulário de termos eficientemente [Mikolov et al. 2013].

Word embeddings visa a representação de palavras através da similaridade que há entre várias outras palavras. Essa representação baseada nos sinônimos busca o mapeamento das palavras em vetores de números reais que aplicadas a um dicionário, a similaridade será notada por haver a correspondência na representação numérica desta.

O método abordado envolve a seleção de palavras-chave de eventos extraídos diretamente do EventKG [Gottschalk e Demidova 2019], uma base de conhecimento multilíngue que incorpora informações relacionadas a eventos.

O EventKG é derivado de vários outras bases de conhecimento de larga escala, como Wikidata [Vrandecic e Krötzsch 2014], DBpedia [Lehmann et al. 2014] e YAGO [Suchanek, Kasneci e Weikum 2007], bem como fontes menos estruturadas, como o Portal de Eventos da Wikipedia e listas de eventos da Wikipedia em 15 idiomas [1].

[1] https://en.wikipedia.org/wiki/Portal:Current_events

As palavras-chave extraídas do EventKG são as queries iniciais, e a indexação do *The New York Times Annotated Corpus* é realizada pelo *Elastic Search* para posterior extração do contexto. O contexto é obtido através de treinamentos de modelos de word embeddings (como o *BM25*) diretamente nos documentos. A decisão final dos termos a serem selecioandos é baseada nos scores obtidos de tais modelos.

Os documentos são usados para obter termos candidatos à expansão da consulta e os termos serão assim obtidos pela recuperação dos documentos. A seleção final dos termos para expansão de consulta terá base nas pontuações dos termos no modelo de incorporação.

As contribuições deste trabalho são as seguintes:

- Um novo algoritmo para geração de queries expandidas que considera palavras topicamente relacionadas a eventos;
- Uma estrutura extensível de código aberto para geração automática das queries para facilitar a manutenção e atualizações.

2 Objetivos

2.1 Gerais

Este trabalho tem como principal objetivo demonstrar que a utilização de *word embeddings* para expansão de *queries* orientadas é eficiente no que diz respeito ao aumento de informação recuperada.

2.2 Específicos

Como parte do processo de expansão de *queries*, este trabalho pretende demonstrar que:

1. A escolha de termos semanticamente similares a eventos deve aumentar as chances de recuperação de novos documentos, também relevantes ao tópico e/ou evento em destaque;
2. O treinamento local de modelos para expansão de termos similares influencia na qualidade dos termos expandidos.

3 Justificativa

De acordo com a literatura, existem diversos trabalhos que são dedicados à análise de desempenho de métodos que utilizam *word embeddings* com *query expansion*. Porém, o uso de *word embeddings* em eventos ainda é pouco explorado na literatura. Deste modo, faz-se necessária tal análise para que a pesquisa e desenvolvimento de novas técnicas orientadas a eventos sejam desenvolvidas no futuro.

Adicionalmente, as bases de conhecimento existentes (*Knowledge Bases*) são principalmente desenvolvidas para entidades e não para eventos, o que significa que eles não cobrem suficientemente eventos e relações entre entidades. Como consequência, os *datasets* existentes, com exemplos recentes, concentram-se principalmente em queries não orientada a eventos. Recentemente, grafos de conhecimento centrados em eventos como EventKG [Gottschalk e Demidova 2018] e grafos de conhecimento extraídos de notícias (por exemplo, [Leetaru e Schrodt 2013, Rospocher et al. 2016]) foram propostos. Portanto, este trabalho dá os passos no sentido de analisar uma abordagem para expansão de queries voltada a eventos.

4 Trabalhos Relacionados

Este trabalho utiliza word embeddings com suporte à expansão da *query*. Embora hajam trabalhos recentes relacionados à expansão da *query* com *word embeddings*, os métodos e abordagem realizada se diferem dos trabalhos já existentes.

Existe um trabalho recente [Rosin, Guy e Radinsky 2021], que possui abordagem similar, mas se difere deste trabalho quanto à introdução de uma fase de detecção de eventos relacionados ao evento da *query* em seu *pipeline*, além do uso de características temporais.

4.1 Recuperação de Informações

A incorporação de aspectos temporais tem sido amplamente explorada em recuperação de informação (*Information Retrieval*). A forma como as *queries* e os documentos são modelados, levando em consideração características temporais de documentos, pode aumentar a qualidade dos resultados.

Recentemente, Rao e Lin [Rao e Lin 2016] implementaram a ideia de expandir termos que deveriam ser retirados de documentos massivos. Eles utilizaram um modelo contínuo de Markov Escondida (cHMM), que melhor explica a distribuição do conjunto inicial de documentos recuperados, levando a significantes melhoramentos ao explorar traços temporais para o ranqueamento. Wang et al. [Wang, Huang e Feng 2017] propôs um modelo conceitual de feedback a fim de extrair informações implícitas apresentadas em textos curtos de bases de conhecimento externas para o modelo de expansão de *queries*. Eles também incorporaram informações temporais no método proposto para satisfazer necessidades de informação em tempo real.

4.2 Expansão de Queries com Word Embeddings

A incorporação de palavras vem sendo explorada pelas abordagens de expansão de *queries* [0001, Mitra e Craswell 2016] [Almasri, Berrut e Chevallet 2016] [Kuzi, Shtok e Kurland 2016] [Roy et al. 2016]. Diaz et al. [0001, Mitra e Craswell 2016] sugerem o treinamento de *word embeddings*, tal como *Word2Vec* e GloVe, localmente. Documentos sobre subtópicos em uma coleção apresentam diferentes distribuições de unigramas quando comparadas com todo o corpus. Logo, o treinamento em todo o corpus aumentaria a probabilidade de gerar um contexto não relacionado de termos.

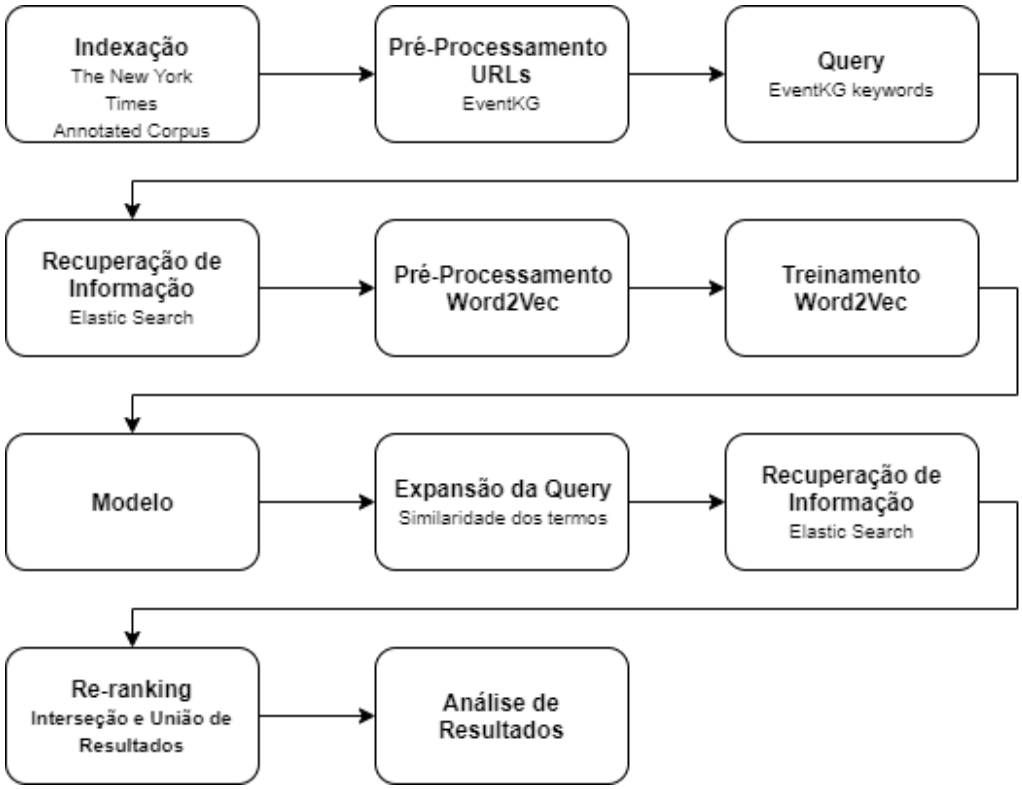
Os resultados sugerem que os benefícios alcançados a partir de incorporações globais

também podem ser alcançados em incorporações locais. ALMasri et al. [Almasri, Berrut e Chevallet 2016] compararam uma aprendizagem profunda baseada na expansão de *queries* contra o *feedback* de pseudo-relevância e informações mútuas. Os resultados confirmam que os vetores de aprendizado profundo são fonte promissora para a expansão de *queries*. Quando usados os termos fornecidos pelo *Word2Vec* para expandir a consulta original, o desempenho da recuperação é significativamente melhorado. Roy et al. [Roy et al. 2016] propôs uma estrutura para expansão de *queries*, usando modelo de linguagem neural distribuída. A abordagem do vizinho mais próximo para obter os termos expandidos foi utilizada.

5 Metodologia

A proposta principal envolve a expansão de *queries*, encontrando e sugerindo termos semanticamente relacionados fazendo uso de word embeddings de maneira a aumentar a precisão e a quantidade de documentos relevantes recuperados. A Figura 1 demonstra o *pipeline* da metodologia aplicada.

Figura 1 – Pipeline da Metodologia



Primeiramente, ocorre a indexação do *The New York Times Annotated Corpus*, fazendo uso do *BM25*. Em seguida, ocorre o pré-processamento de URLs correspondentes a eventos extraídos do *EventKG*. Estas URLs, uma vez processadas, serão as palavras-chaves da *query* inicial. A construção do Modelo Online é realizada através de um pré-processamento dos documentos recuperados que consistirá em prepará-los para o treinamento no *Word2Vec* e, após conclusa a construção deste modelo, uma nova *query* será realizada com os termos mais similares às palavras-chaves da primeira *query*. Por fim, os resultados obtidos das duas *queries* é comparado.

5.1 Corpus Anotado do New York Times

O The New York Times Annotated Corpus[1] é o dataset utilizado neste trabalho. Ele contém artigos publicados no The New York Times desde 01 de Janeiro de 1987 até 19 de Junho de 2007. Ele é composto por uma coleção de documentos XML versão 3.3 seguindo a especificação da *News Industry Format* (NITF)[2] e distribuído pela *Internet Memory* [Blanco, Mika e Batalla 2010]. Possui uma coleção com cerca de 20GB quando descomprimido contendo cerca de 3.8 milhões de documents de 1500 variados blogs e outras fontes de notícias. A Figura 2 ilustra um documento extraído do corpus.

Figura 2 – Exemplo de Documento


```
<nitf change.date="June 10, 2005" change.time="19:30" version="-//IPTC//DTD NITF 3.3//EN">
 - <head>
    <title>A Seated Tour Of Europe</title>
    <meta content="01JOHN$01" name="slug"/>
    <meta content="1" name="publication_day_of_month"/>
    <meta content="1" name="publication_month"/>
    <meta content="2004" name="publication_year"/>
    <meta content="Thursday" name="publication_day_of_week"/>
    <meta content="House & Home/Style Desk" name="dsk"/>
    <meta content="3" name="print_page_number"/>
    <meta content="F" name="print_section"/>
    <meta content="4" name="print_column"/>
    <meta content="Home and Garden; Style" name="online_sections"/>
  - <docdata>
      <doc-id id-string="1547299"/>
      <doc.copyright holder="The New York Times" year="2004"/>
      <series series.name="CURRENTS: FURNISHINGS"/>
    - <identified-content>
        <classifier class="indexing_service" type="descriptor">Chairs</classifier>
        <org class="indexing_service">Johnson & Hicks (NYC)</org>
        <person class="indexing_service">Louie, Elaine</person>
        <classifier class="online_producer" type="taxonomic_classifier">Top/Features/Home and Garden</classifier>
        <classifier class="online_producer" type="taxonomic_classifier">Top/Features/Style</classifier>
        <classifier class="online_producer" type="general_descriptor">Chairs</classifier>
      </identified-content>
    </docdata>
    <pubdata date.publication="20040101T000000" ex-ref="http://query.nytimes.com/gst/fullpage.html?res=9C07E4DE1E3EF932A35752C0A9629C8B63" item-length="174"
    name="The New York Times" unit-of-measure="word"/>
  </head>
 - <body>
  - <body.head>
    - <hedline>
        <hl1>A Seated Tour Of Europe</hl1>
      </hedline>
      <byline class="print_byline">By ELAINE LOUIE</byline>
      <byline class="normalized_byline">Louie, Elaine</byline>
    - <abstract>
      - <p>
          Chairs of 1920's and 30's are featured at Johnson & Hicks, new home furnishings store in TriBeCa; photos (M)
        </p>
      </abstract>
```


5.2 Construção do Modelo Online

A fim de processar e expandir as *queries*, um processo online de construção do modelo, que consiste em criar o vetor representação dos termos dos documentos (*Word2Vec*) usados em toda o aplicação será realizado.

5.2.1 Pré-processamento de URLs

A construção de um modelo melhor envolve um pré-processamento na coleção de documentos. Isso deve ser feito pela remoção de *stopwords* e caracteres especiais das

[1] https://catalog.ldc.upenn.edu/LDC2008T19
[2] http://www.nitf.org

URLs obtidas de eventos existentes no *EventKG* no sentido de extrair as *keywords* que fornecem um certo significado à fase de Recuperação de Informação. Por fim, para que esses documentos sejam recuperados, os mesmos devem ser indexados.

5.2.2 Indexação de Documentos

Os documentos são indexados com *Elastic Search*[3], que posteriormente serão usados na fase online da abordagem, no momento de realização das consultas.

5.3 Query

Obtidas as palavras-chave, a recuperação de informação será realizada sobre os documentos de notícias indexados. Um *score* é atribuído de acordo com a relevância do documento. Desta forma, um dado documento recuperado pontuará melhor, o valor da pontuação é elevado ao quadrado quando os termos mais relevantes são encontrados no *<title>* e no *<abstract>* de cada notícia ou pior quando os termos mais relevantes são encontrados no restante do documento, não havendo nenhuma alteração na sua pontuação. Durante o processo de recuperação, o algoritmo BM25 é utilizado para realizar o rankeamento desses documentos.

5.4 Algoritmo BM25

O algoritmo de ranqueamento utilizado pelo *Elastic Search* padrão é o BM25, cujo nome deriva de *Best Matching* e tem por objetivo estimar a relevância de documentos de acordo com uma *query* dada. Tal algoritmo é formalizado pela Equação 5.1.

$$\sum_{i=1}^{n} \mathrm{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\mathrm{avgdl}})}, \tag{5.1}$$

onde $f(q_i, D)$ é a frequência de termo q_i no documento D, $|D|$ é o comprimento do documento D em palavras, e avgdl é o comprimento médio do documento na coleção de texto a partir do qual os documentos são extraídos. k_1 e b são parâmetros livres.

Este algoritmo ranqueia o conjunto de documentos baseado nos termos da *query* fornecida independente de haver uma proximidade entre tais termos.

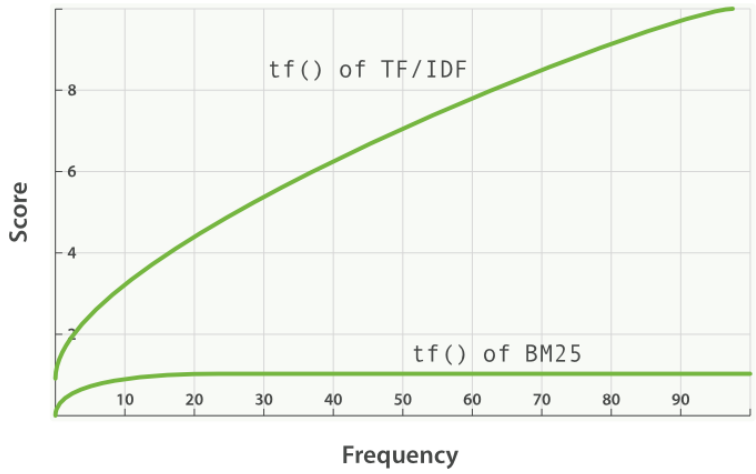
A frequência dos termos referida como *TF* (*Term Frequency*) mede o quão frequente cada termo da *query* ocorre num documento. Pode haver a ocorrência de cada termo várias vezes num documento e, quanto mais frequente este termo for, maior pontuação ele terá.

[3] https://www.elastic.co/

A frequência inversa do documento, denotada por *IDF* (*Inverse Document Frequency*) mede a relevância de cada termo, o que confere peso a termos com pouca ou rara ocorrência.

O ranqueamento baseado apenas no *TFxIDF* apresenta maior grau de relevância nas primeiras ocorrências dos termos ao longo dos documentos e, a partir de então, os incrementos na pontuação de cada termo passa a ser menos significativa quanto mais frequente. Diferentemente, no *BM25*, ocorre uma penalização com a alta ocorrência de modo que uma normalização da pontuação tende a atenuar os aumentos da pontuação a cada nova ocorrência deste termo. O gráfico ilustrado pela Figura 3 demonstra exatamente o quão sensível o *BM25* é no que se refere à ocorrência dos termos.

Figura 3 – TFxIDF e o BM25



Os valores padrões no *Elastic Search* para a variável de saturação *k1* e a variável *b*, que define até que ponto o tamanho do documento influencia na normalização dos valores de *TF*, são respectivamente 1.2 e 0.75, e foram estes os valores utilizados nesta abordagem.

Nesta abordagem, a pontuação mínima para que um documento seja considerado na lista de notícias recuperadas é 20.000, em que há um peso maior na pontuação, dependendo de onde os termos mais significantes da *query* foram encontrados. Quando tais termos foram localizados nas *tags* <*title*> ou <*abstract*>, a pontuação é elevada ao quadrado, nas demais *tags*, não há diferença na pontuação normalmente atribuída.

5.5 Expansão da Query

Esta fase consiste em expandir os termos da pesquisa original a partir das palavras mais similares obtidas após a fase de Treinamento *Word2Vec*.

5.5.1 Pré-processamento das Notícias

O pré-processamento das notícias recuperadas consiste numa série de procedimentos, que possuem o intuito de preparar os documentos para a conversão num espaço vetorial. Assim, todas as sentenças obtidas de cada documento é combinada num único texto, que logo é convertido para a caixa baixa (*lowercase*), em seguida, ocorre a tokenização que visa separar em tokens cada palavra, a remoção de caracteres especiais que não sejam alfanuméricos e a remoção das *stopwords* da língua inglesa. O documento já processado é então anexado a todos os demais que passaram pelo mesmo processo. Findado todo esse processamento, o documento resultante está pronto para a fase de Treinamento *Word2Vec*.

5.5.2 Treinamento Word2vec

Após o pré-processamento, os termos que representam os documentos serão convertidos em vetores num espaço vetorial (fase de treinamento). Para produzir uma representação distribuída das palavras num espaço vetorial, deve-se usar o modelo de uma bag-of-words contínua [Mikolov et al. 2013], dado que se deseja encontrar palavras de contexto, começando pelas palavras-chave da pesquisa. O modelo encontrará representações de palavras úteis para prever o contexto (palavras ao redor) em frases que incorporam um conjunto de documentos e possuem grande relação contextual. Logo, cada tópico será representado por um modelo de onde os termos de expansão serão obtidos. A construção deste modelo envolve a configuração de alguns parâmetros descritos na Tabela 1.

Tabela 1 – Parâmetros do Word2Vec

Variável	**Descrição**	**Valor**
min_count	Frequência mínima dos termos no documento	2
size	Número (N) do espaço N-dimensional usado para mapear os vetores	100
window	Distância máxima entre a palavra alvo e as demais ao redor	5
sg	Algoritmo de Treinamento - CBOW (0) Skip-Gram (1)	0

Construído o modelo, as palavras-chave da *query* original é utilizada como reforço positivo na obtenção das palavras com maior similaridade e, destas, as três mais similares irão se integrar à *query* original.

5.5.3 Query Expandida

As três palavras mais similares às palavras-chaves da *query* original são unidas numa única *query* expandida com estes novos termos. Uma nova recuperação de documentos é realizada obedecendo os mesmos procedimentos e parâmetros realizados na primeira fase do processo com a *query* original.

5.5.4 Comparação de Documentos

Esta é a última fase da abordagem utilizada e, nela, os *IDs* correspondentes aos documentos recuperados pela *query* original são extraídos de cada resultado recuperado, o mesmo ocorre com os *IDs* correspondentes aos documentos da *query* expandida. Em seguida, é realizada a interseção destes *IDs* e a união de ambos dando maior prioridade para os *IDs* pertencentes à interseção dos resultados. Os documentos recuperados pela *query* expandida aparecem em seguida e, por fim, os resultados recuperados da *query* original. Importante destacar que tanto os resultados da interseção entre as duas *queries* quanto os resultados da *query* expandida serão exibidas à frente dos resultados da *query* original, independente de seu *score* que será respeitado apenas em relação a cada conjunto de resultados. Esta etapa, bem como toda a abordagem realizada, termina com a exibição dos resultados expandidos que possuem seus *IDs*, *links* para cada documento e seu respectivo *score* demonstrados junto ao número total de documentos recuperados.

6 Resultados

Neste trabalho, foi apresentado um algoritmo de expansão de queries, que usa Word Embeddings como suporte para geração de termos expandidos.

Desta forma, as URLs do *EventKG* foram extraídas e processadas a fim de tornarem-se palavras-chave para a formulação da primeira *query*, o Corpus Anotado do New York Times foi indexado e fazendo uso do algoritmo *BM25*, de acordo com as regras estabelecidas para o ranqueamento, documentos foram recuperados e processados a fim de compor termos para uma nova *query*. O que se sucedeu, a partir de então, foi o treinamento *Word2Vec* que, utilizando o *CBOW* como algoritmo de treinamento, possibilitou encontrar os três termos mais similares à *query* original. Estes termos mais similares ao comporem a *query* expandida e, após recuperar mais documentos, teve estes comparados aos documentos recuperados pela *query* original de modo a nos fornecer os resultados finais desta abordagem.

Os resultados desta abordagem foram obtidos com o uso dos eventos fornecidos pelo *EventKG* em que, feita uma seleção aleatória de três eventos de uma lista pré-processada de cerca de mil eventos, cada evento teve suas palavras-chaves utilizadas pela *query* original e expandida. Assim, as palavras-chaves escolhidas a partir das URLs são demonstradas na Tabela 2.

Tabela 2 – Palavras-chave escolhidas

Palavra-chave
Athletics Summer Olympics
FIFA World Championship
Hurricane Disaster

O primeiro resultado em relação às palavras-chaves *Athletics Summer Olympics* apontou um total de 9 documentos recuperados pela *query* original, conforme demonstrado na Figura 4, contendo os *IDs* de cada documento.

Figura 4 – Query original: *athletics summer olympics*

```
Número de Documentos Recuperados: 7
--------------------------------
First Search IDs:
['820360', '782647', '855816', '1697647', '840454', '646336', '425368', '1728700',
 '1693535']
--------------------------------
```

A *query* expandida apresentou entre os termos mais similares *event*, *sport* e *performance* a partir de um total de 108 palavras no corpus após realizado todo o pré-processamento. A Tabela 3 demonstra os 10 termos mais similares à *query* original com suas devidas pontuações.

Tabela 3 – Termos mais similares a *athletics summer olympics*

Palavra	**Pontuação**
event	0.20675334334373474
sport	0.16313736140727997
performance	0.15304459631443024
citizens	0.13788382709026337
people	0.13378882408142090
way	0.11608307063579560
yorkers	0.11484740674495697
edward	0.09938810020685196
raising	0.09316247701644897
director	0.08925282955169678

Uma vez expandida a *query*, foram recuperados 7 documentos, o que é visualizado na Figura 5.

Figura 5 – Query expandida: *athletics summer olympics event sport performance*

```
--------------------------------
Second Search IDs:
['820360', '1251141', '899037', '594179', '1191216', '1192944', '1033742']
```

Dentre estes documentos, foi obtida a interseção de um documento recuperado pelas duas *queries* e, realizada a união, um total de 15 documentos foi recuperado conforme demonstra a Figura 6.

Figura 6 – Interseção e União dos documentos recuperados - *athletics summer olympics*

```
--------------------------------
Intersection IDs:
['820360']
--------------------------------
Union IDs:
['820360', '1251141', '899037', '594179', '1191216', '1192944', '1033742', '782647
', '855816', '1697647', '840454', '646336', '425368', '1728700', '1693535']
```

Os documentos recuperados foram ordenados seguindo a prioridade descrita pela abordagem, em que os *IDs* que pertencem à interseção possuem prioridade, seguidos dos documentos recuperados pela *query* expandida e, por fim, os documentos da *query* original foram exibidos, assim como pode ser observado na Figura 7.

A pontuação dos documentos recuperados variou entre o valor máximo de 33.295822 e mínimo de 20.114471 e teve um total de 339.3543 pontos assim apresentando uma pontuação média de 22.623620000000003.

A precisão pôde ser calculada através da inspeção manual das notícias recuperadas. Desta forma, a relevância dos documentos recuperados foi verificada a partir do conteúdo de cada notícia recuperada e, para os 15 documentos, pode-se apontar que dez deles apresentaram relevância em relação à *query* original e expandida. Os cinco últimos documentos

Figura 7 – Todos os documentos recuperados - *athletics summer olympics*

```
Expanded Search IDs:


ID: 820360 - ['http://query.nytimes.com/gst/fullpage.html?res=950DE0DE163FF933A157
55C0A96F948260', '33.295822']

ID: 1251141 - ['http://query.nytimes.com/gst/fullpage.html?res=9805EED8153FF937A25
75BC0A9629C8B63', '26.183187']

ID: 899037 - ['http://query.nytimes.com/gst/fullpage.html?res=950DEFDD153EF930A257
57C0A96F948260', '26.00016']

ID: 594179 - ['http://query.nytimes.com/gst/fullpage.html?res=9A04E7DB1239F932A057
50C0A960958260', '22.938004']

ID: 1191216 - ['http://query.nytimes.com/gst/fullpage.html?res=9E0DEED91F3AF930A15
754C0A9629C8B63', '21.85191']

ID: 1192944 - ['http://query.nytimes.com/gst/fullpage.html?res=9B07EFDC1E3BF93BA25
754C0A9629C8B63', '21.736082']

ID: 1033742 - ['http://query.nytimes.com/gst/fullpage.html?res=9901E1D91738F930A25
752C0A9639C8B63', '20.525259']

ID: 782647 - ['http://query.nytimes.com/gst/fullpage.html?res=9D0DEFD61F30F935A157
51C0A9669C8B63', '22.26447']

ID: 855816 - ['http://query.nytimes.com/gst/fullpage.html?res=950DEEDE1F38F931A257
50C0A96F948260', '21.48298']

ID: 1697647 - ['http://query.nytimes.com/gst/fullpage.html?res=9D0CEFDE123EF933A15
755C0A967958260', '21.087326']

ID: 840454 - ['http://query.nytimes.com/gst/fullpage.html?res=950DE5DB123DF932A257
52C0A96F948260', '21.047972']

ID: 646336 - ['http://query.nytimes.com/gst/fullpage.html?res=9F0CE2DF1638F935A157
55C0A965958260', '20.454815']

ID: 425368 - ['http://query.nytimes.com/gst/fullpage.html?res=9E0CE1D7123EF934A357
52C1A964958260', '20.253693']

ID: 1728700 - ['http://query.nytimes.com/gst/fullpage.html?res=9D0CE4DA123AF931A25
750C0A967958260', '20.118149']

ID: 1693535 - ['http://query.nytimes.com/gst/fullpage.html?res=9D0CE1D71E3AF932A35
753C1A967958260', '20.114471']

Total de documentos recuperados: 15
```

indicaram até uma relevância em relação às *queries*, mas não são notícias tão relevantes em comparação com as demais. Considerando apenas as notícias mais relevantes da *query* original, a precisão desta fica em 0.5 e, considerando a precisão após efetuada a *query* expandida, esta ficou em aproximadamente 0.667, o que representa um aumento de 0.117 em relação às notícias com maior relevância, embora os demais documentos recuperados também apresentem certa relevância à busca.

O segundo resultado em relação às palavra-chaves *FIFA World Championship* obteve um número grande de documentos na recuperação da *query* original, onde um total

de 42 documentos foram recuperados, conforme disposto na Figura 8, contendo os *IDs* de cada documento.

Figura 8 – Query original: *fifa world championship*

```
------------------------------
First Search IDs:
['1311174', '745305', '1798000', '934046', '1568009', '1536502', '990087', '61123', '927664', '158301
4', '566312', '1441296', '428351', '1301031', '587683', '1732760', '161641', '751667', '174829', '173
2195', '683326', '674485', '1398205', '109532', '1787464', '135715', '112450', '389319', '1581501', '
1721411', '1677402', '950095', '1161260', '369952', '460873', '277097', '1464111', '1156173', '145475
1', '325306', '615661', '200580', '1299426', '1772822', '119766']
```

A *query* expandida apresentou como termos mais similares *soccer*, *tournament* e *american* a partir de um total de 247 palavras no corpus após realizado todo o pré-processamento. A Tabela 4 demonstra os 10 termos mais similares à *query* original com suas devidas pontuações.

Tabela 4 – Termos mais similares a *fifa world championship*

Palavra	**Pontuação**
soccer	0.17344261705875397
tournament	0.11099383980035782
american	0.08225916326045990
record	0.08218261599540710
group	0.07391073554754257
team	0.06809275597333908
coach	0.05665260553359985
germany	0.05371212959289551
match	0.04830790311098099
games	0.03832753002643585

Uma vez expandida a *query*, foram recuperados 14 documentos, como visto na Figura 9.

Figura 9 – Query expandida: *fifa world championship soccer tournament american*

```
--------------------------------
Second Search IDs:
['1311174', '751667', '1299426', '1288102', '1440770', '509373', '1038756', '1357379', '1485791',
'1275572', '247839', '1231175', '1787410', '925326']
```

Dentre estes documentos, foi obtida a interseção de três documentos recuperados pelas duas *queries* e, realizada a união, um total de 53 documentos foram recuperados conforme demonstra a Figura 10.

A pontuação dos quinze primeiros documentos recuperados variou entre o valor máximo de 44.683723 e mínimo de 20.085823, tendo um total de 395.208699 pontos assim apresentando uma pontuação média de 26.3472466.

Em virtude de uma grande quantidade de documentos recuperados pelas duas *queries*, a análise da relevância se limitou a verificar os quinze documentos que possuem as

Figura 10 – Interseção e União dos documentos - *fifa world championship*

```
------------------------------
Intersection IDs:
['1311174', '751667', '1299426']
------------------------------
Union IDs:
['1311174', '751667', '1299426', '1288102', '1440770', '509373', '1038756', '1357379', '1485791', '12
75572', '247839', '1231175', '1787410', '925326', '745305', '1798000', '934046', '1568009', '1536502'
, '990087', '61123', '927664', '1583014', '566312', '1441296', '428351', '1301031', '587683', '173276
0', '161641', '174829', '1732195', '683326', '674485', '1398205', '109532', '1787464', '135715', '112
450', '389319', '1581501', '1721411', '1677402', '950095', '1161260', '369952', '460873', '277097', '
1464111', '1156173', '1454751', '325306', '615661', '200580', '1772822', '119766']
```

maiores pontuações. Assim, foi apontada relevância em relação à *query* original e expandida de 11 entre os 15 quinze documentos do intervalo que possuem relevância. Apesar das notícias não classificadas como relevantes ainda possuírem certo grau de relevância em relação a alguns dos termos da *query* original ou expandida. Em relação à *query* original, a precisão dos primeiros quinze documentos P@15 ficou em 0.5 e considerando a precisão após efetuada a *query* expandida, esta ficou em aproximadamente 0.667, o que representa um aumento de 0.117 em relação às notícias com maior relevância. Importante ressaltar que a relevância maior dos documentos da *query* expandida, somada à abordagem utilizada, em que documentos recuperados que estejam na interseção ou que sejam obtidos pela *query* expandida possuem maior prioridade o que desloca para o fim os documentos com melhor pontuação da *query* original pode ter contribuído para o aumento da precisão neste cenário onde um número grande foi recuperado pela *query* original.

O terceiro resultado, em relação às palavra-chaves *Hurricane Disaster*, obteve um total de 12 documentos recuperados pela *query* original. Os *IDs* de cada documento são demonstrados pela Figura 11.

Figura 11 – Query original: *hurricane disaster*

```
------------------------------
First Search IDs:
['1736412', '730238', '1024863', '1829708', '1782377', '1024730', '800532', '1025969',
 '512296', '1065656', '1122529', '571687']
```

A *query* expandida apresentou como termos mais similares *national*, *vulnerabilities* e *season* a partir de um total de 164 palavras no corpus após realizado todo o pré-processamento. A Tabela 5 abaixo demonstra os 10 termos mais similares à *query* original com suas devidas pontuações.

Uma vez expandida a *query*, foram recuperados 12 documentos, como visto na Figura 12.

Figura 12 – Query expandida: *hurricane disaster*

```
------------------------------
Second Search IDs:
['730238', '206515', '802549', '812149', '1739406', '1216483', '187937', '1218901',
'515072', '1310277', '551230', '1041109']
```

Tabela 5 – Termos mais similares a *hurricane disaster*

Palavra	**Pontuação**
national	0.24695888161659240
vulnerabilities	0.21001456677913666
season	0.19449165463447570
tropical	0.13858586549758910
tornado	0.13093462586402893
government	0.12921904027462006
damage	0.12911722064018250
midwest	0.10930404812097550
states	0.10691251605749130
landscape	0.10538785094022751

Realizada a interseção, um documento recuperado pelas duas *queries* foi identificado e, após a união, um total de 23 documentos foram recuperados conforme demonstra a Figura 13.

Figura 13 – Interseção e União dos documentos - *hurricane disaster*

```
--------------------------------
Intersection IDs:
['730238']
--------------------------------
Union IDs:
['730238', '1736412', '206515', '1829708', '812149', '802549', '187937', '1739406',
'1216483', '1218901', '515072', '1310277', '551230', '1041109', '1024863', '1782377'
, '1024730', '800532', '1025969', '512296', '1065656', '1122529', '571687']
--------------------------------
```

A pontuação dos quinze primeiros documentos recuperados obteve o valor máximo de 45.82444, mínimo de 20.519424 no intervalo de quinze documentos totalizando 392.259365, que fornece uma pontuação média de 26.1506243333333. Estes documentos podem ser visualizados na Figura 14.

A análise da relevância verificada nos quinze documentos melhor classificados aponta que em relação à *query original*, apenas quatro tiveram relevância. A *query expandida* obteve sete documentos relevantes entre os quinze recuperados. As notícias classificadas como irrelevantes foram muito mais genéricas e tendiam a ter certo grau de relevância apenas em relação ao segundo termo da *query*.

Desta forma, a respeito da *query original*, a precisão dos primeiros quinze documentos P@15 ficou em 0.267 aproximado. A precisão da *query* expandida dos quinze documentos melhor classificados P@15 em valor aproximado equivale a 0.467. O que demonstra um aumento de 0.2 em relação às notícias com maior relevância da *query* original. Neste caso, é extremamente indispensável comentar que o maior ganho, além do óbvio resultado na precisão, foi no deslocamento dos resultados existentes na interseção e obtidos pela *query* expandida, pois estes documentos são exatamente os três últimos recuperados pela *query* original. Conclui-se então que a abordagem contribuiu consideravelmente, não

Figura 14 – Documentos Recuperados - *hurricane disaster*

```
Expanded Search IDs:

ID: 730238 - ['http://query.nytimes.com/gst/fullpage.html?res=9C03E2DD113CF937A35755C0A9669C8B63', '45.82444']
ID: 1736412 - ['http://query.nytimes.com/gst/fullpage.html?res=9D0CE4DC163BF930A35752C1A967958260', '32.08874']
ID: 206515 - ['http://query.nytimes.com/gst/fullpage.html?res=9E01E0DF173CF936A2575AC0A96F958260', '28.605143']
ID: 1829708 - ['http://query.nytimes.com/gst/fullpage.html?res=9C00EEDC103EF935A1575BC0A9609C8B63', '27.441147']
ID: 812149 - ['http://query.nytimes.com/gst/fullpage.html?res=950DEFD81F3AF931A15753C1A96F948260', '25.537397']
ID: 802549 - ['http://query.nytimes.com/gst/fullpage.html?res=950DEEDF1E31F934A35754C0A96F948260', '26.00935']
ID: 187937 - ['http://query.nytimes.com/gst/fullpage.html?res=9C05E6DB173EF936A15754C0A96F958260', '25.878017']
ID: 1739406 - ['http://query.nytimes.com/gst/fullpage.html?res=9D0CE7DD1F3FF933A1575BC0A967958260', '24.834631']
ID: 1216483 - ['http://query.nytimes.com/gst/fullpage.html?res=9903EED81739F93BA2575AC0A9629C8B63', '24.345827']
ID: 1218901 - ['http://query.nytimes.com/gst/fullpage.html?res=9903E2DF1F39F936A1575AC0A9629C8B63', '24.291533']
ID: 515072 - ['http://query.nytimes.com/gst/fullpage.html?res=9C0DEFDA103EF93AA35752C1A96E958260', '22.773476']
ID: 1310277 - ['http://query.nytimes.com/gst/fullpage.html?res=9F0DE1DD133AF93AA2575AC0A9659C8B63', '21.28004']
ID: 551230 - ['http://query.nytimes.com/gst/fullpage.html?res=9C00E1DA1E39F937A25754C0A960958260', '21.265661']
ID: 1041109 - ['http://query.nytimes.com/gst/fullpage.html?res=9C06EEDC1E31F931A25751C1A9639C8B63', '20.519424']
ID: 1024863 - ['http://query.nytimes.com/gst/fullpage.html?res=9F01E3D91531F935A3575AC0A9639C8B63', '21.564539']
ID: 1782377 - ['http://query.nytimes.com/gst/fullpage.html?res=9901E5D71730F93BA15755C0A9609C8B63', '21.422602']
ID: 1024730 - ['http://query.nytimes.com/gst/fullpage.html?res=9402E3D91531F935A3575AC0A9639C8B63', '21.411268']
ID: 800532 - ['http://query.nytimes.com/gst/fullpage.html?res=9404E6DA133CF931A35757C0A9669C8B63', '21.186525']
ID: 1025969 - ['http://query.nytimes.com/gst/fullpage.html?res=9B02EFDC1731F932A3575AC0A9639C8B63', '21.001972']
ID: 512296 - ['http://query.nytimes.com/gst/fullpage.html?res=9C07EFDE153EF934A35752C1A96E958260', '20.739653']
ID: 1065656 - ['http://query.nytimes.com/gst/fullpage.html?res=9502E6D61631F932A0575BC0A9639C8B63', '20.493181']
ID: 1122529 - ['http://query.nytimes.com/gst/fullpage.html?res=9C0CE5DF1E31F934A35752C0A966958260', '20.102852']
ID: 571687 - ['http://query.nytimes.com/gst/fullpage.html?res=9401E3DB153AF936A2575AC0A960958260', '20.074745']
Total de documentos recuperados: 23
```

apenas para o aumento da precisão, como para melhor experiência no uso do sistema neste cenário.

Importante ressaltar que, embora seja possível calcular a *precisão* por meio da inspeção manual dos documentos, não há como calcular o *recall* desta abordagem em razão de ser desconhecida a quantidade de documentos relevantes presentes no *The New York Times Annotated Corpus* em relação a cada *query* realizada.

Os resultados assim indicam que todos os documentos recuperados após a expansão da *query* apresentam um número maior de documentos independente de haver ou não a interseção de documentos recuperados. As pontuações das notícias recuperadas também tendem a serem maiores que as pontuações dos documentos recuperados pela *query* original.

Comparando os resultados obtidos com outros parâmetros para melhor avaliar não apenas a precisão num escopo maior de documentos recuperados como em escopo menores,

dado que, normalmente, o interesse se encontra concentrado nos primeiros resultados, a Tabela 6 representa a precisão para cada uma das palavras-chaves utilizadas a respeito da *query* original para os escopos de precisão P@5, P@10 e P@15.

Tabela 6 – Precisão da Query Original

Palavras-chave	P@5	P@10	P@15	Média
Athletics Summer Olympics	1	0.5	0.334	0.612
FIFA World Championship	1	0.7	0.467	0.723
Hurricane Disaster	0.8	0.4	0.267	0.489

Utilizando os mesmos parâmetros, a precisão em P@5, P@10 e P@15 da *query* expandida pode ser constatada na Tabela 7.

Tabela 7 – Precisão da Query Expandida

Palavras-chave	P@5	P@10	P@15	Média
Athletics Summer Olympics	1	**1**	**0.667**	**0.889**
FIFA World Championship	1	**0.9**	**0.734**	**0.878**
Hurricane Disaster	**1**	**0.7**	**0.467**	**0.723**

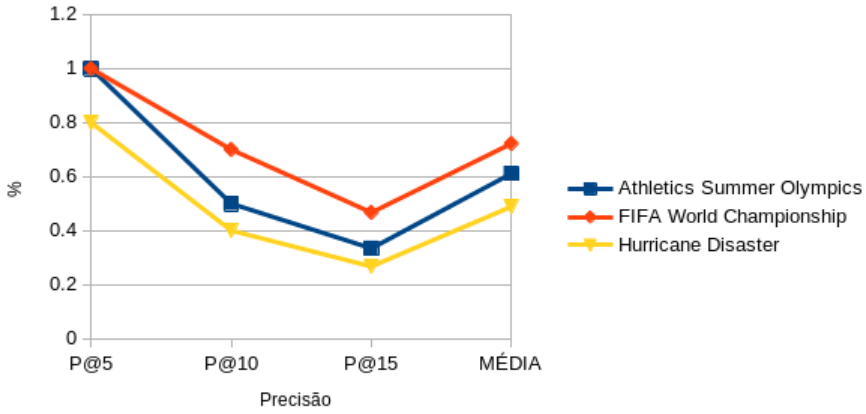
Como se pode observar nos resultados, a abordagem usando a query expandida obtém precisão superior em relação ao BM25, especialmente, em P@10 e P@15 para todas as *queries*, particularmente em P@10 de *Athletics Summer Olympics* onde o aumento foi mais expressivo. Normalmente, foi possível também constatar que os primeiros resultados em todas as *queries*, costuma ter boa acurácia dado que apenas no P@5 de *Hurricane Disaster* não ocorreu empate entre as *queries* normais e expandidas, o que, ainda assim, serve para demonstrar que o reordenamento de resultados feitos no último processo dando prioridade para os resultados expandidos na interseção, ajudou apesar de ambas não terem tido uma boa pontuação quando comparadas com os valores observados nas demais *queries*, mas isso pode ser consequência da especificidade do tópico ou pela quantidade de documentos com este tema no *Corpus*. Em P@15, de modo geral, houveram ganhos, mas há uma redução na relevância dos documentos recuperados. A média geral para cada *query* demonstra os ganhos que de aproximam-se de 0.9 após a expansão enquanto que, sem ela, os resultados tendem a 0.7. Pode-se verificar que a abordagem consegue se manter superior à *baseline* embora seja perceptível uma diminuição esperada na precisão em ambas as abordagens com o aumento do número de documentos. A Tabela 8 detalha os resultados obtidos referente à recuperação dos documentos.

Além disso, certamente há um número maior de documentos recuperados com a expansão da *query* permitindo finalmente incluir entre resultados mais relevantes notícias que tenham no mínimo uma boa similaridade quanto às palavras-chaves da *query* original. O desempenho proporcional da precisão em relação ao número de documentos recuperados quanto à *query* original pode ser observado na Figura 15.

Tabela 8 – Total de Documentos Recuperados pelas *queries*

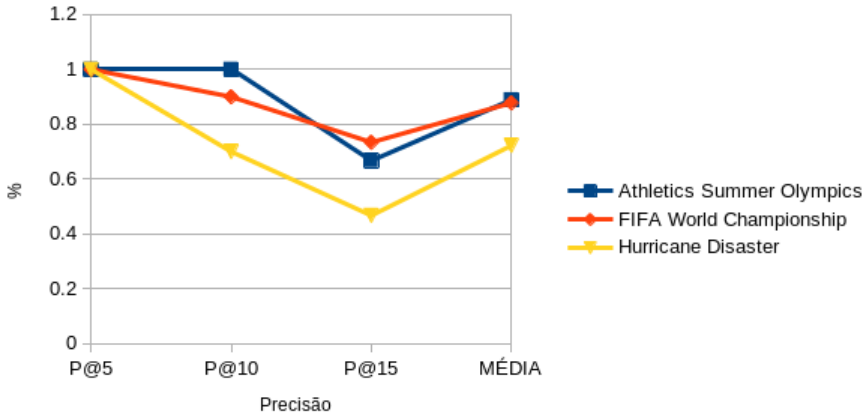
Palavras-chave	Original	Expandida	Interseção	Total
Athletics Summer Olympics	19	17	1	35
FIFA World Championship	42	14	3	53
Hurricane Disaster	12	12	1	23

Figura 15 – Precisão Query Original



Pode-se verificar que a abordagem consegue se manter superior à baseline embora seja perceptível uma diminuição esperada na precisão em ambas as abordagens com o aumento do número de documentos. Este comportamento também é observado após a expansão da *query*, mas é possivel notar que a precisão permanece mais alta em comparação com a precisão da *query* orginal em conformidade com a Figura 16.

Figura 16 – Precisão Query Expandida



7 Conclusão

Neste trabalho, foi apresentado um algoritmo de expansão de *queries*, que usa *Word Embeddings* como suporte para geração de termos expandidos. O algoritmo efetua uma indexação de documentos, fazendo uso do *BM25* e após um pré-processamento numa certa quantidade de eventos extraídas do *EventKG*, palavras-chaves são escolhidas a fim de ser gerada uma *query* que submetida ao sistema de recuperação de informação, retornará notícias que, após uma fase de pré-processamento, servirão no *Treinamento Word2Vec* no intuito de obter os termos mais similares aos termos da *query* original, gerando assim a *query* expandida que será submetida novamente ao sistema de recuperação de informação a fim de obter as notícias mais relevantes a respeito desta nova *query*.

Os resultados desta abordagem mostram que notícias recuperadas pela *query* expandida podem apresentar relevância até maior que as notícias recuperadas pela *query* original apesar de inicialmente poderem até não ter sido recuperadas no primeiro momento e, independente de haver ou não relevância em relação à *query*, observou-se como constante o aumento de documentos recuperados. O que pode ser considerado um fator positivo ao considerar que, mesmo que marginalmente, também houve certo grau de relevância nos resultados recuperados.

7.1 Trabalhos Futuros

Este trabalho apresentou resultados importantes em relação ao aumento de relevância e de documentos recuperados pelo sistema de recuperação de informação construído. Todavia, durante a avaliação de resultados, já na fase final de seu desenvolvimento, foi percebido que poderia ser aumentada a similaridade dos termos para compor a *query* expandida e, consequentemente, aumentada a relevância das notícias recuperadas a partir da expansão do vocabulário que serve como reforço positivo e negativo para o cálculo da similaridade. Isso poderia ser realizado por meio da criação de *profiles* para cada tipo de documento, com aproveitamento da existência de anotações feitas em cada documento - seja por pessoas ou ainda pelos classificadores automáticos utilizados. Como tal, seria possível criar um vocabulário de reforço positivo e até um negativo de acordo com cada tipo de notícia. Como exemplo, uma notícia referente a esportes costuma se utilizar de palavras mais associadas a esportes do que notícias sobre política, pois estas usam um outro conjunto de vocabulário.

A criação de perfis de notícias viria a melhorar significativamente na fase de *Treinamento Word2Vec*, possibilitando o aumento na similaridade entre os termos expandidos e os termos da *query* original. Consequentemente, haveria uma tendência de aumento da

relevância dos resultados recuperados após a expansão da *query*.

Uma nova função de ranqueamento também poderia ser criada após a interseção dos resultados provenientes da *query* original e da *query* expandida, adicionando-se pesos diferentes aos resultados que venham ou não a pertencer à interseção.

Referências

0001, F. D.; MITRA, B.; CRASWELL, N. Query expansion with locally-trained word embeddings. *CoRR*, abs/1605.07891, 2016. Disponível em: <http://arxiv.org/abs/1605.07891>. Citado na página 14.

ALMASRI, M.; BERRUT, C.; CHEVALLET, J.-P. A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. In: FERRO, N. et al. (Ed.). *ECIR*. Springer, 2016. (Lecture Notes in Computer Science, v. 9626), p. 709–715. ISBN 978-3-319-30670-4. Disponível em: <http://dx.doi.org/10.1007/978-3-319-30671-1>. Citado na página 14.

BLANCO, R.; MIKA, P.; BATALLA, J. A. Searching through time in the new york times. In: *Proceedings of HCIR 2010*. [S.l.: s.n.], 2010. Citado na página 17.

DAKKA, W.; GRAVANO, L.; IPEIROTIS, P. G. Answering general time-sensitive queries. *IEEE Trans. Knowl. Data Eng*, v. 24, n. 2, p. 220–235, 2012. Disponível em: <http://dx.doi.org/10.1109/TKDE.2010.187;http://doi.ieeecomputersociety.org/10.1109/TKDE.2010.187>. Citado na página 10.

EFRON, M.; GOLOVCHINSKY, G. Estimation methods for ranking recent information. In: MA, W.-Y. et al. (Ed.). *SIGIR*. ACM, 2011. p. 495–504. ISBN 978-1-4503-0757-4. Disponível em: <http://doi.acm.org/10.1145/2009916>. Citado na página 10.

GOTTSCHALK, S.; DEMIDOVA, E. EventKG+TL: Creating cross-lingual timelines from an event-centric knowledge graph. In: *Proc. of ESWC 2018 Satellite Events*. [S.l.: s.n.], 2018. Citado na página 13.

GOTTSCHALK, S.; DEMIDOVA, E. EventKG - the Hub of Event Knowledge on the Web - and Biographical Timeline Generation. In: . [S.l.]: IOS Press, 2019. v. 10, n. 6, p. 1039–1070. Citado na página 10.

KUZI, S.; SHTOK, A.; KURLAND, O. Query expansion using word embeddings. In: MUKHOPADHYAY, S. et al. (Ed.). *CIKM*. ACM, 2016. p. 1929–1932. ISBN 978-1-4503-4073-1. Disponível em: <http://doi.acm.org/10.1145/2983323>. Citado na página 14.

LEETARU, K.; SCHRODT, P. A. GDELT: Global Data on Events, Location, and Tone, 1979-2012. In: CITESEER. *ISA annual convention*. [S.l.], 2013. v. 2, p. 1–49. Citado na página 13.

LEHMANN, J. et al. DBpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014. Citado na página 10.

LI, X.; CROFT, W. B. Time-based language models. In: *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management (CIKM-03)*. New York: ACM Press, 2003. p. 469–475. Citado na página 10.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. Disponível em: <http://arxiv.org/abs/1301.3781>. Citado na página 20.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013. Disponível em: <http://arxiv.org/abs/1310.4546>. Citado na página 10.

RAO, J.; LIN, J. J. Temporal query expansion using a continuous hidden markov model. In: CARTERETTE, B. et al. (Ed.). *ICTIR*. ACM, 2016. p. 295–298. ISBN 978-1-4503-4497-5. Disponível em: <http://doi.acm.org/10.1145/2970398>. Citado 2 vezes nas páginas 10 e 14.

ROSIN, G.; GUY, I.; RADINSKY, K. Event-driven query expansion. In: . [S.l.: s.n.], 2021. p. 391–399. Citado na página 14.

ROSPOCHER, M. et al. Building Event-Centric Knowledge Graphs from News. *J. Web Sem.*, v. 37-38, p. 132–151, 2016. Citado na página 13.

ROY, D. et al. *Using Word Embeddings for Automatic Query Expansion*. 2016. Comment: 5 pages, 3 tables, 1 figure. Neu-IR '16 SIGIR Workshop on Neural Information Retrieval July 21, 2016, Pisa, Italy. Disponível em: <http://arxiv.org/abs/1606.07608>. Citado na página 10.

ROY, D. et al. Using word embeddings for automatic query expansion. *CoRR*, abs/1606.07608, 2016. Disponível em: <http://arxiv.org/abs/1606.07608>. Citado na página 14.

SUCHANEK, F. M.; KASNECI, G.; WEIKUM, G. Yago: a core of semantic knowledge. In: *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM Press, 2007. p. 697–706. ISBN 978-1-59593-654-7. Disponível em: <http://portal.acm.org/citation.cfm?id=1242572.1242667>. Citado na página 10.

VRANDECIC, D.; KRÖTZSCH, M. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, v. 57, n. 10, p. 78–85, 2014. Citado na página 10.

WANG, Y.; HUANG, H.; FENG, C. Query expansion based on a feedback concept model for microblog retrieval. In: BARRETT, R. et al. (Ed.). *WWW*. ACM, 2017. p. 559–568. ISBN 978-1-4503-4913-0. Disponível em: <http://doi.acm.org/10.1145/3038912>. Citado 2 vezes nas páginas 10 e 14.