CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
CAMPUS TIMÓTEO

Kleverson Roythe Nunes

USO DA CLASSIFICAÇÃO AUTOMÁTICA DE TEXTO NA IDENTIFICAÇÃO DE SIMILARIDADE ENTRE NOTÍCIAS

Timóteo

2021

Kleverson Royther Nunes

USO DA CLASSIFICAÇÃO AUTOMÁTICA DE TEXTO NA IDENTIFICAÇÃO DE SIMILARIDADE ENTRE NOTÍCIAS

Monografia apresentada à Coordenação de Engenharia de Computação do Campus Timóteo do Centro Federal de Educação Tecnológica de Minas Gerais para obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Leonardo Lacerda Alves

Timóteo

2021

Kleverson Roythe Nunes

USO DA CLASSIFICAÇÃO AUTOMÁTICA DE TEXTO NA IDENTIFICAÇÃO DE SIMILARIDADE ENTRE NOTÍCIAS

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Computação do Centro Federal de Educação Tecnológica de Minas Gerais, campus Timóteo, como requisito parcial para obtenção do título de Engenheiro de Computação.

Trabalho aprovado. Timóteo, 17 de Março de 2021.

Prof. Leonardo Lacerda Alves, PhD
Orientador

Assinatura eletrônica em 23/03/2021
Prof. Maurílio Alves Martins da Costa, PhD
Professor Convidado

Assinatura eletrônica em 22/03/2021
Prof. Marcelo de Sousa Balbino, MSc
Professor Convidado

Assinatura eletrônica em 22/03/2021
Prof. Odilon Correa da Silva, MSc
Professor Convidado

Assinaturas eletrônicas:
https://sig.cefetmg.br/public/jsp/documentos/info_documento.jsf?idDocumento=686343

Timóteo
2021

.

.

.

Dedico este trabalho aos sonhadores e inquietos
que nadam contra a corrente para mudar o curso do rio.

Agradecimentos

Primeiramente à minha família por ter me apoiado e torcido por mim em todos esses anos: Meus pais — Cairo e Esmeralda — que batalharam por mim (mais do que eu saberia descrever) desde sempre e sem os quais eu certamente não estaria aqui; ao meu irmão Kayllon, cujos conselhos e experiências ajudaram-me a tomar as decisões que me levaram a este dia.

Agradeço ao meu orientador, Leonardo, pelas conversas e orientações que me ajudaram em muito mais do que este trabalho; ao professor Maurílio por ter me guiado e me aconselhado a tomar boas decisões ao longo do curso e ainda a todo o CEFET-MG Campus Timóteo pelo acolhimento e por tudo que fizeram para que eu pudesse concluir esta graduação a despeito das dificuldades visuais.

Também sou grato aos amigos que fiz nessa caminhada, que contribuíram para meu crescimento e me auxiliaram em diversos momentos de dificuldade; e a todos com quem desempenhei atividades, dentro ou fora do curso, que me ajudaram de alguma forma a chegar aqui.

“Existe um caminho, nenhuma estrada qualquer
entre o amanhecer e o breu da noite
Se você for, ninguém te seguirá
Esse caminho é para seus passos somente.”
Robert Hunter

Resumo

O crescimento no uso problemático de dispositivos está altamente relacionado com o consumo de notícias e traz diversos impactos negativos sobre a vida produtiva e pessoal. Assim, propõe-se um método algorítmico que determina a similaridade de uma nova notícia em relação às presentes no histórico de leitura do usuário. Para isso, o trabalho foi desenvolvido de maneira exploratória e experimental, consistindo em um levantamento bibliográfico sobre as técnicas de sumarização e classificação de texto empregadas, a criação e caracterização de uma coleção de notícias, bem como o desenvolvimento do algoritmo em questão. Cento e quarenta e nove notícias foram coletadas, sumarizadas e classificadas através do cálculo de TF-IDF. Quando utilizado sobre um histórico simulado a partir de trinta das notícias previamente coletadas, o algoritmo identificou elevada coincidência entre sumários de notícias similares para as palavras mais discriminatórias, pouca coincidência para notícias pouco relacionadas e nenhuma coincidência entre notícias sem relação, atribuindo índices de similaridade a cada caso. Com isso, o método algorítmico se mostrou promissor para a identificação da similaridade entre notícias, podendo ser utilizado em trabalhos futuros para averiguar seu impacto sobre o uso problemático de dispositivos.

Palavras-chave: Sumarização de texto. Classificação de texto. TF-IDF.

Abstract

The increase in problematic use of devices is greatly related to news consumption and has several negative impacts on productive and personal life. Thus, an algorithmic method is proposed to determine the similarity between a new publication and what is present in the user's reading history. For such, this work was conducted in an exploratory and experimental way, consisting of a bibliographic survey on the text summarization and classification techniques employed, creation and characterization of a news collection, as well as the development of the algorithm in question. One hundred and forty-nine news items were collected, summarized and classified using the TF-IDF calculation. When used on a simulated history from thirty of the previously collected news, the algorithm identified high coincidence between similar news summaries for the most discriminatory words, low coincidence for unrelated news and no coincidence between unrelated news, attributing similarity indexes to each case. Thus, the algorithmic method has shown promise in identifying the similarity between news, and can be used in future works to investigate its impact on the problematic use of devices.

Keywords: Text summarization. Text classification. TF-IDF.

Lista de ilustrações

Figura 1 – Exemplo de arquivo XML gerado 23
Figura 2 – Planilha de teste inicial 24
Figura 3 – Trecho da planilha de *tags* criada 25
Figura 4 – Gráficos de sumarização manual 25
Figura 5 – Segmento da planilha com os resultados do TF-IDF 26
Figura 6 – Segmento da planilha com as palavras ordenadas pelo TF-IDF 26
Figura 7 – Gráficos de *tags* vs palavras com maior TF-IDF 27
Figura 8 – Gráficos após normalização das *tags* 27
Figura 9 – Representação das comparações de sumários realizadas 28
Figura 10 – Tabela produzida pelo algoritmo 31
Figura 11 – Conteúdo da notícia de nº 116 32
Figura 12 – Resultados dos métodos de cálculo de similaridade 33
Figura 13 – Resultados da soma de frações para onze e sete palavras 35

Sumário

1 INTRODUÇÃO 11
1.1 Contextualização 11
1.2 Justificativa 11
1.3 Problema 12
1.4 Objetivo 12
1.5 Estrutura 12

2 REVISÃO DA LITERATURA 13
2.1 Uso problemático de dispositivos 13
2.1.1 *Fear of Missing Out* e *checking habits* 13
2.1.2 As redes sociais e o consumo de notícias 14
2.1.3 O comportamento de leitura de notícias 15
2.2 Webjornalismo e características das notícias na web 15
2.2.1 Hipertextualidade 16
2.2.2 Multimidialidade 16
2.2.3 Interatividade 16
2.3 Sumarização e classificação de texto 17

3 PROCEDIMENTOS METODOLÓGICOS 19
3.1 Levantamento bibliográfico 19
3.2 Coleção de notícias 20
3.3 Caracterização do corpus 20
3.4 Desenvolvimento do algoritmo 21

4 DESENVOLVIMENTO 22
4.1 Levantamento bibliográfico 22
4.2 Coleção de notícias 22
4.2.1 Definição das fontes 22
4.2.2 Construção e utilização dos *web crawlers* 22
4.2.3 Tratamento dos arquivos XML 23
4.3 Caracterização do corpus 23
4.3.1 Teste inicial do TF-IDF 23
4.3.2 Análise superficial e diversificação do corpus 24
4.3.3 Sumarização das notícias 25
4.4 Comparação sobre notícias reconhecidamente similares 27
4.5 Desenvolvimento do algoritmo 29

5 ANÁLISE DE RESULTADOS 30
5.1 Análise dos sumários 30

5.2 Análise dos vetores 30
5.3 Definição do método 33
5.4 Definição dos intervalos 34

6 CONCLUSÃO 36
6.1 Resultados 36
6.2 Limitações 37
6.3 Trabalhos futuros 38

REFERÊNCIAS 39

ANEXOS 42

ANEXO A – ALGORITMO DESENVOLVIDO 43

ANEXO B – ARQUIVOS DISPONIBILIZADOS 47

1 Introdução

“Pensamentos valem e vivem pela observação exata ou nova, pela reflexão aguda ou profunda; não menos querem a originalidade, a simplicidade e a graça do dizer.”.
Machado de Assis

1.1 Contextualização

O crescente uso de smartphones e computadores ao redor do mundo trouxe consigo discussões sobre o tempo que é despendido neles e o impacto que esse comportamento gera em nosso cotidiano. Boa parte disso é descrita por Oulasvirta et al. (2012) pelo termo “hábitos de checagem” ou “*checking habits*” — sessões breves de navegação no dispositivo, espalhadas pelo dia, para que o usuário se mantenha atualizado. Tal impulso é construído pela necessidade de estar conectado e ciente das novidades e dos assuntos relevantes ao indivíduo (OULASVIRTA et al., 2012).

A maior parte das sessões envolve visitas às redes sociais e leitura de notícias, entre outras aplicações (OULASVIRTA et al., 2012). Todavia, mesmo os intervalos de tempo despendidos nas mídias sociais também estão atrelados à obtenção de notícias (WHITING; WILLIAMS, 2013; MITCHELL et al., 2016). Outro aspecto a se considerar é que a maioria dos leitores de notícias em dispositivos móveis as consome de múltiplas fontes, segundo Chan-Olmsted, Rim e Zerba (2013), Schmidt et al. (2017), o que dificulta aos leitores reconhecerem se já leram essas notícias ou não.

Esses comportamentos trazem implicações negativas sobre a produtividade, com quedas no tempo de trabalho ininterrupto e outras repercussões que impactam não somente o trabalho, como a vida pessoal. Ademais, esses comportamentos são as principais causa de psicopatologias relacionadas ao uso problemático de dispositivos (DUKE; MONTAG, 2017).

1.2 Justificativa

O leitor pode encontrar variações de um mesmo conteúdo escritas por vários jornalistas. Significa que, após ler uma determinada notícia, o leitor pode ser exposto novamente a ela, agora publicada por outro veículo, com título e conteúdo distintos o suficiente para motivar uma nova leitura. Nesse cenário, o leitor despendeu tempo para ler um conteúdo que já era de seu conhecimento.

Este trabalho se motiva na subsequente necessidade de minimizar o tempo de leitura destinado a notícias já conhecidas, que pode contribuir para a redução do tempo das sessões relacionadas ao consumo de notícias e, potencialmente, o combate a diversos problemas, tais como perda de produtividade, hábitos de checagem, ansiedade e outros.

1.3 Problema

O problema que esta pesquisa se destina a responder consiste na seguinte pergunta: É possível determinar um método algorítmico que identifique, com base no histórico de leitura do usuário em questão, a similaridade de uma nova notícia com as previamente lidas?

1.4 Objetivo

Assim, o objetivo-geral deste trabalho consiste em desenvolver um algoritmo que determine, com base no histórico de leitura do usuário em questão, a similaridade de uma nova notícia em relação às previamente lidas. Dessa forma, o indivíduo poderá optar por ler a notícia em questão — se essa for suficientemente distinta da anterior — ou ignorá-la em caso contrário. Especificamente, objetiva-se:

- Investigar as características da notícia como gênero textual que possam ser utilizadas;
- Identificar quais características descrevem o assunto de uma notícia;
- Observar as características que discriminam uma notícia de outra;
- Elaborar um algoritmo para determinar a similaridade entre notícias.

1.5 Estrutura

O capítulo 1 é uma visão introdutória sobre o que será abordado neste trabalho, contextualizando a temática, o problema, descrevendo os objetivos e justificativas.

O capítulo 2 diz respeito ao levantamento bibliográfico realizado, que expõe com mais profundidade o uso problemático de dispositivos, bem como uma caracterização sobre o que diferencia as notícias publicadas na web e as técnicas de sumarização e classificação textual que foram empregadas no desenvolvimento.

No capítulo 3 referente aos procedimentos metodológicos, o tipo de pesquisa é caracterizado quanto aos objetivos e procedimentos técnicos. Além disso, as etapas adotadas ao longo deste trabalho são listadas e descritas.

O desenvolvimento do trabalho, incluindo os gráficos e segmentos das planilhas produzidas, está localizado no capítulo 4. Nesse capítulo, os procedimentos de criação da coleção de notícias, sumarizações, comparações e o desenvolvimento do algoritmo que automatiza esses procedimentos são detalhados.

O capítulo 5 apresenta e discute os resultados obtidos neste trabalho. Ademais, o método utilizado para calcular a similaridade entre notícias é definido, além dos intervalos que a determinam.

Por fim, no capítulo 6, a conclusão deste trabalho é apresentada, juntamente com as limitações existentes e possibilidades de trabalhos futuros.

2 Revisão da literatura

“Se muito vale o já feito, mais vale o que será.
E o que foi feito é preciso conhecer para melhor prosseguir.”.
Milton Nascimento

O levantamento bibliográfico apresentado a seguir foi desenvolvido conforme procedimento descrito no capítulo 3. Assim, este capítulo está organizado da seguinte forma: a seção 2.1 trata do uso problemático de dispositivos e justifica a necessidade da busca por uma solução tecnológica para reconhecer notícias similares; a caracterização das notícias na web como objeto de pesquisa é abordada na seção 2.2; e, por fim, a seção 2.3 contempla as técnicas de sumarização e classificação de texto, utilizadas neste trabalho, que tornam possível o reconhecimento da similaridade entre notícias.

2.1 Uso problemático de dispositivos

O levantamento bibliográfico discorrido nesta seção é embasado na necessidade de compreender o cenário em que este trabalho se insere e no qual ele pode ser utilizado. Além disso, esta pesquisa apresenta informações que podem ser uteis em trabalhos futuros.

O uso problemático de dispositivos parece estar atrelado a psicopatologias que, com base nos artigos estudados neste trabalho, aparentam estar distantes de uma solução que dependa apenas da psicologia. Portanto, a busca por soluções tecnológicas pode contribuir no combate a essas psicopatologias.

2.1.1 *Fear of Missing Out* e *checking habits*

Em primeiro lugar, é necessário visualizar o cenário dos estudos referentes ao uso problemático de dispositivos que justificam o problema abordado neste trabalho. Diversos autores descrevem esse uso através da expressão *Fear of Missing Out* — ou medo de ficar de fora, na tradução para o português.

O termo *fear of missing out* (FoMO) trata de “uma construção de personalidade recente que envolve a relutância em perder informações importantes, incluindo informações sociais. O FoMO resulta na necessidade de permanecer frequentemente conectado às redes sociais” (ELHAI et al., 2016, p.510, tradução nossa). No estudo realizado por esses autores, esse era o principal fator que levava ao uso problemático de smartphones. Esse estudo acrescenta ainda que:

> O uso problemático de smartphones pode interferir em outras atividades agradáveis e interromper as atividades sociais, reduzindo assim a ativação comportamental e subsequentemente aumentando a depressão. (ELHAI et al., 2016, p.514, tradução nossa)

Os resultados obtidos por Gezgin et al. (2017, p.165, tradução nossa), em um estudo sobre professores em treinamento, corroboram essa afirmação: “Considerando-se que os indivíduos têm altos níveis de FoMO em ambientes sociais, isso pode ser visto como um risco para futuros professores que pode ter um impacto negativo sobre os processos de ensino e aprendizagem”.

Oulasvirta et al. (2012) descrevem essa problemática no termo *checking habits* ou *checking behaviors* — hábitos ou comportamentos de checagem, na tradução para o português — que o artigo define como “comportamentos automatizados nos quais o dispositivo é aberto rapidamente para verificar a tela de espera ou o conteúdo de informações em um aplicativo específico”.

Além das consequências relatadas nos estudos sobre FoMO, os hábitos de checagem também têm impacto sobre outros aspectos do cotidiano de seus usuários. Duke e Montag (2017, p.94, tradução nossa) concluem que “o vício em smartphones foi associado a efeitos negativos autorrelatados sobre a produtividade, tanto no local de trabalho quanto na vida diária dos participantes”. Nos resultados apresentados, os mesmos autores comentam:

> Observamos relações moderadas entre o vício em smartphones, medido por [índices de vício em smartphones] e variáveis associadas à natureza distrativa do smartphone. Especificamente, houve uma relação moderada entre [índices mais altos de vício em smartphones] e os participantes que relatavam diminuição da produtividade devido ao tempo gasto no smartphone enquanto trabalhavam, além de pessoas gastando menos minutos trabalhando sem interrupção do smartphone. (DUKE; MONTAG, 2017, p.93, tradução nossa)

Outros estudos, como o realizado por Alt (2015, p.116), relacionam o consumo de mídias sociais causado por esse fenômeno ao consumo de notícias e conteúdo comercial. Segundo Oulasvirta et al. (2012, p.112), os aplicativos associados ao hábito de checagem incluem tela inicial, lista de contatos, email, mídias sociais e notícias. Essa associação acrescenta credibilidade à importância de se analisar o relacionamento entre redes sociais, notícias e os hábitos de checagem motivados pelo FoMO.

2.1.2 As redes sociais e o consumo de notícias

Nesse contexto, mesmo o tempo despendido nas mídias sociais também está intrincado com a leitura de notícias. Oitenta por cento dos respondentes da pesquisa conduzida por Whiting e Williams (2013, p.366) relataram que usam mídias sociais para adquirir informação. Ademais, oitenta e um por cento dos americanos, segundo Mitchell et al. (2016, p.2), dizem obter pelo menos parte das notícias que leem através de sites, aplicativos ou redes sociais.

O consumo de notícias através de plataformas como o Facebook e o Twitter é caracterizado por uma distinta estrutura de comunidades e elevada polarização no que diz respeito às fontes. Schmidt et al. (2017, p.3036, tradução nossa) apontam que "há uma tendência natural dos usuários em restringir suas atividades a um conjunto limitado de páginas. De acordo com nossas descobertas, o consumo de notícias no Facebook é dominado pela exposição seletiva".

O compartilhamento entre contatos nessas redes sociais também é um fator que interfere nos comportamentos de leitura discorridos anteriormente: Dentre os respondentes da pesquisa conduzida por Whiting e Williams (2013, p.367), quarenta por cento citaram o compartilhamento de informações como um tipo de uso em seus cotidianos. Complementar a esse dado, Hermida et al. (2012, p.821, tradução nossa) conclui que "a maioria dos usuários de mídias sociais acredita que seus círculos sociais lhes fornecem uma gama mais ampla de notícias e informações do que se eles confiassem apenas na mídia tradicional".

2.1.3 O comportamento de leitura de notícias

O consumo de notícias em dispositivos móveis, conforme Molyneux (2018, p.645, tradução nossa) é caracterizado pela brevidade, frequência, distração e baixa atenção. No estudo em questão, o autor aponta ainda que "o tempo médio gasto pelas pessoas em notícias no celular antes de mudar para outra coisa foi inferior a 12 minutos".

Além das redes sociais, usuários de smartphones também tendem a ler notícias em aplicativos móveis e sites, conforme dados levantados por Mitchell et al. (2016, p.2) citados na seção anterior. Ademais, Chan-Olmsted, Rim e Zerba (2013, p.140) citam estudos que sugerem a grande quantidade de fontes e plataformas de mídia consumidas por leitores em dispositivos móveis.

Unindo esse dado às afirmações de Schmidt et al. (2017, p.3036), pode-se inferir que, embora haja tal pluralidade de páginas de notícias, a diversidade de perspectivas consumidas pelo leitor é geralmente baixa. Todavia, não foram encontrados estudos que ratifiquem essa inferência, o que constitui importante trabalho futuro.

Esses estudos implicam, por conseguinte, na necessidade de conhecer as características específicas das notícias publicadas na web. Essa caracterização auxiliará no entendimento do objeto de estudo e no desenvolvimento do método algorítmico contido neste trabalho.

2.2 Webjornalismo e características das notícias na web

O desenvolvimento deste trabalho exige um estudo sobre as características que descrevem como as notícias na Web se diferenciam dos demais meios de comunicação. Esta seção, portanto, tem como objetivo adentrar nos conceitos relativos a esse tipo de jornalismo;

Canavilhas (2001, p.64) atribui o termo webjornalismo ao "jornalismo que se pode fazer na web" e outros autores, como Mielniczuk (2003) também utilizam o termo da mesma forma. O webjornalismo possui características capazes de tornar a leitura uma atividade personalizável de acordo com o comportamento do leitor, que passa a participar do fluxo informacional (CANAVILHAS, 2014, p.181).

Rocha, Dantas et al. (2017) discutem o relacionamento entre o webjornalismo e redes sociais, levantando pontos que remetem aos dados discutidos na seção 2.1.2, como a polarização e exposição seletiva dos usuários em relação às fontes. Os autores alertam inclusive sobre a necessidade de selecionar notícias que realmente merecem a atenção do leitor:

> Perante a grande oferta de informações disponível na internet atualmente, e da disponibilidade de menos tempo para leitura diante das tarefas do dia-dia, os usuários passaram a necessitar de novos filtros para indicar o que é realmente importante, e, consequentemente, merece sua atenção. (ROCHA; DANTAS et al., 2017, p.392)

Embora outros autores como Canavilhas (2014) considerem diversos paradigmas que caracterizam as notícias na web, para Deuze (2001), apenas três paradigmas fundamentais são considerados para distinguir esse de outros meios de veiculação: Hipertextualidade, multimidialidade e interatividade.

2.2.1 Hipertextualidade

A hipertextualidade diz respeito ao uso de hipertexto ou links, que conectam uma página a outras. A hipertextualidade designa "um processo de escrita/leitura não-linear e não hierarquizada que permite o acesso ilimitado a outros textos de forma instantânea" (FACHINETTO, 2005, p.247).

Mielniczuk (2003, p.202) define que a hipertextualidade se sobrepõe às demais características como condicionante das notícias na web. A autora acrescenta ainda que "qualquer informação jornalística, que está na web, se inscreve através do hipertexto e é regida pelos seus princípios, características, possibilidades e limitações."

Deuze (2001) elucida que a maioria dos sites de notícias utilizam links internos — que apontam para outras páginas do mesmo site. Essa característica, segundo o autor, leva o leitor "a uma espiral decadente de conteúdo" (DEUZE, 2001, tradução nossa).

2.2.2 Multimidialidade

Para Mielniczuk (2003), no contexto do webjornalismo, "a multimidialidade caracteriza a convergência dos formatos das mídias tradicionais (imagem, texto e som) na narração do fato jornalístico em um mesmo suporte.". Dessa forma, uma mesma notícia pode conter imagens, vídeos e arquivos de áudio que contribuem para a transmissão da mensagem (MIELNICZUK, 2003, p.48).

De forma geral, dois tipos de multimidialidade são definidos por Deuze (2001): convergente, em que a multimídia é vista como a soma de diferentes formatos de mídia que compõem o conteúdo; e divergente, na qual cada modalidade de mídia oferece um caminho diferente para consumir o mesmo conteúdo. Segundo o autor, a maioria dos sites de notícia opta pelo formato convergente.

2.2.3 Interatividade

A interatividade é a característica que dá às notícias na web a possibilidade de se adequar às preferências do leitor, permitindo-lhe escolher quais elementos são exibidos e quais caminhos seguir. Assim a interatividade tem "a capacidade de fazer com que o leitor/usuário sinta-se parte do processo jornalístico" (RIBAS, 2004, p.4).

Este conceito está fortemente atrelado ao uso do hipertexto e pode ser dividido em três tipos de interação: navegacional, feita através de links para outras páginas (próxima página, notícias relacionadas); funcional, relacionando-se diretamente com o conteúdo (como seções de comentários); e adaptativa, em que a página é personalizada para as necessidades do usuário. Tanto interações navegacionais quanto funcionais são amplamente empregadas em notícias na web (DEUZE, 2001).

A seguir, a seção 2.3 diz respeito às técnicas que serão empregadas ao longo deste trabalho para realizar a análise e comparação das notícias.

2.3 Sumarização e classificação de texto

Esta seção detalha as técnicas adotadas neste trabalho para desenvolver o método algorítmico para análise e comparação das notícias.

Sumarização de texto pode ser definida como uma maneira de condensar grandes quantidades de informações de forma concisa pelo processo que seleciona informações importantes e descarta informações redundantes. Esse tipo de processo pode ser caracterizado, segundo sua natureza, pela sumarização extrativa ou abstrativa (GHOLAMREZAZADEH; SALEHI; GHOLAMZADEH, 2009, p.1).

A sumarização extrativa — utilizada neste trabalho — retira saliências do texto-fonte para criar um sumário que não altera os termos originais, enquanto a sumarização abstrativa consiste do entendimento linguístico da interpretação para produzir um resumo generalizado e conciso do texto-fonte, que pode substituir os termos originalmente utilizados (KIYANI; TAS, 2017, p.7889).

No caso da sumarização extrativa, segundo Nenkova e McKeown (2012), uma pontuação é dada para cada palavra do texto-fonte e as palavras com maior pontuação são utilizadas na construção do sumário. Essa pontuação determina a relevância da palavra em questão para o conteúdo do texto-fonte em si.

Dentre os algoritmos que automatizam esse processo, o *Term Frequency-Inverse Document Frequency* (ou TF-IDF) — que pode ser traduzido como frequência do termo-inverso da frequência nos documentos — foi escolhido para este trabalho. Esse método foi previamente utilizado por Oh et al. (2014) para criar recomendações personalizadas de notícias e por Kim e Gil (2019) para classificar trabalhos de pesquisa, se provando eficaz em ambos os trabalhos.

Nenkova e McKeown (2012, p.50) descrevem o algoritmo de TF-IDF através de uma equação. Nela, o índice de TF-IDF para uma palavra w é calculado pela quantidade $c(w)$ de vezes que ela aparece em um documento, multiplicada pelo número de documentos que contém essa palavra ($d(w)$) contidos no conjunto total de D documentos. Ou seja:

$$TF.IDF(w) = c(w) * ln\frac{D}{d(w)}$$

Como a frequência de cada palavra em outros documentos no conjunto é levada em

consideração, o TF-IDF dispensa a construção de uma lista de *stop words* — também conhecidas como palavras vazias, que são palavras comuns da linguagem natural. (NENKOVA; MCKEOWN, 2012, p.49).

Esse algoritmo, embora possua aplicação relacionada à classificação de texto, pode ser utilizado para automatizar o processo de sumarização designando pesos às palavras a partir de sua frequência em um texto e o inverso da sua frequência relativa aos documentos em análise (KIYANI; TAS, 2017, p.50). Sobre TF-IDF, Nenkova e McKeown (2012) afirmam:

> Esse método explora contagens de um corpus prévio, que é uma enorme coleção de documentos, normalmente do mesmo gênero do documento que será sumarizado; o corpus prévio serve como indicativo de quantas vezes uma palavra deve aparecer em um texto arbitrário. (NENKOVA; MCKEOWN, 2012, p.49, tradução nossa)

Entretanto, Ramos et al. (2003, p.51) ressaltam a limitação semântica desse algoritmo, como a incapacidade de identificar sinônimos e mesmo variações de gênero ou número de uma palavra.

Classificação de texto, em contrapartida, diz respeito a um método de aprendizado supervisionado em que um algoritmo é treinado para atribuir classes aos documentos que recebe. Aggarwal e Zhai (2012) definem a técnica de classificação da seguinte forma:

> Tomamos um conjunto de registros de treinamento $D = \{X_1, ..., X_N\}$, tal que cada registro esteja etiquetado com um valor de classe extraído de um conjunto de k valores discretos distintos indexados por $\{1...k\}$. Os dados de treinamento são usados para construir um modelo de classificação, que relaciona os atributos no registro em questão a uma das classes. Para uma dada instância de teste cuja classe é desconhecida, o modelo de treinamento é usado para prever uma classe para essa instância. (AGGARWAL; ZHAI, 2012, p.163, tradução nossa)

Embora não possua o melhor desempenho semântico, o algoritmo TF-IDF também pode ser utilizado nessa etapa, como observado por Zhang, Yoshida e Tang (2011, p.2764), obtendo qualidade estatística superior aos demais métodos estudados no artigo. Outro estudo, conduzido por Dai (2018, p.584, tradução nossa), acrescenta que "o algoritmo TF-IDF é amplamente utilizado na categorização de textos devido às suas vantagens de rapidez, simplicidade e precisão."

3 Procedimentos metodológicos

“Crie espaço — e então coloque algo dentro dele.
Comece, desenvolva e termine com intenção.”.
Chick Corea

Esta pesquisa possui caráter exploratório em seus objetivos. Wazlawick (2010) define a pesquisa exploratória como um método de pesquisa “em que o autor não necessariamente tem uma hipótese ou objetivo definido em mente”, embora não seja totalmente o caso desta pesquisa.

Nos procedimentos técnicos, foi adotado o formato experimental. Wazlawick (2010) define que “a pesquisa experimental caracteriza-se pela manipulação de um aspecto da realidade pelo pesquisador”.

Quanto aos procedimentos metodológicos necessários para o desenvolvimento deste trabalho, os seguintes passos foram adotados:

1. Levantamento bibliográfico referente às técnicas de sumarização e classificação de texto empregadas. Os resultados desse levantamento são apresentados no capítulo 2 referente à revisão de literatura;
2. Criação de coleção de notícias que será utilizada para o teste do algoritmo, envolvendo os passos a seguir:
    - Definição das fontes utilizadas para a coleção;
    - Criação de *web crawlers* para buscar e armazenar as notícias das fontes previamente escolhidas;
3. Caracterização do corpus obtido na etapa anterior, cujos passos seguem abaixo:
    - Sumarização manual das notícias coletadas;
    - Sumarização automática das notícias através do algoritmo TF-IDF;
    - Comparação e análise das classificações realizadas;
4. Desenvolvimento do algoritmo de classificação com base nos resultados obtidos.

3.1 Levantamento bibliográfico

Primeiramente, uma pesquisa sobre sumarização e classificação de texto foi necessária para o desenvolvimento deste trabalho. A necessidade de um levantamento bibliográfico é fundada na busca por técnicas e métodos que tornam possível reconhecer a similaridade entre notícias de maneira algorítmica, que poderiam nem ser possíveis.

3.2 Coleção de notícias

A criação de uma coleção de notícias visa a construção de um para a utilização e análise ao longo deste trabalho. Fixar um corpus, em vez de lidar com coletas de notícias em tempo de execução, permite a progressão dos passos listados neste capítulo — uma vez que os dados são sempre obtidos das mesmas notícias.

Segundo Aluísio e Almeida (2006), um corpus é uma amostra finita de enunciados que caracterizem uma dada linguagem, disponível digitalmente para análise. O corpus criado e utilizado neste trabalho é característico do gênero de linguagem jornalística

Assim, as fontes de notícias devem ser definidas antes da construção do corpus, priorizando os veículos mais acessados do Brasil. O site Alexa Traffic Rank (2020) foi escolhido como métrica sobre quais veículos são mais acessados.

A opção por coletar as notícias apenas dos veículos mais acessados é fundada na proximidade maior com situações de mundo real. Nesse cenário, as notícias dos veículos em questão são mais prováveis de serem acessadas em um comportamento de leitura típico.

Dentre os métodos disponíveis para a criação da coleção, o uso de *web crawlers* apresentou maior praticidade ao coletar notícias automaticamente. Assim, uma nova coleção de notícias pode ser criada em cada execução do algoritmo em questão.

Ademais, o emprego de *feeds* RSS foi uma opção mais prática do ponto de vista da construção dos *web crawlers*, visto que este formato apresenta uma estrutura simplificada de onde os URLs das notícias podem ser obtidos. Outras opções experimentadas incluem o uso de *sitemaps* e a navegação direta na página inicial de cada veículo, entretanto ambos apresentaram complicações para o desenvolvimento do *web crawler*.

3.3 Caracterização do corpus

A caracterização do corpus é necessária para identificar os assuntos das notícias coletadas. Este passo permite ainda reconhecer quais notícias no corpus são similares entre si e é fundamental para a proposta do algoritmo.

A sumarização manual, feita em seguida, permitiu uma classificação prévia das notícias que foi determinante para nortear a análise da sumarização automática por TF-IDF realizada subsequentemente. Com isso, os assuntos das notícias já são conhecidos e se sabe quais notícias possuem maior probabilidade de serem similares.

Já a comparação entre os sumários produzidos possibilita investigar a eficácia da sumarização e classificação por TF-IDF como método para reconhecer a similaridade entre notícias. Para este fim, os sumários de notícias similares de cada assunto reconhecidas previamente foram comparados entre si.

3.4 Desenvolvimento do algoritmo

Neste trabalho, um algoritmo que automatiza os passos anteriores foi necessário para simular a indexação de notícias numa ordem que represente um comportamento típico de leitura de notícias. Este algoritmo executa a sumarização e classificação automática, bem como a comparação das notícias reconhecidamente similares.

Por último, com a análise dos resultados produzidos pelo algoritmo desenvolvido, um método foi definido para calcular a similaridade entre notícias. Esse método possibilitou a utilização de intervalos numéricos que discriminam notícias similares das demais. Ambos os passos estão detalhados no capítulo 5.

4 Desenvolvimento

“Sua mente vai responder a maioria das perguntas
se você aprender a relaxar e esperar pela resposta.”
William S. Burroughs

Este capítulo detalha o desenvolvimento das etapas envolvidas neste trabalho, conforme definido na seção 3, cujos resultados serão analisados no capítulo 5.

4.1 Levantamento bibliográfico

Primeiramente, uma pesquisa foi realizada sobre o uso problemático de dispositivos e comportamento de leitura de notícias, além de seus impactos sobre a saúde e produtividade dos indivíduos. Essa pesquisa visou adentrar na motivação deste trabalho para averiguar a necessidade de uma solução de engenharia que contribua no combate aos problemas estudados.

Outro estudo foi feito sobre técnicas que possibilitem reconhecer que uma notícia é similar a outra. Essa pesquisa implicou a leitura sobre sumarização e classificação de texto e os métodos algorítmicos que automatizam esse processo.

Ambas as pesquisas foram realizadas, de forma exploratória, sobre os assuntos em português e em inglês no Google Acadêmico e na ferramenta de buscas do Mendeley. O levantamento bibliográfico resultante desses estudos está apresentado no capítulo 2.

4.2 Coleção de notícias

4.2.1 Definição das fontes

Para a coleção de notícias, foram escolhidos os três sites de notícias mais acessados do Brasil segundo o Alexa Traffic Rank (2020). Na consulta realizada em 24/9/2020, os veículos Globo, UOL e Metrópoles foram os três primeiros na lista. No caso dos portais Globo e UOL, que possuem subdomínios de notícias e outros serviços, foi decidido realizar a coleta para os sites G1 e Folha de S.Paulo.

Em seguida, foram localizados os *feeds* RSS de cada veículo para a subsequente extração dos URLs das notícias. Uma página de notícia de cada veículo foi consultada manualmente para identificar as classes CSS que continham o título e o corpo da notícia.

4.2.2 Construção e utilização dos *web crawlers*

Após a escolha dos veículos de notícias, a biblioteca Scrapy (versão 2.3.0) foi utilizada para a construção dos *web crawlers* responsáveis por coletar as notícias. Um *web crawler* foi construído para cada veículo. O algoritmo foi configurado para acessar os respectivos *feeds*

RSS, ingressar nos URLs das notícias, encontrar as classes CSS previamente identificadas e recuperar seus conteúdos.

A coleta foi realizada no dia 25/9/2020, totalizando 76 notícias da Folha de S.Paulo, 38 do G1 e 20 do Metrópoles. Foi constatado que a discrepância entre as quantidades de notícias de cada site se deve as configurações realizadas pelos respectivos veículos.

O título e corpo das notícias de cada veículo foram então armazenados em arquivos no formato XML. Um arquivo foi criado para cada portal, para manter distinção entre eles. Além disso, foi armazenada localmente uma cópia do arquivo HTML de cada notícia para futuras inspeções e para refazer a coleta caso ocorressem erros.

4.2.3 Tratamento dos arquivos XML

Para melhorar a eficiência em identificar as tuplas e em processar o arquivo do XML, tanto para a leitura ocular quanto pelo algoritmo de TF-IDF desenvolvido na etapa seguinte, foram removidas as quebras de linha no conteúdo das notícias, de forma que cada uma passou a ocupar apenas uma linha do arquivo. A figura 1 apresenta o arquivo XML que contém as notícias obtidas do site Metrópoles.

A biblioteca utilizada possui a funcionalidade de remover *tags* HTML ao armazenar as notícias nos arquivos XML. Entretanto, ao fazê-lo, o texto contido nos hyperlinks era perdido. Portanto, a coleta foi refeita com essa funcionalidade desabilitada e, em vez disso, foi utilizada a ferramenta “localizar e substituir” do editor Visual Studio Code com a expressão regular `<.*?>` para eliminar códigos HTML dos arquivos, mas mantendo o conteúdo em texto dos hyperlinks.

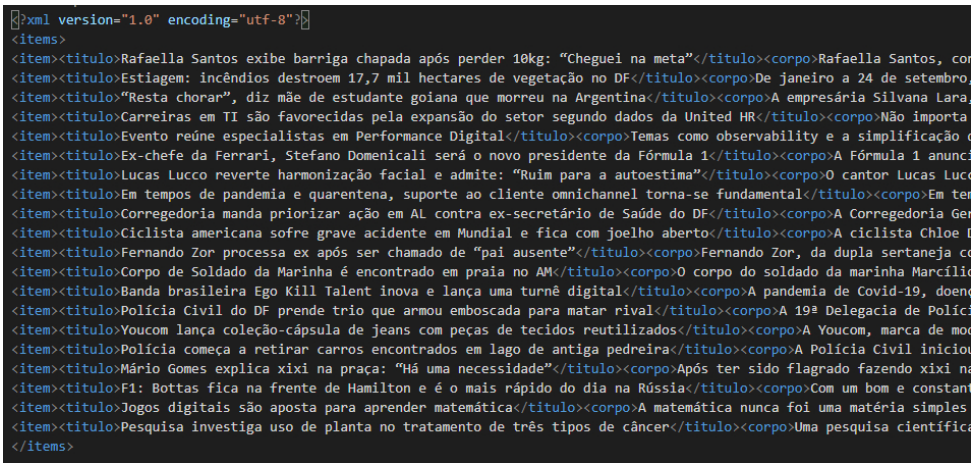

Figura 1 – Exemplo de arquivo XML gerado

4.3 Caracterização do corpus

4.3.1 Teste inicial do TF-IDF

Para proporcionar um ambiente controlado em que os primeiros testes de comparação possam ser realizados com os resultados de TF-IDF, foram coletadas três notícias através do

portal Google Notícias, que agrupa notícias sobre o mesmo assunto publicadas em diversos veículos através da ferramenta Cobertura Completa.

Nesse momento, os títulos e conteúdos das notícias foram copiados a partir do modo de leitura do navegador, coladas em um documento XML formatado aos moldes dos gerados anteriormente pelo Scrapy e fornecidas ao algoritmo de TF-IDF para que os resultados pudessem ser comparados.

Utilizando a ferramenta *TfIdfVectorizer* contida na biblioteca Scikit-Learn, os resultados de TF-IDF para cada notícia do arquivo XML foram gerados e organizados em uma planilha. Dessa forma, foi possível constatar a similaridade entre eles e o funcionamento do algoritmo. A figura 2 apresenta a planilha com os resultados desse teste.

Notícia 1	TF-IDF
causadas	0.407866
ultrapassa	0.407866
marca	0.407866
135	0.310193
mil	0.310193
covid	0.310193
19	0.310193
brasil	0.240892
mortes	0.240892
casos	0.000000
consórcio	0.000000
coronavírus	0.000000
13h	0.000000
diz	0.000000
imprensa	0.000000
atualização	0.000000
passa	0.000000
segundo	0.000000
setembro	0.000000
18	0.000000
veículos	0.000000

Notícia 2	TF-IDF
veículos	0.318554
18	0.318554
atualização	0.318554
casos	0.318554
setembro	0.318554
coronavírus	0.318554
13h	0.318554
segundo	0.318554
consórcio	0.242269
imprensa	0.242269
mortes	0.188143
brasil	0.188143
ultrapassa	0.000000
passa	0.000000
135	0.000000
mil	0.000000
marca	0.000000
diz	0.000000
causadas	0.000000
19	0.000000
covid	0.000000

Notícia 3	TF-IDF
passa	0.402648
diz	0.402648
135	0.306225
consórcio	0.306225
mil	0.306225
imprensa	0.306225
covid	0.306225
19	0.306225
brasil	0.237811
mortes	0.237811
causadas	0.000000
casos	0.000000
coronavírus	0.000000
13h	0.000000
atualização	0.000000
marca	0.000000
18	0.000000
segundo	0.000000
setembro	0.000000
ultrapassa	0.000000
veículos	0.000000

Figura 2 – Planilha de teste inicial

4.3.2 Análise superficial e diversificação do corpus

As notícias obtidas através dos *web crawlers* também foram organizadas em uma planilha para uma análise intelectual da relação entre elas, buscando notícias repetidas entre os três veículos. Com isso, seria possível visualizar que tipo de resultados o algoritmo deste trabalho traria — ou seja: quantas notícias com alta, média e baixa similaridade havia dentro do corpus que poderiam ser identificadas posteriormente pelo algoritmo.

Entretanto, nessa análise, foram encontradas apenas notícias moderadamente similares ou sem relação. Não houve notícias com alta similaridade entre os três veículos (um mesmo acontecimento publicado nos três veículos). As únicas exceções foram algumas notícias parafraseadas e republicadas pela Folha de S. Paulo.

Para obter a diversidade desejada, o Google Notícias foi usado novamente para encontrar notícias reconhecidamente similares. Cinco acontecimentos publicados pelos três veículos escolhidos neste trabalho — totalizando quinze notícias — foram coletadas manualmente na data 5/10/2020 usando os mesmos procedimentos adotados na seção 4.3.1. O corpus, agora com o acréscimo das quinze notícias referidas acima, foi introduzido em uma nova planilha organizada com o veículo e título de cada uma.

4.3.3 Sumarização das notícias

O passo seguinte foi uma sumarização manual, em que foi feita uma leitura instrumental do conteúdo das notícias coletadas para determinar os assuntos de cada uma. Assim, foram atribuídas de três a cinco *tags* que representam o conteúdo das notícias. A figura 3 ilustra as *tags* atribuídas a um subconjunto de notícias.

Fonte	Título	Tags				
Metrópoles	"Resta chorar", diz mãe de estudante goiana que morreu na Argentina	estudante	goiana	morte	buenos aires	argentina
Folha de S. Paulo	1970: Presidente dos EUA afirma que situação se estabilizou na Jordânia	EUA	nixon	guerra	jordânia	
G1	A importância do diagnóstico para o câncer infantojuvenil	câncer	campanha	criança	adolescente	setembro amare
Folha de S. Paulo	A pergunta a ser feita é se dará para jogar o Brasileiro até o fim	brasileiro	futebol	flamengo	palmeiras	covid-19
Folha de S. Paulo	A queda do teto Temer	teto de gastos	bolsonaro	economia	temer	
Folha de S. Paulo	A saúde mental na pandemia	covid-19	saúde	psiquiatria	isolamento	
G1	Adolescente de 16 anos é atingido com quatro tiros em Arinos	adolescente	bicicleta	tiros	arinos	minas gerais
Metrópoles	Análises encontram fungos, ácaro e bactérias em sementes misteriosas	sementes	china	fungos	ácaros	bactérias
Folha de S. Paulo	André Heller-Lopes assina opereta inspirada em Carmen Miranda que será encenada na Estônia	carmen miranda	heller-lopes	opereta	arte	estônia
Folha de S. Paulo	Anitta posa como nova garota carioca em show de Jimmy Fallon	anitta	jimmy fallon	álbum		
Folha de S. Paulo	Antonio Fagundes fala de história e futuro da TV no país no Ao Vivo em Casa	fagundes	globo	entrevista		
Folha de S. Paulo	Ao menos 21 mil candidatos mudaram declaração de cor para eleição de 2020	eleições	racismo	cotas		
G1	Após assalto em comércio, quatro homens são presos dentro de carro com armas e celulares em M	assalto	comércio	manaus	prisão	
Folha de S. Paulo	As vantagens da reeleição	reeleição	FHC	política		
Folha de S. Paulo	Ataque a faca em Paris deixa ao menos 2 feridos perto de antiga sede do Charlie Hebdo	paris	charlie hebdo	facada	feridos	
Folha de S. Paulo	Banco diz que idosa de 85 anos foi 'infantil' ao cair em golpe do motoboy	golpe	idosa	banco	motoboy	
Metrópoles	Banda brasileira Ego Kill Talent inova e lança uma turnê digital	banda	ego kill talent	turnê digital		
Folha de S. Paulo	Bittar diz que apresentação do Renda Cidadã fica para a semana que vem 'se Deus quiser'	renda cidadã	bittar	PEC	auxílio emergencial	
Metrópoles	Bittar: Renda Cidadã "respeitará teto de gastos" e será apresentado na 4ª	renda cidadã	bittar	PEC	auxílio emergencial	

Figura 3 – Trecho da planilha de *tags* criada

Cuidados foram tomados para minimizar flexões de gênero, número e sinônimos entre as *tags* utilizadas. Entretanto, casos como "morte", "mortes" e "falecimento" foram mantidos por possuírem semântica e utilização diferentes nas notícias em que apareciam.

Em seguida, dados estatísticos e gráficos acerca das *tags* atribuídas foram levantados, podendo ser visualizados na figura 4. Isso possibilitou observar a distribuição das incidências de cada *tag*, bem como a quantidade de *tags* ao longo do corpus.

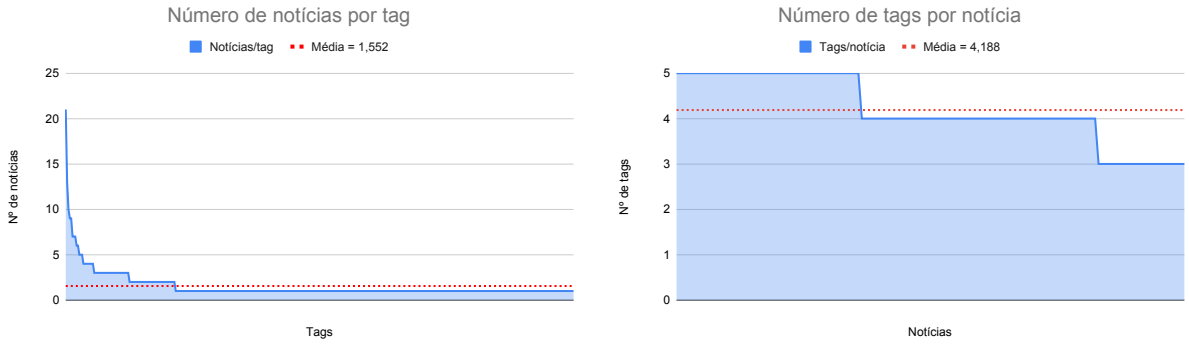

Figura 4 – Gráficos de sumarização manual

A distribuição dos dados em cauda longa do gráfico de notícias por *tag* demonstra a diversidade de assuntos contida no corpus, condizente com o comportamento de leitura de notícias discutido no capítulo 2 referente à Revisão de Literatura.

Utilizando novamente o *TfIdfVectorizer*, foi desenvolvido um algoritmo que armazena, em formato CSV ("*Comma-Separated Values*", ou valores separados por vírgula), o resultado do cálculo de TF-IDF para cada notícia. O arquivo resultante foi então convertido em uma planilha contendo linhas para as notícias e colunas que representavam as diversas palavras

existentes em todo o corpus. As linhas foram ordenadas alfabeticamente por título e as colunas foram ordenadas pelas palavras do corpus. Este resultado para um subconjunto de notícias e palavras é ilustrado na figura 5.

Fonte:	titulo	desses	desta	destaca	destacam	destacar
Metrópoles	“Resta chorar”, diz mãe de estudante goiana que morreu na Argentina	0.0	0.0	0.0	0.0	0.0
Folha de S. Paulo	1970: Presidente dos EUA afirma que situação se estabilizou na Jordânia	0.0	0.0	0.0	0.0	0.0
G1	A importância do diagnóstico para o câncer infantojuvenil	0.0	0.0	0.0	0.0	0.0
Folha de S. Paulo	A pergunta a ser feita é se dará para jogar o Brasileiro até o fim	0.0	0.0	0.0	0.0	0.0
Folha de S. Paulo	A queda do teto Temer	0.0	0.0	0.0	0.0	0.0
Folha de S. Paulo	A saúde mental na pandemia	0.0	0.0	0.0	0.1690158260	0.0
G1	Adolescente de 16 anos é atingido com quatro tiros em Arinos	0.0	0.0	0.0	0.0	0.0
Metrópoles	Análises encontram fungos, ácaro e bactérias em sementes misteriosas	0.0	0.0325679500	0.0	0.0	0.0
Folha de S. Paulo	André Heller-Lopes assina opereta inspirada em Carmen Miranda que será encenada na Estônia	0.0	0.0	0.0	0.0	0.0
Folha de S. Paulo	Anitta posa como nova garota carioca em show de Jimmy Fallon	0.0	0.0	0.0	0.0	0.0
Folha de S. Paulo	Antonio Fagundes fala de história e futuro da TV no país no Ao Vivo em Casa	0.0	0.0575360590	0.0	0.0	0.0
Folha de S. Paulo	Ao menos 21 mil candidatos mudaram declaração de cor para eleição de 2020	0.0	0.0	0.0203895430	0.0	0.0
G1	Após assalto em comércio, quatro homens são presos dentro de carro com armas e celulares em Manau	0.0	0.0332081100	0.0	0.0	0.0
Folha de S. Paulo	As vantagens da reeleição	0.0304482700	0.0	0.0	0.0	0.0
Folha de S. Paulo	Ataque a faca em Paris deixa ao menos 2 feridos perto de antiga sede do Charlie Hebdo	0.0	0.0	0.0	0.0	0.0
Folha de S. Paulo	Banco diz que idosa de 85 anos foi 'infantil' ao cair em golpe do motoboy	0.0	0.0	0.0412393070	0.0	0.0
Metrópoles	Banda brasileira Ego Kill Talent inova e lança uma turnê digital	0.0	0.0374170480	0.0	0.0	0.0
Folha de S. Paulo	Bittar diz que apresentação do Renda Cidadã fica para a semana que vem 'se Deus quiser'	0.0	0.0233047790	0.0	0.0	0.0
Metrópoles	Bittar: Renda Cidadã “respeitará teto de gastos” e será apresentado na 4ª	0.0	0.0	0.0	0.0	0.0
Folha de S. Paulo	Bolsonaro é rejeitado por 46% dos moradores de São Paulo, e Doria, por 39%, diz Datafolha	0.0	0.0	0.0	0.0	0.0
Folha de S. Paulo	Bolsonaro ignorou impacto da Covid-19 por decisão consciente, diz Mandetta	0.0265615620	0.0	0.0	0.0	0.0
Folha de S. Paulo	Bolsonaro nunca aceitou sentar comigo para ver a realidade', afirma Mandetta em livro	0.0569531920	0.0	0.0	0.0	0.0
G1	Bombeiros encerram buscas por banhista desaparecido há mais de 10 dias em MT	0.0	0.0	0.0	0.0	0.0
G1	Brasil tem média móvel de mortes por Covid abaixo de 700 há 2 semanas; casos se aproximam de 5 milh	0.0486541230	0.0339217100	0.0	0.0	0.0
Folha de S. Paulo	Brasil ultrapassa 147 mil mortes pela Covid-19, mostra consórcio de imprensa	0.0	0.0225263720	0.0	0.0	0.0
G1	Briga entre vizinhos deixa um casal morto e dois jovens feridos em Santa Maria, diz polícia	0.0	0.0	0.0	0.0	0.0
G1	Campinas tem 18ª semana sem novas internações em UTI desde o início da pandemia, diz prefeitura	0.0	0.0332743010	0.0	0.0	0.0

Figura 5 – Segmento da planilha com os resultados do TF-IDF

Em seguida, outro algoritmo foi desenvolvido utilizando o *TfIdfVectorizer*, dessa vez armazenando em um arquivo CSV as palavras ordenadas, de forma decrescente, a partir do respectivo cálculo de TF-IDF para cada notícia. Esse arquivo CSV foi também convertido em uma planilha para comparar as *tags* da sumarização manual com os resultados obtidos por esse método. A figura 6 apresenta um segmento da planilha resultante.

Fonte	Título	Feature names						
Metrópoles	“Resta chorar”, diz mãe de estudante goiana que morreu na Argentina	buenos	aires	karol	argentina	setembro	cursava	fernandez
Folha de S. Paul	1970: Presidente dos EUA afirma que situação se estabilizou na Jordânia	jordânia	estabilizou	palestinos	conjunta	pressionaram	guerras	hussein
G1	A importância do diagnóstico para o câncer infantojuvenil	câncer	tumor	células	tumores	crianças	tratamento	diagnóstico
Folha de S. Paul	A pergunta a ser feita é se dará para jogar o Brasileiro até o fim	flamengo	palmeiras	goiás	continente	protocolo	jogar	equador
Folha de S. Paul	A queda do teto Temer	temer	teto	servidores	gasto	desoneração	gerações	futuras
Folha de S. Paul	A saúde mental na pandemia	mental	psiquiatria	saúde	social	destacam	população	mídia
G1	Adolescente de 16 anos é atingido com quatro tiros em Arinos	tiros	adolescente	16	bairro	arinos	crispim	bicicleta
Metrópoles	Análises encontram fungos, ácaro e bactérias em sementes misteriosas	sementes	goulart	pacotes	sanidade	amostras	comerciantes	insumos
Folha de S. Paul	André Heller-Lopes assina opereta inspirada em Carmen Miranda que será encenada na Estônia	opereta	assina	inspirada	heller	estônia	carmen	lopes
Folha de S. Paul	Anitta posa como nova garota carioca em show de Jimmy Fallon	mulher	fallon	ator	nenhuma	filho	jimmy	synchronic
Folha de S. Paul	Antonio Fagundes fala de história e futuro da TV no país no Ao Vivo em Casa	fagundes	tv	teatro	globo	novela	antonio	personagem
Folha de S. Paul	Ao menos 21 mil candidatos mudaram declaração de cor para eleição de 2020	cor	pardo	preto	cota	candidatos	negro	branco
G1	Após assalto em comércio, quatro homens são presos dentro de carro com armas e celulares em	assalto	carro	comércio	armas	suspeitos	manaus	dentro
Folha de S. Paul	As vantagens da reeleição	reeleição	cargos	fhc	legislativos	bons	argumentos	argumento
Folha de S. Paul	Ataque a faca em Paris deixa ao menos 2 feridos perto de antiga sede do Charlie Hebdo	charlie	hebdo	paris	perto	ataque	faca	feridos
Folha de S. Paul	Banco diz que idosa de 85 anos foi 'infantil' ao cair em golpe do motoboy	aposentada	banco	cartão	cartões	85	bradesco	saques
Metrópoles	Banda brasileira Ego Kill Talent inova e lança uma turnê digital	turnê	banda	talent	ego	kill	on	theo
Folha de S. Paul	Bittar diz que apresentação do Renda Cidadã fica para a semana que vem 'se Deus quiser'	bittar	renda	cidadã	emergencial	pec	programa	teto
Metrópoles	Bittar: Renda Cidadã “respeitará teto de gastos” e será apresentado na 4ª	bittar	renda	cidadã	programa	social	teto	guedes
Folha de S. Paul	Bolsonaro é rejeitado por 46% dos moradores de São Paulo, e Doria, por 39%, diz Datafolha	bolsonaro	pesquisa	covas	aprovação	tucano	presidente	doria
Folha de S. Paul	Bolsonaro ignorou impacto da Covid-19 por decisão consciente, diz Mandetta	sr	livro	bolsonaro	falava	escutar	fase	saúde
Folha de S. Paul	Bolsonaro nunca aceitou sentar comigo para ver a realidade', afirma Mandetta em livro	mandetta	ministro	livro	bolsonaro	ex	presidente	braga
G1	Bombeiros encerram buscas por banhista desaparecido há mais de 10 dias em MT	bombeiros	tangará	serra	corpo	rio	buscas	sepotuba
G1	Brasil tem média móvel de mortes por Covid abaixo de 700 há 2 semanas; casos se aproximam	mortes	média	14	móvel	variação	covid	dias
Folha de S. Paul	Brasil ultrapassa 147 mil mortes pela Covid-19, mostra consórcio de imprensa	mortes	habitantes	média	óbitos	100	mil	mortos
G1	Briga entre vizinhos deixa um casal morto e dois jovens feridos em Santa Maria, diz polícia	polícia	vizinhos	casal	feridos	faca	pegou	2ª

Figura 6 – Segmento da planilha com as palavras ordenadas pelo TF-IDF

Ambas as planilhas foram analisadas lado a lado, de forma que cada linha contivesse as *tags* e as palavras ordenadas pelo TF-IDF, ou seja, as mais discriminatórias. A princípio, optou-se por utilizar apenas as onze primeiras colunas geradas pelo algoritmo supracitado, pois esse era o menor número de palavras resultante em uma notícia.

A figura 7 foi gerada sobre o número de palavras que apareceram tanto nas *tags* quanto nas onze palavras mais discriminatórias para cada notícia. O total de palavras coincidentes de cada notícia foi dividido pela quantidade de *tags* designadas a ela. A média de coincidência foi de aproximadamente 42,4%.

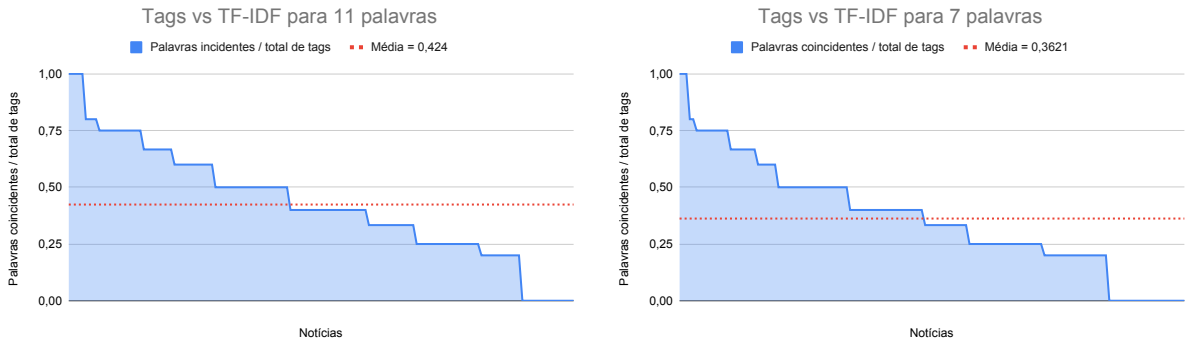

Figura 7 – Gráficos de *tags* vs palavras com maior TF-IDF

Um segundo gráfico foi gerado para fim de teste (também na figura 7), utilizando apenas as sete palavras mais discriminatórias, em vez de onze. O objetivo foi visualizar se uma quantidade menor de palavras afetaria a média, sob a hipótese de que esse pode ser um número suficiente para determinar a similaridade entre as notícias. Nesse caso, a média caiu de 42,4% para 36,2%.

Entretanto, observou-se que as *tags* compostas por mais de uma palavra produziam impacto negativo sobre as estatísticas de coincidência e uma nova planilha foi gerada. Nela, as *tags* compostas foram divididas em novas *tags* como cada palavra utilizada. Com isso, a média de *tags* por notícia subiu de 4,188 *tags* para 5,020.

Essa planilha de comparação normalizada foi então submetida ao mesmo procedimento previamente adotado para gerar novamente os gráficos de coincidência para onze e sete palavras com maior TF-IDF, que estão apresentados na figura 8. Após a normalização, a média subiu de 42,4% para 51,2%, na comparação com onze palavras do TF-IDF, e de 36,2% para 43,4% com sete palavras.

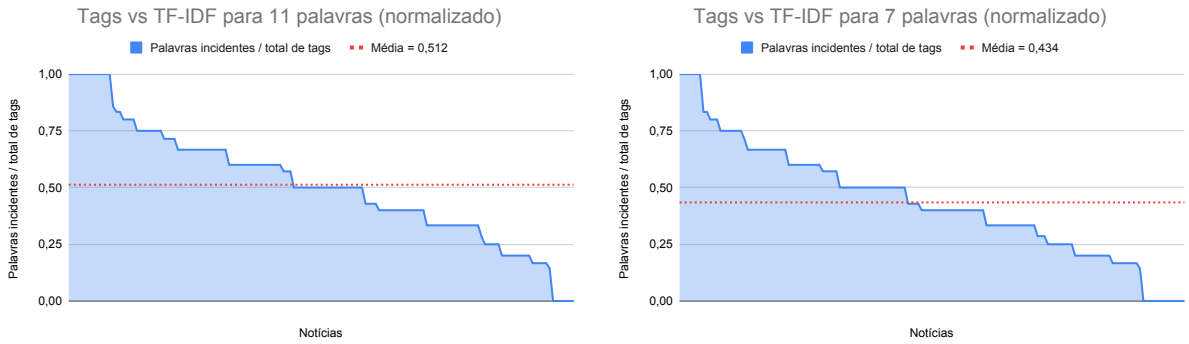

Figura 8 – Gráficos após normalização das *tags*

4.4 Comparação sobre notícias reconhecidamente similares

Com as comparações entre a sumarização manual e a sumarização produzida pelo TF-IDF concluidas, uma nova análise foi realizada apenas entre as notícias reconhecidamente similares coletadas na seção 4.3.2. Esta análise teve o objetivo de investigar as coincidên-

cias presentes entre os dois sumários e observar o que acontece quando essas notícias são comparadas entre si.

Nessa análise, observou-se coincidências de 33% a 100% entre a sumarização manual e as palavras mais discriminatórias para as notícias em questão. Além disso, houve elevada concentração de coincidências entre as primeiras palavras da sumarização manual.

Depois disso, as notícias reconhecidamente similares foram agrupadas de acordo com o assunto retratado — nesse caso, cinco assuntos com três notícias cada. Assim, as sumarizações por TF-IDF de cada assunto puderam ser comparadas entre si de forma a simular o comportamento do algoritmo diante de uma jornada de leitura que envolva notícias similares.

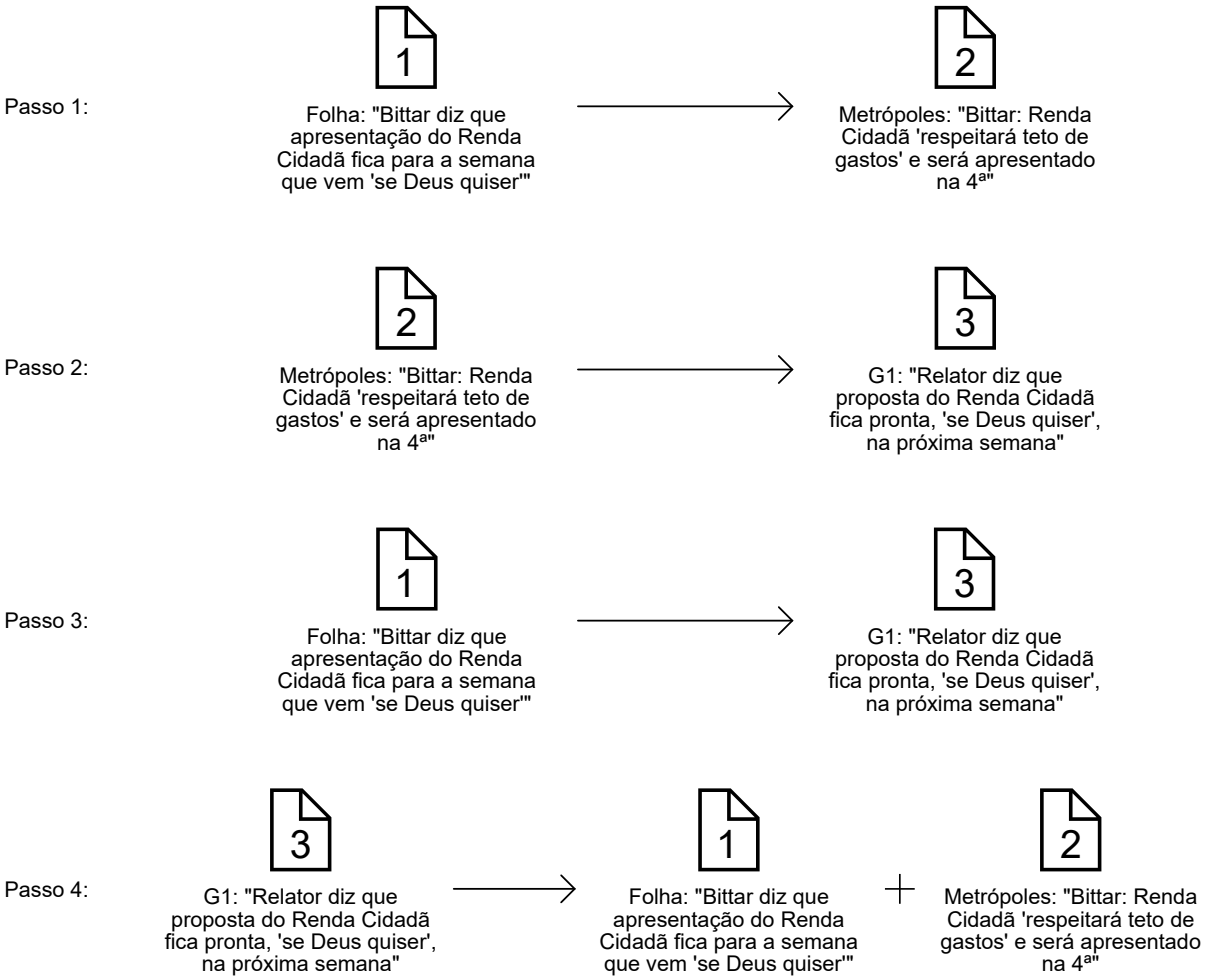

Figura 9 – Representação das comparações de sumários realizadas

Dessa forma, as sumarizações por TF-IDF feitas sobre cada notícia foram comparadas entre si conforme figura 9. Para cada uma dessas comparações, buscava-se o número de vezes em que cada palavra de um sumário apareciam também no outro. Esse procedimento foi realizado para cada grupo de notícias. No passo 4, em que há a soma de dois sumários, todas as repetições de palavras foram mantidas.

Nessa análise, números de incidência maiores foram observados entre as primeiras palavras com maior TF-IDF, em particular nas sete primeiras palavras — corroborando com a

hipótese levantada na seção 4.3.3 de que este número é suficiente para determinar a similaridade entre notícias.

4.5 Desenvolvimento do algoritmo

O método algorítmico desenvolvido nesta seção utilizou a linguagem de programação *Python* (versão 3.8.3 64-bit) no ambiente de trabalho *Microsoft Visual Studio Code* (versão 1.52.1) e está disponível para leitura no anexo A.

As notícias armazenadas nos arquivos XML foram inseridas em um vetor de forma que cada notícia possa ser acessada através de um índice. Esse vetor ficou organizado da seguinte forma: Nas primeiras posições, as notícias do Folha de S. Paulo, G1 e Metrópoles — nessa ordem — foram inseridas, deixando as notícias obtidas através do Google Notícias nas últimas posições para facilitar o reconhecimento posteriormente.

Uma função *adicionar(indice)* foi criada para representar o ato de leitura de uma notícia. Ao inserir um número ao parâmetro *indice*, a notícia correspondente é adicionada ao vetor *historico*, enquanto a data e hora de leitura são adicionadas ao vetor *timestamps* para permitir a exclusão de notícias lidas há mais de uma semana.

Embora a ordenação por palavras mais discriminatórias, com o tempo, desloque as *stop words* para o final (por serem pouco discriminatórias), os cálculos de TF-IDF nesse algoritmo foram realizados descartando antecipadamente as *stop words* da lista publicada por Lopes (2013) com o objetivo de reduzir o número de palavras processadas.

O procedimento desenvolvido na seção 4.4 foi mimetizado na criação da função *comparar(indice)*. Essa função representa um cenário em que o usuário está prestes a abrir uma notícia e é onde o algoritmo determina a similaridade dessa notícia às previamente lidas.

O índice da notícia que se pretende ler é fornecido à função, que verifica se há notícias lidas há mais de uma semana, as exclui do histórico e, em seguida, gera um corpus temporário contendo as notícias do histórico e a notícia a ser comparada. Em seguida, o TF-IDF de cada notícia para esse novo corpus é calculado e o resultado é transformado em um sumário de onze palavras para cada notícia — através da ordenação por palavras mais discriminatórias.

A partir desse resultado, a função verifica quais notícias possuem pelo menos duas palavras em seus sumários que aparecem também no sumário da notícia que se pretende ler. Essas notícias são agrupadas e um vetor é gerado contendo, em suas posições, quantas vezes cada palavra do sumário da notícia que se pretende ler aparece nesse subconjunto.

Um histórico simulado foi criado para representar um possível comportamento de leitura. Trinta notícias foram escolhidas a partir do corpus obtido na seção 4.3.2 e fornecidas ao algoritmo. Cada notícia nesse histórico foi submetida à função *comparar* e, em seguida, à função *adicionar*. Os vetores resultantes de cada notícia, juntamente com o índice e título, foram então armazenados em um arquivo CSV para inspeção e definição do método que determina a similaridade entre as notícias.

5 Análise de resultados

“Nossas cabeças são redondas para que o pensamento possa mudar de direção.”
Allen Ginsberg

Este capítulo aborda os resultados obtidos pelo algoritmo desenvolvido e documentado na seção 4.5, bem como discussões acerca da escolha do método que converte esses resultados em um número que determina a similaridade entre notícias.

5.1 Análise dos sumários

Esta seção detalha os resultados observados na tabela criada a partir do histórico simulado da seção 4.5, que simula a indexação de notícias em uma ordem que representa um possível comportamento de leitura de um usuário. Dois sumários foram produzidos: o sumário inicial, realizado no ato da comparação (ou seja, realizado apenas sobre o corpus desse histórico naquele momento); e o sumário final, feito após a inserção das trinta notícias no histórico.

Os sumários em questão, na maioria dos casos, foram representativos dos conteúdos das notícias adicionadas. Por exemplo, o sumário da notícia de nº 76 presente na figura 10 permite inferir que se trata do diagnóstico e tratamento do câncer em crianças e adolescentes. Uma exceção, contudo, acontece no sumário da notícia de nº 116 — ilustrada na figura 11 — que não permite uma inferência condizente com o conteúdo da notícia.

Ademais, ao comparar os sumários iniciais e finais, nota-se que comumente são alteradas as posições das palavras nas primeiras notícias lidas. Além disso, a presença de *stop words* não contempladas na lista publicada por Lopes (2013) denota a dificuldade da sumarização por TF-IDF em identificar palavras discriminatórias em um corpus pequeno. Essa dificuldade se deve ao fato de *stop words* serem consideradas discriminatórias nesse primeiro momento. Conforme o corpus cresce, essa variação diminui e, com ela, também diminui a ocorrência de *stop words* nos sumários.

A tabela produzida pelo algoritmo contém os vetores resultantes do processo de comparação de notícias e duas colunas referentes aos sumários conforme figura 10. Os dois sumários são ordenados da primeira até a décima primeira palavra mais discriminatória segundo seu resultado de TF-IDF. Isso significa que essas são as palavras que discriminam a notícia em questão das demais contidas no corpus. Ou seja, a última palavra do sumário não é, necessariamente, a menos discriminatória.

5.2 Análise dos vetores

Quanto aos vetores, também apresentados na figura 10, observa-se elevada quantidade de elementos zerados devido à criação de subconjuntos para notícias do mesmo assunto

#	Título	Vetor	Sumário no momento da inclusão	Sumário no final
76	A importância do diagnóstico para o câncer int	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['câncer' 'tumor' 'células' 'crianças' 'tratamento' 'causa' 'infantil"diagnóstico' 'tumores' 'ainda' 'adolescentes']	['câncer' 'tumor' 'células' 'crianças' 'infantil' 'tumores' 'tratamento"diagnóstico' 'causa' 'adolescentes' 'sinais']
146	Bittar: Renda Cidadã “respeitará teto de gasto	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['bittar' 'renda' 'social' 'programa' 'teto' 'recursos' 'proposta"cidadã' 'novo' 'guedes' '10']	['bittar' 'renda' 'programa' 'teto' 'recursos' 'cidadã' 'social"proposta' 'guedes' 'última' 'agenda']
61	Dona do Magalu abriu vaga para negros e des	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['rebouças' 'governo' 'abolicionista' 'enquanto' 'magalu' 'família' 'nome"negros' 'presidencial' 'cotegipe' 'cargos']	['rebouças' 'abolicionista' 'magalu' 'presidencial' 'negros' 'governo"nome' 'barão' 'liberal' 'bolsonaristas' 'conservador']
148	Relator diz que proposta do Renda Cidadã fica	[1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0]	['cidadã' 'teto' 'gastos' 'governo' 'renda' 'dinheiro' 'bittar' 'semana"precatórios' 'proposta' 'duas']	['cidadã' 'renda' 'teto' 'gastos' 'bittar' 'governo' 'precatórios"proposta' 'dinheiro' 'orçamento' 'despesas']
116	“Resta chorar”, diz mãe de estudante goiana c	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['aires' 'karol' 'argentina' 'buenos' 'ano' 'anos' 'ainda' 'setembro' '22"moradora' 'mãe']	['aires' 'argentina' 'buenos' 'karol' 'setembro' 'ano' 'anos' 'ainda"moradora' 'fernandez' 'goiana']
133	Pesquisa investiga uso de planta no tratament	[0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0]	['pesquisa' 'câncer' 'espécie' 'três' 'amazonas' 'planta' 'uso"tratamento' 'tipos' 'investiga' 'investigando']	['pesquisa' 'câncer' 'planta' 'tipos' 'espécie' 'amazonas' 'três"tratamento' 'uso' 'substâncias' 'preliminar']
1	Bolsonaro nunca aceitou sentar comigo para v	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['mandetta' 'livro' 'ministro' 'ex' 'bolsonaro' 'presidente' 'caso"covid' 'mil' 'saúde' 'afirma']	['mandetta' 'ministro' 'livro' 'ex' 'bolsonaro' 'presidente' 'mil' 'braga"moro' 'caso' 'saúde']
19	Mandetta conta bastidores do Ministério da Sa	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['episódio' 'manhã' 'podcast' 'durante' 'publicado' 'café' 'ouça"bastidores' 'sexta' 'mandetta' 'pandemia']	['episódio' 'ouça' 'podcast' 'café' 'publicado' 'mandetta' 'bastidores"manhã' 'sexta' 'durante' 'pandemia']
136	Morre Eddie Van Halen, um dos maiores guita	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['halen' 'van' 'eddie' 'banda' 'rock' 'roth' 'wolfgang' 'filho' 'anos"câncer' 'grupo']	['van' 'eddie' 'halen' 'banda' 'roth' 'rock' 'wolfgang' 'câncer' 'filho"anos' 'lee']
29	Santos vence no Equador e deixa vaga encam	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['equatoriano' 'santos' 'time' 'dez' 'marinho' 'delfín' 'libertadores"carlos' 'partida' 'placar' 'vaga']	['santos' 'equatoriano' 'time' 'placar' 'delfín' 'libertadores' 'partida"vaga' 'marinho' 'minutos' 'garantiu']
38	No epicentro da pandemia no México, vírus aç	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['pessoas' 'vírus' 'méxico' 'mercado' 'arriaga' 'autoridades' 'disse"aquino' 'cidade' 'trabalhadores' 'mundo']	['méxico' 'vírus' 'mercado' 'arriaga' 'pessoas' 'aquino' 'trabalhadores"morreram' 'iztapalapa' 'autoridades' 'disse']
23	Largada eleitoral - A campanha municipal está	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['pesquisas' 'sinal' 'informativas' 'eleitor' 'pleito' 'ocorre' 'disputa"bolsonaro' 'pouco' 'importante' 'campanha']	['pesquisas' 'pleito' 'ocorre' 'informativas' 'disputa' 'eleitor' 'sinal"bolsonaro' 'pouco' 'importante' 'campanha']
90	Radioamador André Sampaio morre aos 70 ar	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['radioamador' 'sampaio' 'noronha' 'ilha' 'andré' 'fernando' 'amigo"marenga' 'radioamadorismo' 'comunicações' '70']	['radioamador' 'sampaio' 'noronha' 'ilha' 'andré' 'fernando' 'amigo"radioamadorismo' 'marenga' 'comunicações' '70']
137	Conselho de Educação permite aula remota a	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['ensino' '2021' 'presencial' 'cne' 'educação' 'aulas' 'atividades"superior' 'ano' 'presenciais' 'texto']	['ensino' '2021' 'presencial' 'cne' 'educação' 'atividades' 'superior"aulas' 'presenciais' 'ano' 'prazo']
83	Homem é preso suspeito de revender carga d	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['produtos' 'abreu' 'casimiro' 'carga' 'polícia' 'homem' 'farmacêuticos"preso' 'policiais' 'revender' 'região']	['produtos' 'casimiro' 'abreu' 'carga' 'farmacêuticos' 'homem' 'polícia"revender' 'preso' 'suspeito' 'civil']
36	Novatos na berlinda - Até a semana passada,	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['moisés' 'política' 'witzel' 'processo' 'assembleia' 'acusação"defender' 'novatos' 'vice' 'legislativa' 'ambos']	['moisés' 'witzel' 'política' 'processo' 'assembleia' 'legislativa"novatos' 'juiz' 'acusação' 'vice' 'defender']
101	Motorista que dirigia a 180 km/h vira réu na Ju	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['motorista' 'réu' 'trânsito' 'racha' 'df' 'homicídio' 'km' 'acidente"tentativa' 'vira' 'josé']	['motorista' 'réu' 'racha' 'trânsito' 'df' 'km' 'acidente' 'homicídio"tentativa' 'vira' 'josé']
84	Homem é preso por suspeita de estuprar uma	[0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0]	['criança' 'suspeito' 'estuprar' 'curuá' 'homem' 'delegada' 'delegacia"suspeita' 'então' 'município' 'prisão']	['criança' 'suspeito' 'estuprar' 'curuá' 'homem' 'delegada' 'delegacia"preso' 'prisão' 'município' 'suspeita']
2	Bolsonaro ignorou impacto da Covid-19 por de	[1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0]	['livro' 'sr' 'saúde' 'bolsonaro' 'fase' 'doença' 'escutar' 'falava"decisão' 'cenário' 'números']	['livro' 'sr' 'bolsonaro' 'saúde' 'fase' 'escutar' 'falava' 'doença"entorno' 'decisão' 'cenário']
30	Próxima crise financeira pode chegar em brev	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['bancos' 'crise' 'unidos' 'banco' 'financeira' 'estados' 'financeiro"oxford' 'reinhart' 'economics' 'podem']	['bancos' 'crise' 'banco' 'unidos' 'financeira' 'estados' 'financeiro"reinhart' 'oxford' 'economics' 'podem']
49	Gente detergente - Eu fico aqui meio pasma c	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['profissão' 'detergente' 'emprego' 'curador' 'escreve' 'geração' 'porra"aproveitar' 'foda' 'oportunidade' 'criar']	['profissão' 'detergente' 'emprego' 'velha' 'porra' 'escreve' 'geração"foda' 'fico' 'curador' 'oportunidade']
57	EUA tem nova noite de protestos contra o raci	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['breonna' 'policiais' 'apartamento' 'polícia' 'agentes' 'manifestantes"louisville' 'noite' 'namorado' 'protestos' 'porta']	['breonna' 'policiais' 'apartamento' 'polícia' 'agentes' 'manifestantes"louisville' 'noite' 'namorado' 'porta' 'disse']
5	Ataque a faca em Paris deixa ao menos 2 feri	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['antiga' 'charlie' 'faca' 'hebdo' 'paris' 'perto' 'ataque' 'feridos"sede' 'deixa' 'menos']	['antiga' 'charlie' 'faca' 'hebdo' 'paris' 'perto' 'ataque' 'sede"feridos' 'deixa' 'menos']
147	Bittar diz que apresentação do Renda Cidadã	[2, 2, 0, 2, 0, 1, 0, 2, 0, 0, 1]	['bittar' 'renda' 'emergencial' 'cidadã' 'pec' 'programa' 'vem' 'teto"diretor' 'ifi' 'gastos']	['bittar' 'renda' 'emergencial' 'cidadã' 'pec' 'programa' 'vem' 'teto"gastos' 'ifi' 'diretor']
11	A saúde mental na pandemia - Os problemas	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['mental' 'saúde' 'psiquiatria' 'população' 'social' 'destacam' 'mídia"revista' 'instituto' 'pandemia' 'médicos']	['mental' 'saúde' 'psiquiatria' 'social' 'população' 'mídia' 'destacam"instituto' 'revista' 'isolamento' 'pandemia']
135	Guitarrista Eddie Van Halen morre aos 65 ano	[1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0]	['halen' 'van' 'eddie' 'tempos' 'câncer' 'guitarristas' 'banda"guitarrista' 'garganta' 'contra' 'batalha']	['van' 'halen' 'eddie' 'tempos' 'câncer' 'banda' 'garganta' 'guitarristas"guitarrista' 'maiores' 'batalha']
108	Campinas tem 1ª semana sem novas internaç	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['campinas' 'coronavírus' 'uti' 'internações' 'carmino' 'leitos' 'souza"semana' 'jonas' 'secretário' 'vivo']	['campinas' 'coronavírus' 'uti' 'leitos' 'internações' 'carmino' 'souza"semana' 'permanência' 'positivos' 'jonas']
18	Menos mortes - O número de pessoas mortas	[1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['polícia' 'agentes' 'policial' 'mortes' 'aumento' 'ano' 'letalidade"confronto' 'casos' 'agosto' 'redução']	['polícia' 'agentes' 'policial' 'mortes' 'ano' 'aumento' 'letalidade"confronto' 'casos' 'agosto' 'redução']
37	Nós humanos às vezes somos histéricos - No	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['transtorno' 'funcional' 'algo' 'histeria' 'termo' 'sintoma' 'histéricos"estar' 'faz' 'sintomas' 'vezes']	['transtorno' 'funcional' 'algo' 'histeria' 'termo' 'sintoma' 'histéricos"estar' 'faz' 'sintomas' 'vezes']
62	Diversidade em Hollywood é tema de reunião	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]	['diversidade' 'cultura' 'audiência' 'hollywood' 'cancelamento"congresso' 'papéis' 'filmes' 'americano' 'pantera' 'forma']	['diversidade' 'cultura' 'audiência' 'hollywood' 'cancelamento"congresso' 'papéis' 'filmes' 'americano' 'pantera' 'forma']

Figura 10 – Tabela produzida pelo algoritmo

METRÓPOLES

BRASIL

“Resta chorar”, diz mãe de estudante goiana que morreu na Argentina

Corpo de Ana Karolina, que cursava medicina em Buenos Aires, ainda passa pelos trâmites necessários para chegar em solo brasileiro

MAIS GOIÁS
25/09/2020 12:04,ATUALIZADO 25/09/2020 12:06

A empresária Silvana Lara, moradora do município goiano de Chapadão do Céu ainda encontra dificuldade para falar sobre a filha, Ana Karolina Fernandez, de 22 anos. Isso porque a jovem, que cursava o terceiro ano de medicina na Universidade de Buenos Aires (UBA), na Argentina, veio a óbito tragicamente no início de setembro, ao cair no poço de um elevador do prédio de amigos.

O corpo de Karol, ainda passa pelos trâmites necessários e só deve chegar em solo brasileiro no fim da próxima semana. O fato ocorreu no dia 4 de setembro deste ano. Karol, que morava há 4 anos em Buenos Aires, havia saído para uma celebração de uma boa nota conquistada no curso

Leia mais em **Mais Goiás**, parceiro do **Metrópoles**.

Figura 11 – Conteúdo da notícia de nº 116

pelo algoritmo. Como a maioria das notícias no histórico simulado não possui relação entre si, um subconjunto de assunto não pode ser criado.

Além disso, chamam a atenção notícias tais como as de nº 83 e 84 que, apesar de possuírem o mesmo assunto sob o critério definido no algoritmo (ou seja, pelo menos duas palavras devem aparecer em ambos os sumários), não possuem sumários similares e, portanto, as notícias também não são similares. Essa conclusão é corroborada pela inspeção do conteúdo completo da notícia, em que se percebe a pequena relação entre as duas notícias (ambas relatando prisões de homens suspeitos).

Esse ponto também pode ser observado nas notícias reconhecidamente similares (coletadas na seção 4.3.2) que constam nesse histórico: na figura 10, há três notícias sobre a proposta da Renda Cidadã (nº 146, 147 e 148) e duas sobre a morte do guitarrista Eddie Van Halen (nº 135 e 136). Essas notícias, além de apresentarem vetores não-nulos, possuem elevada concentração de coincidências nas primeiras posições.

Entretanto, também houve uma situação em que notícias que não são similares (especificamente as notícias de nº 57 e 18) apresentaram coincidência nas primeiras posições. A diferença deste caso para os exemplos citados no parágrafo anterior é que, embora haja essa característica em comum, a primeira possui coincidência somente em duas palavras. Essa observação é relevante para a escolha dos intervalos na seção 5.4.

#	Título	1	2	3	4	5	6	7	8	9	10	11	Pesos lineares	Soma de frações	Pesos exp.	Pesos log.
76	A importância do diagnóstico para o câncer infantojuvenil -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
146	Bittar: Renda Cidadã “respeitará teto de gastos” e será apre	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
61	Dona do Magalu abriu vaga para negros e desapontou liber	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
148	Relator diz que proposta do Renda Cidadã fica pronta, 'se [	1	1	0	0	1	0	1	0	0	1	0	0,53030303	1,94285714	0,87790195	0,51130260
116	“Resta chorar”, diz mãe de estudante goiana que morreu na	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
133	Pesquisa investiga uso de planta no tratamento de três tipo	0	1	0	0	0	0	0	1	0	0	0	0,21212121	0,625	0,23312447	0,21076531
1	Bolsonaro nunca aceitou sentar comigo para ver a realidad	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	Mandetta conta bastidores do Ministério da Saúde na pand	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
136	Morre Eddie Van Halen, um dos maiores guitarristas do roc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	Santos vence no Equador e deixa vaga encaminhada na Li	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	No epicentro da pandemia no México, vírus agrava pobreza	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	Largada eleitoral - A campanha municipal está para começ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	Radioamador André Sampaio morre aos 70 anos, em Sergi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
137	Conselho de Educação permite aula remota até fim de 202	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
83	Homem é preso suspeito de revender carga de produtos fa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	Novatos na berlinda - Até a semana passada, apenas um g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
101	Motorista que dirigia a 180 km/h vira réu na Justiça por tent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
84	Homem é preso por suspeita de estuprar uma criança de 5	0	1	0	0	1	0	0	0	0	0	0	0,25757575	0,7	0,24412592	0,24273914
2	Bolsonaro ignorou impacto da Covid-19 por decisão consci	1	0	1	1	0	0	0	0	0	0	0	0,42424242	1,58333333	0,74915271	0,38135321
30	Próxima crise financeira pode chegar em breve - Qual é a p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	Gente detergente - Eu fico aqui meio pasma com a capacid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
57	EUA tem nova noite de protestos contra o racismo, e 24 sã	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Ataque a faca em Paris deixa ao menos 2 feridos perto de a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
147	Bittar diz que apresentação do Renda Cidadã fica para a se	2	2	0	2	0	1	0	2	0	0	1	1,10606060	4,00757575	1,79774305	1,03553155
11	A saúde mental na pandemia - Os problemas emocionais n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
135	Guitarrista Eddie Van Halen morre aos 65 anos - O icônico	1	1	1	0	1	0	1	0	0	0	0	0,63636363	2,17619047	0,96337358	0,59723855
108	Campinas tem 1ª semana sem novas internações em UTI d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	Menos mortes - O número de pessoas mortas por policiais	1	1	0	0	0	0	0	0	0	0	0	0,31818181	1,5	0,86467915	0,26856346
37	Nós humanos às vezes somos histéricos - No pronto-socor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
62	Diversidade em Hollywood é tema de reunião no Congress	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 12 – Resultados dos métodos de cálculo de similaridade

5.3 Definição do método

Nesta seção, diferentes métodos foram experimentados, comparados e, dentre eles, um foi escolhido para determinar a similaridade entre notícias. Os resultados de cada método sobre os vetores produzidos podem ser observados na figura 12.

Dentre os métodos experimentados, o cálculo por soma de frações mostrou-se mais adequado por dois motivos. Primeiramente, a construção dos índices mimetiza a interpretação realizada intelectualmente, o que favorece a probabilidade de que uma notícia identificada como similar pelo comportamento humano também seja identificada dessa forma pelo algoritmo. Em segundo lugar, o custo computacional é reduzido em relação aos demais métodos, favorecendo a execução deste método em segundo plano com menor impacto para o usuário. Neste método, cada valor no vetor foi dividido por sua respectiva posição (o primeiro dividido por um, o segundo por dois e assim sucessivamente) e somado à divisão do valor anterior.

Os índices gerados por este método utilizam predominantemente a casa das unidades e, com isso, duas casas decimais são suficientes para representar a similaridade de notícias. Como a faixa de valores aumenta juntamente com a quantidade de notícias similares lidas sobre um assunto, o cálculo pode ser interrompido quando a similaridade ultrapassar um determinado valor.

O método de pesos lineares — que consiste em uma média ponderada na qual se

atribui peso onze para a primeira posição e dez para a segunda, decrescendo linearmente até a última posição — também apresentou índices promissores, mas o custo computacional é penalizado pela elevada dependência de casas decimais. Os resultados estão contidos em uma curta faixa de valores, tornando necessário o uso de duas ou mais casas decimais e utilizando a casa das unidades apenas quando há duas ou mais notícias similares no corpus.

Outro método experimentado foi o de pesos exponenciais — que utiliza a mesma lógica dos pesos lineares, mas utiliza expoentes da base e como pesos para a média ponderada (a primeira posição recebe peso e^{11}, a segunda recebe e^{10} e assim por diante) — mas este não obteve índices nem custo computacional desejáveis. O índice calculado sofreu elevada penalidade nos casos em que os valores estão dispersos pelas posições dos vetores, o que dificulta a definição de um intervalo de valores de similaridade para notícias previamente lidas que apresente poucos falsos-positivos para o corpus utilizado.

Por fim, o método de pesos logarítmicos — que substitui as potências na base e por logaritmos naturais ($ln(11)$ para a primeira posição, $ln(10)$ para a segunda posição, sucessivamente até a última posição) — apresentou índices comparáveis aos do método de pesos lineares, mas com custo computacional ainda maior. O comportamento dos índices também foi condizente com as comparações realizadas intelectualmente sem a probabilidade de falsos-positivos do método anterior, mas raramente utilizando a casa das unidades. Adicionalmente, o uso de operações logarítmicas apresenta custo computacional demasiado para um índice que deve ser calculado a cada nova notícia.

5.4 Definição dos intervalos

A escolha do método de soma de frações torna possível calcular a similaridade entre a notícia a ser lida e as contidas no histórico e, consequentemente, determinar intervalos que permitem dizer ao usuário se determinada notícia oferece informações novas ou não. Esses intervalos são discutidos na presente seção.

Para a análise deste corpus, o intervalo de 0 a 1,5 contempla todas as notícias completamente novas para o usuário. Nota-se, por exemplo, que o cenário abordado na seção 5.1 referente à notícia de nº 18 da figura 12 produz o maior índice para uma notícia que não possui outra similar no histórico simulado. Além disso, todas as notícias desse histórico que possuem outras similares apresentam valor superior a este.

É possível que notícias com índice superior a 1,5 possuam alguma informação desconhecida para o usuário, embora não tenha ocorrido no corpus deste trabalho. Uma situação parecida acontece na notícia de nº 148 (presente na figura 12) ao ser comparada com a nº 146 — ambas possuem o mesmo assunto, mas as diferenças de escrita das duas notícias causam dispersão dos números de coincidências no vetor.

Todavia, a similaridade calculada aumenta juntamente com a quantidade de notícias similares lidas pelo usuário e, consequentemente, diminui a probabilidade de que a notícia apresente alguma novidade. Isso acontece porque os sumários são concatenados no processo de agrupamento por assunto e, portanto, há mais palavras sendo comparadas. Ao calcular o

índice para a notícia de nº 147, o histórico simulado já conta com outras duas notícias do mesmo assunto, implicando maior similaridade dentre as comparações.

Um último teste foi realizado para investigar a hipótese descrita na seção 4.3.3 sobre qual seria o impacto de reduzir o tamanho dos sumários de onze para sete palavras. Conforme esperado, não houve mudanças expressivas nos índices e os valores dos intervalos puderam ser mantidos. Os resultados para onze e sete palavras podem ser observados e comparados na figura 13.

#	Título	1	2	3	4	5	6	7	8	9	10	11	Soma de frações (11)	Soma de frações (7)
76	A importância do diagnóstico para o câncer infantojuvenil -	0	0	0	0	0	0	0	0	0	0	0	0	0
146	Bittar: Renda Cidadã “respeitará teto de gastos” e será apr	0	0	0	0	0	0	0	0	0	0	0	0	0
61	Dona do Magalu abriu vaga para negros e desapontou libe	0	0	0	0	0	0	0	0	0	0	0	0	0
148	Relator diz que proposta do Renda Cidadã fica pronta, 'se	1	1	0	0	1	0	1	0	0	1	0	1,942857143	1,842857143
116	“Resta chorar”, diz mãe de estudante goiana que morreu n	0	0	0	0	0	0	0	0	0	0	0	0	0
133	Pesquisa investiga uso de planta no tratamento de três tipo	0	1	0	0	0	0	0	1	0	0	0	0,625	0,5
1	Bolsonaro nunca aceitou sentar comigo para ver a realidad	0	0	0	0	0	0	0	0	0	0	0	0	0
19	Mandetta conta bastidores do Ministério da Saúde na pand	0	0	0	0	0	0	0	0	0	0	0	0	0
136	Morre Eddie Van Halen, um dos maiores guitarristas do roc	0	0	0	0	0	0	0	0	0	0	0	0	0
29	Santos vence no Equador e deixa vaga encaminhada na Li	0	0	0	0	0	0	0	0	0	0	0	0	0
38	No epicentro da pandemia no México, vírus agrava pobrez	0	0	0	0	0	0	0	0	0	0	0	0	0
23	Largada eleitoral - A campanha municipal está para começ	0	0	0	0	0	0	0	0	0	0	0	0	0
90	Radioamador André Sampaio morre aos 70 anos, em Serg	0	0	0	0	0	0	0	0	0	0	0	0	0
137	Conselho de Educação permite aula remota até fim de 202	0	0	0	0	0	0	0	0	0	0	0	0	0
83	Homem é preso suspeito de revender carga de produtos fa	0	0	0	0	0	0	0	0	0	0	0	0	0
36	Novatos na berlinda - Até a semana passada, apenas um g	0	0	0	0	0	0	0	0	0	0	0	0	0
101	Motorista que dirigia a 180 km/h vira réu na Justiça por tent	0	0	0	0	0	0	0	0	0	0	0	0	0
84	Homem é preso por suspeita de estuprar uma criança de 5	0	1	0	0	1	0	0	0	0	0	0	0,7	0,7
2	Bolsonaro ignorou impacto da Covid-19 por decisão consci	1	0	1	1	0	0	0	0	0	0	0	1,583333333	1,583333333
30	Próxima crise financeira pode chegar em breve - Qual é a p	0	0	0	0	0	0	0	0	0	0	0	0	0
49	Gente detergente - Eu fico aqui meio pasma com a capacid	0	0	0	0	0	0	0	0	0	0	0	0	0
57	EUA tem nova noite de protestos contra o racismo, e 24 sã	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Ataque a faca em Paris deixa ao menos 2 feridos perto de	0	0	0	0	0	0	0	0	0	0	0	0	0
147	Bittar diz que apresentação do Renda Cidadã fica para a se	2	2	0	2	0	1	0	2	0	0	1	4,007575758	3,666666667
11	A saúde mental na pandemia - Os problemas emocionais n	0	0	0	0	0	0	0	0	0	0	0	0	0
135	Guitarrista Eddie Van Halen morre aos 65 anos - O icônico	1	1	1	0	1	0	1	0	0	0	0	2,176190476	2,176190476
108	Campinas tem 1ª semana sem novas internações em UTI c	0	0	0	0	0	0	0	0	0	0	0	0	0
18	Menos mortes - O número de pessoas mortas por policiais	1	1	0	0	0	0	0	0	0	0	0	1,5	1,5
37	Nós humanos às vezes somos histéricos - No pronto-socor	0	0	0	0	0	0	0	0	0	0	0	0	0
62	Diversidade em Hollywood é tema de reunião no Congress	0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 13 – Resultados da soma de frações para onze e sete palavras

6 Conclusão

“Há sempre uma estrada em qualquer lugar, para qualquer pessoa, em qualquer circunstância. Como, onde, por quê?”
Jack Kerouac

O trabalho em questão visou propor um método algorítmico que determinasse a similaridade de uma nova notícia em relação às presentes no histórico de leitura do usuário. Isso possibilita que o indivíduo decida se lerá ou não uma dada notícia com base na probabilidade de que ela lhe apresente novas informações.

A necessidade de elaborar tal algoritmo é fundada no crescimento do uso problemático de dispositivos. Esse comportamento pode trazer impactos negativos sobre a produtividade e vida pessoal dos usuários e está altamente conectado ao consumo de notícias.

A seção 6.1 é referente aos resultados obtidos neste trabalho. Na seção 6.2 são abordadas limitações referentes aos procedimentos utilizados no desenvolvimento, bem como limitações sobre a natureza das técnicas empregadas. Por fim, possibilidades e recomendações de trabalhos futuros são contemplados na seção 6.3.

6.1 Resultados

Especificamente, este trabalho encontrou características particulares das notícias que são úteis para o algoritmo de classificação automática. Para esse gênero textual, o título e o corpo das notícias foram os elementos necessários para a identificação do assunto abordado.

Além disso, este trabalho percebeu elevada recorrência de determinadas palavras em notícias com o mesmo assunto. Por exemplo, em notícias que abordavam política, era comum observar termos relacionados à localização geográfica e aos nomes das pessoas envolvidas. Este resultado é promissor para o processo de agrupamento de notícias por assunto e permite que técnicas de classificação automática como o TF-IDF sejam utilizadas.

Ademais, observou-se um padrão em que palavras altamente discriminatórias eram mais repetidas em notícias similares. Nas três notícias coletadas que tratavam da apresentação do programa Renda Cidadã, a título de exemplo, houve repetição das palavras “Bittar”, “renda”, “cidadã” e “teto” entre as sete mais discriminatórias. Este padrão é o que possibilita o reconhecimento da similaridade entre notícias pelo algoritmo de classificação.

Este trabalho também elaborou o método algorítmico que se beneficia dos padrões encontrados para realizar a classificação automática das notícias e, assim, determinar a similaridade entre elas. O método em questão utilizou a técnica de TF-IDF para sumarização e classificação de notícias, cujos resultados são descritos a seguir.

A execução desse método sobre um corpus de trinta notícias produziu sumários que, na maioria dos casos, permitem a inferência intelectual do conteúdo abordado na notícia. Este

resultado é útil para averiguar a qualidade dos sumários utilizados nas comparações, visto que um sumário pouco relacionado com seu conteúdo impactaria negativamente os resultados da etapa de comparação.

Entretanto, para as primeiras notícias do corpus, houve elevada variação e ocorrência de *stop words*. Na perspectiva da implementação em trabalhos futuros, isso implica um cuidado com a expiração do *cache* de leitura para prevenir que ele se esvazie.

Nos vetores de comparação de sumários produzidos pelo método algorítmico, três resultados foram observados. Primeiramente, notícias similares produziram vetores com alta concentração de coincidências entre as palavras mais discriminatórias. Além disso, notícias pouco relacionadas produziram vetores com poucas coincidências e essas eram dispersas entre si. Por fim, notícias sem relação alguma produziram vetores vazios. Esses resultados possibilitaram o cálculo da similaridade para este corpus, já que essas características permitem distinguir notícias similares das não-similares.

Para converter os vetores de comparação em valores que determinam a similaridade entre notícias, o cálculo por soma de frações foi capaz de representá-lo com apenas duas casas decimais. Portanto, o cálculo apresenta custo computacional reduzido em comparação com os demais analisados na seção 5.4. Além disso, nesse cálculo, a similaridade aumenta de acordo com a quantidade de notícias similares lidas, o que é uma característica positiva para distinguir numericamente notícias já conhecidas das demais.

Além de todos os resultados previstos, um resultado adicional, que consiste em um arquivo no formato XML, foi obtido no desenvolvimento deste trabalho. Este arquivo, fruto da construção do corpus utilizado na seção 4.2, contém as notícias de exemplo coletadas dos três veículos mais populares no Brasil e está disponível para trabalhos futuros no anexo B.

6.2 Limitações

Este trabalho se limitou a utilizar apenas notícias em português brasileiro, de veículos de notícias com conteúdos variados e sem restrição de assunto. Como consequência, os resultados obtidos não se aplicam a outros idiomas, tampouco a comportamentos de leitura específicos que exijam o adentramento em particularidades de determinados assuntos ou de padrões de escrita.

O método para a produção do corpus utilizado neste trabalho obteve, após o acréscimo manual através do Google Notícias, um total de cento e quarenta e nove notícias. Assim sendo, é provável que os resultados não representem situações de mundo real.

Além disso, devido à utilização dos *feeds* RSS para tal, a coleção de notícias foi restrita a três veículos com diferentes quantidades de cada um — sendo mais de metade das notícias oriundas da Folha de S. Paulo — aumentando a probabilidade de enviesamento dos resultados obtidos.

Outra limitação se encontra na natureza da sumarização e classificação por TF-IDF, que é incapaz de identificar sinônimos, flexões de gênero e de número, entre outras variações

que possuam a mesma semântica para o leitor. Isso aumenta a ocorrência de falsos-positivos, já que duas notícias podem oferecer sumários semanticamente similares e que se diferenciam apenas por plurais ou sinônimos e, em situações como essa, o método algorítmico as trataria como notícias distintas.

6.3 Trabalhos futuros

As limitações citadas previamente também abrem possibilidades para trabalhos futuros. Por exemplo, este trabalho pode ser desenvolvido em outros idiomas ou adaptado a categorias de notícias que necessitem uma ótica diferente sobre o comportamento de leitura do usuário em questão.

Como o trabalho foi desenvolvido sobre um corpus pequeno, recomenda-se a utilização de um corpus mais expressivo em trabalhos futuros, proveniente de mais veículos de notícias e com outro método de obtenção dos URLs.

A limitação do TF-IDF no aspecto semântico poderia ser mitigada em um trabalho futuro através do uso de um dicionário e/ou um *gazetteer*. Assim, o algoritmo poderá agrupar sinônimos relacionados a entidades e localizações.

Também é possível utilizar técnicas de *stemming* para reduzir as palavras das notícias a seus radicais e utilizá-las dessa forma na etapa de comparação. Esse procedimento contornaria os problemas com flexões de gênero e número que ocasionalmente poluíram os sumários.

Outra possibilidade está no emprego de outras técnicas, alternativas ao TF-IDF, para a sumarização e classificação das notícias. Por exemplo, algoritmos de redes neurais ou *Support-Vector Machines* podem ser utilizadas para classificação de texto para avaliar a eficiência e eficácia em comparação com a técnica utilizada.

Recomenda-se ainda a implementação e teste deste algoritmo em aplicações que possam ser disponibilizados ao público, como uma extensão para navegadores web ou aplicativos de leitura por RSS. O algoritmo desenvolvido está disponível juntamente com o arquivo XML das notícias coletadas nos anexos deste trabalho.

Sugere-se ainda a elaboração de testes com usuários finais para averiguar a eficácia do método em identificar notícias já lidas em situações cotidianas e para refinar os intervalos definidos na seção 5.4. Esse tipo de teste também poderia ser conduzido em pessoas impactadas pelo uso problemático de dispositivos com a intenção de investigar como o consumo de notícias é afetado pela utilização do método algorítmico desenvolvido neste trabalho.

Referências

AGGARWAL, C. C.; ZHAI, C. X. *A survey of text classification algorithms*. [S.l.], 2012. v. 9781461432234, 163–222 p. Citado na página 18.

Alexa Traffic Rank. *Top Sites in Brazil - Alexa*. Alexa Internet Inc., 2020. Disponível em: <https://www.alexa.com/topsites/countries/BR>. Citado nas páginas 20 e 22.

ALT, D. College students' academic motivation, media engagement and fear of missing out. *Computers in Human Behavior*, v. 49, p. 111–119, 2015. ISSN 07475632. Citado na página 14.

ALUÍSIO, S. M.; ALMEIDA, G. M. de B. O que é e como se constrói um corpus? lições aprendidas na compilação de vários corpora para pesquisa linguística. *Calidoscópio*, v. 4, n. 3, p. 156–178, 2006. Citado na página 20.

CANAVILHAS, J. Webjornalismo: considerações gerais sobre jornalismo na web. Livros Labcom, 2001. Citado na página 15.

CANAVILHAS, J. Webjornalismo: 7 caraterísticas que marcam a diferença. 2014. Citado nas páginas 15 e 16.

CHAN-OLMSTED, S.; RIM, H.; ZERBA, A. Mobile news adoption among young adults: Examining the roles of perceptions, news consumption, and media usage. *Journalism and Mass Communication Quarterly*, v. 90, n. 1, p. 126–147, 2013. ISSN 10776990. Disponível em: <http://jmcq.sagepub.com>. Citado nas páginas 11 e 15.

DAI, W. Improvement and Implementation of Feature Weighting Algorithm TF-IDF in Text Classification. In: . [S.l.: s.n.], 2018. Citado na página 18.

DEUZE, M. Online journalism. Valauskas, Edward J., 2001. Citado nas páginas 16 e 17.

DUKE, É.; MONTAG, C. Smartphone addiction, daily interruptions and self-reported productivity. *Addictive Behaviors Reports*, 2017. ISSN 23528532. Citado nas páginas 11 e 14.

ELHAI, J. D. et al. Fear of missing out, need for touch, anxiety and depression are related to problematic smartphone use. *Computers in Human Behavior*, Elsevier Ltd, v. 63, p. 509–516, 2016. ISSN 07475632. Disponível em: <http://dx.doi.org/10.1016/j.chb.2016.05.079>. Citado na página 13.

FACHINETTO, E. A. O hipertexto e as práticas de leitura. *Revista Querubim*, p. 27–45, 2005. Citado na página 16.

GEZGIN, D. M. et al. Social networks users: fear of missing out in preservice teachers. *Journal of Education and Practice*, v. 8, n. 17, p. 156–168, 2017. ISSN 2222-1735. Citado na página 14.

GHOLAMREZAZADEH, S.; SALEHI, M. A.; GHOLAMZADEH, B. A comprehensive survey on text summarization systems. In: *Proceedings of the 2009 2nd International Conference on Computer Science and Its Applications, CSA 2009*. [S.l.: s.n.], 2009. ISBN 9781424449460. Citado na página 17.

HERMIDA, A. et al. SHARE, LIKE, RECOMMEND: Decoding the social media news consumer. *Journalism Studies*, v. 13, n. 5-6, p. 815–824, 2012. ISSN 1461670X. Disponível em: <http://dx.doi.org/10.1080/1461670X.2012.664430>. Citado na página 15.

KIM, S.-W.; GIL, J.-M. Research paper classification systems based on tf-idf and lda schemes. *Human-centric Computing and Information Sciences*, Springer, v. 9, n. 1, p. 1–21, 2019. Citado na página 17.

KIYANI, F.; TAS, O. *A survey automatic text summarization*. [S.l.], 2017. v. 5, n. 1, 205–213 p. Disponível em: <www.ijcsit.com>. Citado nas páginas 17 e 18.

LOPES, A. *Portuguese stop words - GitHub*. GitHub, 2013. Disponível em: <https://gist.github.com/alopes/5358189>. Citado nas páginas 29 e 30.

MIELNICZUK, L. Jornalismo na web: uma contribuição para o estudo do formato da notícia na escrita hipertextual. 2003. Citado nas páginas 15 e 16.

MITCHELL, A. et al. The Modern News Consumer. p. 47, 2016. Disponível em: <http://www.journalism.org/2016/07/07/the-modern-news-consumer/>. Citado nas páginas 11, 14 e 15.

MOLYNEUX, L. Mobile News Consumption: A habit of snacking. *Digital Journalism*, v. 6, n. 5, p. 634–650, 2018. ISSN 2167082X. Disponível em: <https://doi.org/10.1080/21670811.2017.1334567>. Citado na página 15.

NENKOVA, A.; MCKEOWN, K. A survey of text summarization techniques. *Mining Text Data*, v. 9781461432234, p. 43–76, 2012. Disponível em: <http://dx.doi.org/10.1561/1500000015>. Citado nas páginas 17 e 18.

OH, K.-J. et al. Personalized news recommendation using classified keywords to capture user preference. In: IEEE. *16th International Conference on Advanced Communication Technology*. [S.l.], 2014. p. 1283–1287. Citado na página 17.

OULASVIRTA, A. et al. Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing*, v. 16, n. 1, p. 105–114, 2012. ISSN 16174909. Disponível em: <https://www.researchgate.net/publication/220141400_Habits_make_smartphone_use_more_pervasive>. Citado nas páginas 11 e 14.

RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: PISCATAWAY, NJ. *Proceedings of the first instructional conference on machine learning*. [S.l.], 2003. v. 242, p. 133–142. Citado na página 18.

RIBAS, B. Características da notícia na web-considerações sobre modelos narrativos. *Encontro Nacional de Pesquisadores em Jornalismo–SBPJor*, v. 2, 2004. Citado na página 16.

ROCHA, H. C. L. da; DANTAS, I. H. F. de A. et al. Webjornalismo: Dos portais às redes sociais. *Revista Observatório*, v. 3, n. 1, p. 374–395, 2017. Citado nas páginas 15 e 16.

SCHMIDT, A. L. et al. Anatomy of news consumption on Facebook. *Proceedings of the National Academy of Sciences of the United States of America*, v. 114, n. 12, p. 3035–3039, 2017. ISSN 10916490. Disponível em: <www.pnas.org/lookup/suppl/doi:10.1073/pnas.1617052114/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1617052114>. Citado nas páginas 11, 14 e 15.

WAZLAWICK, R. S. Uma Reflexão sobre a Pesquisa em Ciência da Computação à Luz da Classificação das Ciências e do Método Científico. *Revista de Sistemas de Informação da FSMA*, nº 6, n. January 2010, p. 3–10, 2010. Citado na página 19.

WHITING, A.; WILLIAMS, D. Why people use social media: a uses and gratifications approach. *Qualitative Market Research: An International Journal*, v. 16, n. 4, p. 362–369, 2013. ISSN 13522752. Disponível em: <www.emeraldinsight.com/1352-2752.htm>. Citado nas páginas 11, 14 e 15.

ZHANG, W.; YOSHIDA, T.; TANG, X. A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, v. 38, n. 3, p. 2758–2765, mar 2011. ISSN 09574174. Citado na página 18.

Anexos

ANEXO A – Algoritmo desenvolvido

```
# funcao para extrair texto dos arquivos XML
def getText(nodelist):
    rc = []
    for node in nodelist:
        if node.nodeType == node.TEXT_NODE:
            rc.append(node.data)
    return ''.join(rc)

# criando corpus com todas as noticias
corpus = buildCorpus()

# lista de stop words
with open("stopwords.txt", encoding='utf-8') as file:
    stopwords = file.read().splitlines()

# funcao para construcao do corpus com todas as noticias
def buildCorpus():

    # lendo arquivos xml
    uolTree = xml.dom.minidom.parse('uol.xml').documentElement
    g1Tree = xml.dom.minidom.parse('g1.xml').documentElement
    metropolesTree = xml.dom.minidom.parse('metropoles.xml').
       ↪ documentElement
    gNewsTree = xml.dom.minidom.parse('gnews.xml').documentElement

    # inicializando variaveis locais
    corpus = []
    noticia = ''

    # adicionando noticias do Uol
    items = uolTree.getElementsByTagName("item")

    for item in items:
        noticia = getText(item.getElementsByTagName('titulo')[0].
           ↪ childNodes) + '␣-␣' + getText(item.getElementsByTagName
           ↪ ('corpo')[0].childNodes)
        corpus.append(noticia)
```

```
    # adicionando noticias do G1
    items = g1Tree.getElementsByTagName("item")

    for item in items:
        noticia = getText(item.getElementsByTagName('titulo')[0].
            ↪ childNodes) + ' - ' + getText(item.getElementsByTagName
            ↪ ('corpo')[0].childNodes)
        corpus.append(noticia)

    # adicionando noticias do Metropoles
    items = metropolesTree.getElementsByTagName("item")

    for item in items:
        noticia = getText(item.getElementsByTagName('titulo')[0].
            ↪ childNodes) + ' - ' + getText(item.getElementsByTagName
            ↪ ('corpo')[0].childNodes)
        corpus.append(noticia)

    # adicionando noticias do google news
    items = gNewsTree.getElementsByTagName("item")
    for item in items:
        noticia = getText(item.getElementsByTagName('titulo')[0].
            ↪ childNodes) + ' - ' + getText(item.getElementsByTagName
            ↪ ('corpo')[0].childNodes)
        corpus.append(noticia)

    # retorna todas as noticias contidas nos arquivos
    return corpus

# classe que representa uma sessao de navegacao do usuario
class Sessao:

    # inicializando objeto
    def __init__(self):
        self.historico = []
        self.timestamps = []

    # funcao que adiciona uma nova noticia ao historico
    def adicionar(self, indice):
        self.historico.append(corpus[indice])

    # funcao que calcula a similaridade de uma nova noticia sobre o
        ↪ historico
```

```
def comparar(self, indice):

    # anexando temporariamente a noticia a ser comparada ao
      ↪ historico para calcular o TF-IDF
    temp = self.historico[0:len(self.historico)]
    temp.append(corpus[indice])

    # iniciando variaveis de comparacao
    sumarios = []
    mesmoAssunto = []
    vetContagem = [0] * 11
    num = 0

    # gerando TF-IDF do historico + noticia a ser comparada
    vectorizer = TfidfVectorizer(stop_words=stopwords)
    tfidf = vectorizer.fit_transform(temp).toarray()

    # gerando sumarizacao do historico + noticia a ser comparada
    for i in tfidf:
        df = pd.DataFrame(i, index=vectorizer.get_feature_names()
          ↪ , columns=["TF-IDF"])
        df = df[(df.T != 0).any()]
        df = df.sort_values('TF-IDF', ascending=False)
          ↪ sumarios.append(df.index[:11].values)

    # criando subconjunto de noticias do mesmo assunto da noticia
      ↪ a ser lida
    for i in reversed(range(len(sumarios) - 1)):
        cont = 0
        for j in range(len(sumarios[i])):
            if sumarios[-1][j] in sumarios[i]:
                cont = cont + 1
            if cont >= 2:
                mesmoAssunto.extend(sumarios[i])
                break

    # calculando coincid ncias entre o sumario da noticia a ser
      ↪ comparada e o subconjunto
    for i in range(len(sumarios[-1])):
        vetContagem[i] = vetContagem[i] + Counter(mesmoAssunto)[
          ↪ sumarios[-1][i]]

    # calculando similaridade
```

```
for i in range(len(vetContagem)):
    num = num + vetContagem[-i]/(i+1)
```

ANEXO B – Arquivos disponibilizados

Algoritmo desenvolvido neste trabalho: Disponível em: <https://github.com/KRoyther/similaridade/blob/main/classificador.py>

Arquivos XML com as notícias coletadas: Disponível em: <https://github.com/KRoyther/similaridade/tree/main/Arquivos%20XML>