Um framework orientado a artigos para análise semântica automática de pesquisas sobre COVID-19

Antonio Alves[1], Antônio Pereira[1], Pablo Cecilio[1], Nayara Pena[1],
Felipe Viegas[2], Elisa Tuler[1], Diego Dias[1], Leonardo Rocha[1*]
[1]Universidade Federal de São João del-Rei
(antoniopedro,antoniopereira,cecilio,nayara.p.pena)@aluno.ufsj.edu.br
(etuler,diegodias,lcrocha)@ufsj.edu.br
[2]Universidade Federal de Minas Gerais
frviegas@dcc.ufmg.br

Abstract

In this work, we propose a framework that automatically extracts semantic topics from scientific publications related to research on COVID-19. The framework has four main building blocks: (i) pre-processing, (ii) topic modeling, (iii) topic correlation with authors and institutions, and (iv) summarization interface. The first block corresponds to the application of pre-processing strategies in texts on the considered articles and the definition of their semantic representation. The topic modeling block aims to find semantic topics in the publications used. The third block correlates these topics with the articles themselves and the authors, institutions, and countries related to each article. The summary interface provides an intuitive view for all these analyses. Our evaluation shows that our framework is capable of automatically extracting relevant characteristics from the articles, identifying the main themes addressed by them, as well as the correlation of researchers, institutions and countries for different topics of research on COVID-19.

Keywords

Word Embeddings, Topic Modeling, COVID-19

1 Introdução

A Internet apresentou significativa expansão e popularização nos últimos anos. A cada dia são criados novas aplicações que geram e consomem uma quantidade cada vez maior de dados. Um dos maiores desafios é fornecer ferramentas que possam realizar análises inteligentes e automáticas sobre esses dados. Um cenário que ilustra este contexto é da produção científica, cujo crescimento é exponencial em termos da quantidade de artigos, pesquisadores e de pesquisas em andamento. Ademais, este cenário vem recebendo ainda mais contribuições em decorrência do intenso trabalho de pesquisadores em todo o mundo para conter o avanço da COVID-19. Nesse sentido, organizar semanticamente as informações fornecidas pelos artigos científicos publicados sobre esta família de vírus, pode ser extremamente importante para apontar rumos de pesquisa mais promissores, identificar abordagens ainda pouco exploradas e sugerir colaborações de pesquisadores com destaque em diferentes áreas.

*This work was partially funded by CAPES, CNPq, and Fapemig.

In: XIX Workshop de Ferramentas e Aplicações (WFA 2020), São Luís, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2020.

ISSN 2596-1683

Neste artigo, propomos um *framework* que permite filtrar, resumir e analisar artigos científicos e publicados em diferentes plataformas digitais. Nosso *framework* é capaz de identificar os principais tópicos abordados por esses artigos, correlacionando-os aos autores, suas instituições e países destas. Ele é composto por quatro blocos de construção principais, a saber, (i) pré-processamento, (ii) modelagem de tópicos, (iii) correlação dos tópicos com autores, instituições e países e (iv) interface de sumarização.

Para avaliar nossa proposta, instanciamos o *framework* proposto e fornecemos uma extensa análise considerando um *dataset* composto por artigos publicados sobre COVID-19, SARS-CoV-2 e outros coronavírus, disponíveis em Kaggle [1]. Instanciamos o primeiro bloco de nosso *framework* aplicando algumas estratégias de pré-processamento: conversão para letras minúsculas, remoção de pontuação, acento e *stopwords* e uma abordagem de reconhecimento de entidade [6]. Além disso, utilizamos os conceitos de CluWords [8], consideradas o estado da arte, para representar semanticamente esses dados. No segundo bloco, adotamos a estratégia de modelagem de tópicos NMF [3] para inferir os diferentes tópicos do nosso conjunto de trabalhos. Para o terceiro bloco, propomos uma estratégia que consiste na manipulação das matrizes fornecidas pelo NMF, que permite correlacionar os temas descobertos a artigos, seus autores, suas instituições e países. Por fim, apresentamos uma proposta de interface visual que resume todas as informações geradas, destacando os principais tópicos obtidos e suas correlações. Esta interface está disponível em labpi.ufsj.edu.br/covpapers/ [1]. Analisando os resultados alcançados pela instanciação de nosso *framework*, pudemos responder a várias questões como: *Quais os autores que mais se correlacionam com um determinado tópico?*; *Quais linhas de pesquisa são mais exploradas por pesquisadores de uma instituição?*, *Quais os países que conduzem pesquisas em uma determinada área?* e *Quais tópicos são mais explorados por um determinado autor?*. Além disso, com a definição de tópicos relacionados a um conjunto de artigos, discutimos como é possível auxiliar: (i) empresas a constituírem equipes multidisciplinares a partir dos pesquisadores que mais contribuem para a área desejada; (ii) órgãos governamentais, com foco em Educação e Pesquisa, para dar suporte na determinação de áreas mais e menos estudadas, permitindo um redirecionamento mais igualitário dos recursos financeiros.

2 Trabalhos Relacionados

A proposta de estratégias que visam fornecer mecanismos para filtrar, sintetizar e analisar a produção científica de uma área específica

[1]nome de usuário: user, senha: covid-19

não é nova [2]. No entanto, todos esses trabalhos se concentram em considerar somente metadados para fornecer esse tipo de análise, como título, palavras-chave e *tags*, argumentando que são suficientes para fornecer uma sumarização de dados eficiente. Neste artigo, levantamos algumas questões sobre esses trabalhos: *Até que ponto os metadados são bons descritores em uma base ampla?*; *Palavras-chave, por si só, são bons descritores sobre um grande repositório de artigos?* e *Um pesquisador que deseja se manter atualizado sobre a produção científica de sua área pode contar apenas com palavras-chave para obter essas informações?*. Na verdade, acreditamos que a resposta a todas essas questões seja **"Não!"**, motivando-nos a apresentar nosso *framework* que fornece todas essas análises considerando outras informações textuais que compõem os artigos. Embora potencialmente capaz de fornecer informações mais ricas em detalhes, o uso de informações textuais maiores também é mais desafiador, implicando na aplicação de estratégias de pré-processamento para minimizar o ruído textual que pode prejudicar etapas posteriores de um processamento de aprendizado de máquina. No contexto da Modelagem de Tópicos, em [7] os autores mostram que, apesar de não haver uma sequência correta, genérica e única de técnicas de pré-processamento, conversão para letras minúsculas e a remoção de pontuação e de *stopwords* podem alcançar resultados mais eficazes.

Outra etapa que também é considerada no processo de pré-processamento textual é a representação dos dados. Basicamente, é necessário construir uma representação de matriz $m \times n$ artigo-termo (ou alguma codificação de termo latente) do conjunto de dados. A representação amplamente adotada desta matriz na mineração de texto explora o paradigma TF-IDF (e suas variantes). Os principais problemas com a representação do TF-IDF têm a ver com sua alta dimensionalidade, dispersão e falta de informações úteis, como contexto. Outras representações alternativas recentes têm conseguido produzir espaços mais compacto em termos de dimensões latentes (por exemplo, *embeddings* de palavras), tais como [8].

A partir do pré-processamento e representação dos dados textuais é possível utilizar o conjunto de estratégias mais usadas para resumir, organizar e analisar grandes volumes de dados textuais: Modelagem de Tópicos (MT). A MT aborda o problema de descobrir relações entre documentos (D) e tópicos (Z), bem como relações entre as palavras (W) que compõem os documentos e tópicos. Cada tópico $z_i \in D$ consiste em uma distribuição probabilística entre palavras que coocorrem de forma frequente, e os documentos, por sua vez, são representados por distribuições probabilísticas entre os tópicos. Uma técnica de MT baseada em decomposição de matrizes que tem ganhado destaque é a *Non-Negative Matrix Factorization* (NMF) [3]. Resumidamente, ela realiza uma fatoração de matrizes, cujas matrizes resultantes não possuem valores negativos. Essa técnica, que tem como parâmetro apenas o número de tópicos, decompõe a matriz $m \times n$ (artigo-termo) em duas outras matrizes, W (de documentos por tópicos) e H (de tópicos por termos). Do ponto de vista do nosso trabalho, as matrizes decompostas são muito valiosas para análises posteriores. Em [4], por exemplo, os autores utilizam essas matrizes para realizar uma análise de sentimento sobre os tópicos obtidos dos *reviews* realizados em aplicativos móveis na Google PlayStore.

3 Trabalho Proposto

Nesta seção apresentamos nosso *framework* capaz de extrair automaticamente características relevantes dos artigos, identificando os principais tópicos por eles abordados, bem como os principais tópicos relacionados aos autores, instituições de pesquisa e países.

3.1 Pré-processamento

Propomos a instanciação do bloco de pré-processamento utilizando quatro estratégias: conversão de maiúsculo para minúsculo, remoção de *stopwords*, tokenização e remoção de algumas entidades identificadas no processo de Reconhecimento de Entidades Nomeadas (ou *Named Entity Recognition*, NER). A conversão dos caracteres para minúsculo, como mostrado por [7], assegura melhores resultados para algoritmos de classificação de texto. Em seguida, a remoção de *stopwords* do texto, que são palavras que pouco contribuem para análises textuais relevantes. Como essa lista de palavras indesejadas varia para cada contexto, nossa proposta consiste em utilizar a lista de *stopwords* da biblioteca spaCy, adicionando algumas outras palavras ruidosas e específicas do contexto do *dataset* a ser analisado. Com um texto mais conciso, é possível avançar para a terceira estratégia que é a tokenização - processo de transformação das palavras do texto em tokens únicos. A próxima etapa é um recurso amplamente utilizado na literatura para categorizar pessoas, lugares, organizações e outras entidades de interesse no texto [6], conhecido como *Named Entity Recognition* (NER). Como nosso foco é trabalhar com bases de dados acadêmicas, compostas por artigos científicos publicados, é comum o uso de expressões numéricas, períodos de tempo, descrições de pessoas ou espaços físicos, tais como "*na **segunda semana** os resultados foram melhores*" ou "*o experimento foi realizado no **laboratório Turing***". Para remover essas entidades textuais, que pouco acrescentam para técnicas de mineração textual, fizemos uso do NER para removê-las. Por fim, define-se a representação do conjunto de artigos como uma matriz $m \times n$ (matriz de artigo-termo). Para isso, adotamos o conceito de *CluWords* apresentados em [8], isto é, cada documento tem uma nova representação em que as palavras originais são substituídas por um grupo de palavras chamadas de *CluWords*, que hoje correspondem aos melhores resultados relatados na literatura para tarefas de Modelagem de Tópicos [8].

3.2 Modelagem de Tópicos

Este bloco é baseado na estratégia *Non-Negative Matrix Fatorization* (NMF), que realiza simultaneamente redução de dimensão e clusterização, com aplicação bem-sucedida na modelagem de tópico [3]. O NMF realiza a decomposição da matriz (não-negativa) $A \in \mathbb{R}^{n \times m}$, onde n é o número de documentos e m o número de termos, em função de fatores não-negativos $H \in \mathbb{R}^{n \times k}$ e $W \in \mathbb{R}^{k \times m}$. H codifica a relação entre documentos e tópicos, enquanto W codifica a relação entre termos e tópicos. A Figura 1 ilustra este processo.

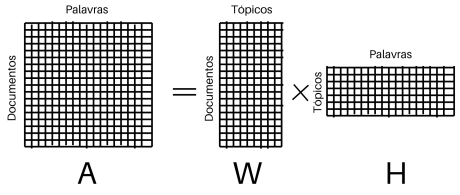

Figura 1: Processo de Decomposição NMF

3.3 Correlação dos tópicos com autores, instituições e países

Um ponto chave para enriquecer ainda mais nossas análises é a descoberta de quais personagens (autores, instituições e países) mais se aproximam sobre cada tópico. Vamos considerar que as duas etapas anteriores de nosso *framework* tenham sido aplicadas em uma base de dados contendo artigos científicos relacionados à área de Ciência da Computação. Assim, ao final das duas primeiras etapas, o resultado são as matrizes H e W que relacionam artigos com tópicos e tópicos por palavras, respectivamente. Seguindo o exemplo, considere que o i-ésimo artigo na matriz H trata, majoritamente, do tópico "Information Retrieval", enquanto que o j-ésimo artigo trata do tópico "Data Mining". Um questionamento pertinente que pode ser formulado seria: quem são os autores do artigo que trata, majoritariamente, do tópico "Information Retrieval"? É evidente que, para este exemplo, cada artigo possui um ou mais autores, logo é possível, por meio das relações entre artigos e tópicos produzidas, destacar quais autores mais abordam os tópicos obtidas pela MT. De forma análoga, como os autores de um artigo pertencem à instituições de pesquisa, também é possível destacar as instituições por tópicos, também considerando a relação entre artigos e tópicos. Finalmente, uma vez que toda instituição está localizada em algum país, é possível relacionar países a tópicos de pesquisa.

Ainda com o exemplo anterior, considerando que as matrizes H e W tenham sido geradas para três tópicos. Primeiramente, é possível identificar do que se trata cada um dos tópicos analisando a matriz H e encontrando quais as palavras mais fortemente associadas a cada um dos tópicos. Vamos supor o exemplo em que o primeiro tópico esteja majoritariamente associado a "Information Retrieval", o segundo tópico relacionado à "Artificial Intelligence" e o terceiro relacionado à "Data Mining". Feito isso, o próximo passo é analisar a matriz W que relaciona documentos e tópicos. Para isso, vamos considerar como exemplo a primeira matriz da Figura 1, contendo três artigos, onde cada posição apresenta a "relevância"do tópico para o documento. A partir dessa matriz, podemos agrupar e somar os valores de tópicos obtidos para os artigos que pertençam a um mesmo autor, levando-nos à segunda matriz da Figura 1. Supondo que todos os três artigos da primeira matriz pertença ao primeiro autor da segunda matriz, podemos inferir como esse autor se relaciona com cada um dos tópicos (em termos dos resultados numéricos do NMF): 145 (60+45+40) para o tópico "Information Retrieval", 160 (80+30+50) para o tópico "Artificial Intelligence" e 130 (20+70+40) para o tópico "Data Mining". Considerando a matriz W completa, esse mesmo processo pode ser aplicado a todos os autores da coleção.

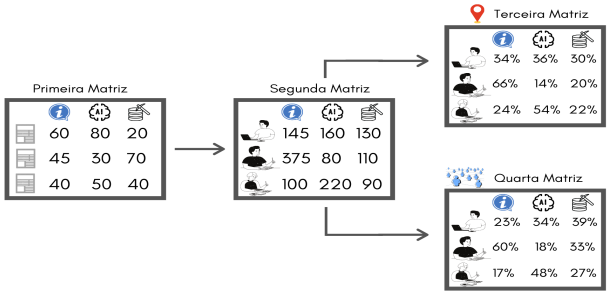

Figura 2: Cálculo de Contribuições de Autores para Tópicos

A partir desses dados, podemos realizar duas análises distintas: (i) calcular, para cada autor, a distribuição entre os tópicos pelos quais seus trabalhos se relacionam; ou (ii) avaliar, dentre todos os autores, aqueles que possuem seus trabalhos mais correlacionados a cada um dos tópicos. A primeira análise (i) se trata de uma informação local, no sentido que se baseia somente no que foi escrito por um mesmo autor. Assim, realizando uma normalização em cada uma das linhas que representa cada autor na segunda matriz da Figura 1 é possível saber o quanto as pesquisas de cada autor estão relacionadas a cada um dos tópicos. O resultado dessa análise é ilustrada na terceira matriz da Figura 2. Analisando a primeira linha que representa o primeiro autor, podemos observar que seus artigos estão 34% relacionados ao tópico "Information Retrieval", 36% ao tópico "Artificial Intelligence" e 30% ao tópico "Data Mining". Podemos afimar que os artigos desse autor tratam mais de assuntos relacionados à "Artificial Intelligence". Por sua vez, a segunda análise (ii) é baseada em uma informação global e considera tudo o que foi escrito para um mesmo tópico para distinguir aqueles autores que tem seus artigos mais alinhados com os tópicos obtidos. Nesse contexto, a quarta matriz da Figura 2 é construída normalizando cada uma das colunas da segunda matriz, as quais representam os tópicos obtidos. Dessa forma, seguindo nosso exemplo, podemos observar que o primeiro autor, apesar de ter seus artigos tratando, majoritariamente, de "Artificial Inteligence", não é o que tem um maior alinhamento com este tópico no contexto global, mas sim o terceiro autor. Por seu turno, mesmo o primeiro autor tendo seus trabalhos menos relacionados com o tópico "Data Mining", dentre os autores avaliados, é o que apresenta trabalhos mais alinhados a este tópico. Considerando que cada autor registra em seus artigos a instituição de pesquisa na qual está vinculado e que cada instituição está localizada em um país, uma análise análoga à apresentada acima pode ser feita para instituições e países.

3.4 Interface de sumarização

Para que todas as análises propostas e apresentadas na seção anterior possam ser validades de uma forma rápida e eficiente, é necessário prover uma interface de visualização que sumarize de forma intuitiva os tópicos encontrados e suas associações com pesquisadores, instituições e países. Dessa forma, desenvolvemos uma *Web App* (disponível em: labpi.ufsj.edu.br/covpapers/ [2]) que apresenta os dados relacionados aos tópicos e suas correlações com artigos, pesquisadores, instituições e países por meio de um grande conjunto de metáforas visuais. Na Figura 3 ilustramos nossa aplicação com uma das várias metáforas visuais disponibilizadas, mais especificamente aquela que detalha os tópicos relacionados a um pesquisador.

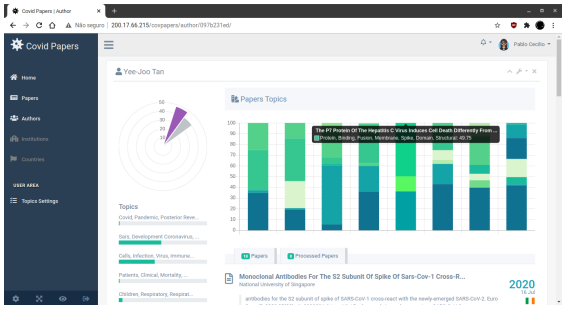

Figura 3: Visualização de um Autor na base do COVID

[2] username: user, password: covid-19

4 Resultados Experimentais

4.1 Coleção de Dados

Validamos o *framework*, considerando um *dataset* de artigos e trabalhos publicados sobre o COVID-19, SARS-CoV-2 e outros coronavírus do mundo, disponível no Kaggle [1]. Consideramos apenas trabalhos escritos em inglês publicados no ano de 2020, resultando em mais de 16 mil artigos, para os quais extraímos títulos e *abstracts* para uma análise específica sobre a COVID-19.

4.2 Correlacionando tópicos e autores

Inicialmente, um pesquisador poderia verificar como a ferramenta proposta define suas áreas de pesquisa, ou com quais tópicos sua pesquisa mais se correlaciona. Para ilustrar essa análise, consideraremos a pesquisadora holandesa Marion Koopmans, que é virologista com especialização em epidemiologia molecular na Erasmus University Medical Center em Rotterdam, Holanda. Na Figura 4 apresentamos a distribuição de tópicos para seus quatro artigos disponíveis no nosso *dataset* (barras do gráfico da direita), bem como uma análise global da pesquisadora (gráfico da esquerda). Podemos observar a pesquisadora tem sua pesquisa voltada em 34% para o Tópico 1 (que trata de virologia de modo geral); 22% para o Tópico 2 (sistema imunológico); 14% para o Tópico 3 (impactos sócio-econômicos da pandemia). Essa associação, identificada automaticamente por nossa ferramenta, pode ser corroborada tanto pela descrição da própria autora em sua página institucional [3], bem como pela reportagem recente da revista científica Nature [5], na qual destaca uma lista de pesquisadores que estão à frente da pesquisa sobre a origem da COVID-19, dentre elas Marion Koopmans.

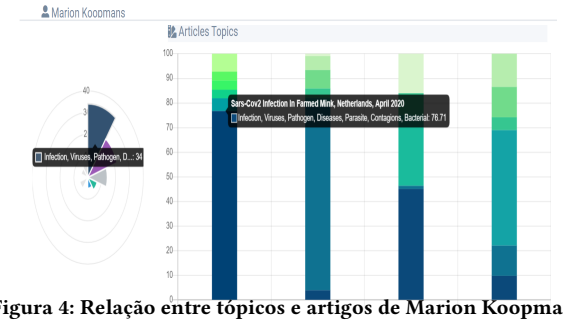

Figura 4: Relação entre tópicos e artigos de Marion Koopmans.

Um segunda possível análise é avaliar, dentre todos os autores, aqueles que possuem seus trabalhos mais alinhados a um determinado tópico. Retomando nosso exemplo, apesar da pesquisa de Marion Koopmans ter como principal enfoque o tópico 1, temos que, considerando nosso *dataset*, outro pesquisador se destaca nessa área. Trata-se do pesquisador japonês Hiroshi Nishiura. Na Figura 5 ilustramos a distribuição de tópicos de seu artigo intitulado "Incubation Period And Other Epidemiological Characteristics Of 2019 Novel Coronavirus Infections With Right Truncation: A Statistical Analysis Of Publicly Available Case Data", o qual tem 52.89% de seu conteúdo vinculado ao Tópico 1.

Em linhas gerais, não é somente um artigo que define, de maneira global, a associação de um pesquisador para um tópico de pesquisa em específico, todavia, vários trabalhos publicados e, majoritariamente, relacionados a um tópico de pesquisa, leva a ferramenta

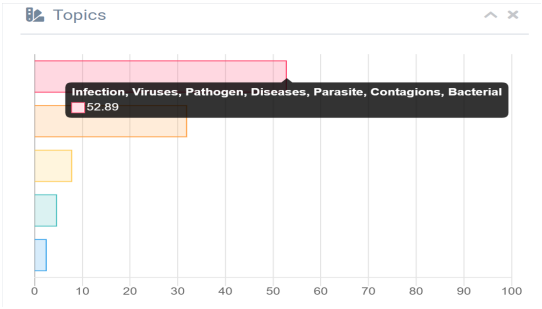

Figura 5: Tópicos Relacionados a um Artigo

inferir um maior correlacionamento do pesquisador para com este tópico - o que ocorreu com Hiroshi Nishiura, que surge como o pesquisador mais alinhado ao tópico 1(labpi.ufsj.edu.br/covpapers/ [4]).

5 Conclusão e Trabalhos Futuros

Observamos recentemente um intenso trabalho de pesquisadores de todo o mundo para conter o avanço do COVID-19. Organizar semanticamente as informações fornecidas pelos artigos científicos publicados para esta família de vírus pode ser extremamente importante para apontar rumos de pesquisa mais promissores, identificar abordagens ainda pouco exploradas, bem como sugerir possíveis colaborações de pesquisadores. Apresentamos um novo *framework* orientado a artigos que visa extrair, automaticamente, tópicos semânticos de publicações científicas relacionadas às pesquisas sobre COVID-19. É disponibilizada uma interface de sumarização que fornece uma visualização intuitiva para todas as análises propostas sobre um *dataset* composto de artigos publicados sobre COVID-19 [1]. Esta demonstrou que o nosso *framework* é capaz de extrair automaticamente características relevantes dos artigos, identificando as principais temáticas por eles abordados, bem como a correlação de pesquisadores, instituições e países para diferentes tópicos das pesquisas sobre COVID-19. Como trabalho futuro, pretendemos propor e avaliar métodos mais dinâmicos e interativos na interface de visualização, bem como aplicá-las a outras bases de dados em trabalhos científicos.

Referências

[1] Allen Institute For AI. 2020. COVID-19 Open Research Dataset Challenge. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
[2] S. Fathalla, S. Vahdati, S. Auer, and C. Lange. 2018. Metadata Analysis of Scholarly Events of Computer Science, Physics, Engineering, and Mathematics. In *TPDL*.
[3] Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791.
[4] Washington Luiz, Felipe Viegas, Rafael Alencar, Fernando Mourão, Thiago Salles, Dárlinton Carvalho, Marcos Andre Gonçalves, and Leonardo Rocha. 2018. A Feature-Oriented Sentiment Rating for Mobile App Reviews. In *Proceedings of the 2018 World Wide Web Conference*. 1909–1918.
[5] Smriti Mallapaty. 2020. Meet the scientists investigating the origins of the COVID pandemic. https://www.nature.com/articles/d41586-020-03402-1
[6] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes* 30, 1 (2007), 3–26. https://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad
[7] Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information Processing & Management* 50, 1 (2014), 04 – 112. http://www.sciencedirect.com/science/article/pii/S0306457313000964
[8] Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Luiz, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. CluWords: exploiting semantic word clustering representation for enhanced topic modeling. In *Proceedings of the Twelfth ACM WSDM*. 753.

[3] www.erasmusmc.nl/en/research/researchers/koopmans-marion

[4] nome de usuário: user, senha: covid-19