UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ
UNIOESTE - CAMPUS DE FOZ DO IGUAÇU
CENTRO DE ENGENHARIAS E CIÊNCIAS EXATAS
CURSO DE CIÊNCIA DA COMPUTAÇÃO

TCC - TRABALHO DE CONCLUSÃO DE CURSO

Proposta de Trabalho de Conclusão de Curso

Mineração de texto da literatura médica sobre a COVID-19

Jedson Gabriel Ferreira de Paula
Orientador: Rômulo César Silva

Foz do Iguaçu, 20 de Dezembro de 2021

1 Identificação

1.1 Ciência da Computação

Grande área: Ciência da Computação
Código: 1.03.00.00-7

Linha de Pesquisa: Matemática da Computação
Código: 1.03.02.00-0

Especialidade: Modelos Analíticos e de Simulação
Código: 1.03.02.02-6

1.2 Palavras-chave

1. mineração de conhecimento
2. covid-19
3. coronavírus

2 Introdução e Justificativa

Segundo informações do Ministério da Saúde (BRASIL, 2020), a COVID-19 é uma doença causada pelo coronavírus SARS-CoV-2, que apresenta um quadro clínico que varia de infecções assintomáticas a quadros respiratórios graves. De acordo com a Organização Mundial de Saúde (OMS), a maioria dos pacientes com COVID-19 (cerca de 80%) podem ser assintomáticos e cerca de 20% dos casos podem requerer atendimento hospitalar por apresentarem dificuldade respiratória e desses casos aproximadamente 5% podem necessitar de suporte para o tratamento de insuficiência respiratória (suporte ventilatório). Nos três primeiros meses de 2020, a pandemia desse coronavírus atigiu vários países, incluindo o Brasil. As maiores preocupações atuais dos governos de diferentes países é com o colapso nos sistemas de saúdes devido à sobrecarga de pacientes necessitando internação, e os prejuízos econômicos e sociais devido ao confinamento e restrição de deslocamento dos cidadãos.

A Mineração de Dados tem sido aplicada nas mais diversas áreas do conhecimento com o intuito de descobrir novas informações e subsidiar as tomadas de

decisões (BRAMER, 2007; AGGARWAL, 2015). Atualmente existem diversas ferramentas e frameworks desenvolvidos especialmente para a mineração de dados (data mining) e aprendizado de máquina (machine learning) tais como: Scikit-Learn, Weka, R, Python e Pandas (GERON, 2019). Uma das aplicações de mineração de dados é o Processamento de Linguagem Natural (do inglês NLP – Natural Language Processing) e a mineração de texto com o objetivo de mapeamento, extração de informação, entendimento de linguagem humana (natural) (DINOV, 2018). Também têm sido desenvolvidas diversas ferramentas open-source voltadas especificamente para mineração de textos tais como: Aika, Rapidminer Text Mining, Data Science Toolkit e KNIME (KAUR; CHOPRA, 2016). A mineração de texto examina grande volumes de texto não estruturado (corpus), auxiliando a extração de novas informações, descoberta de contexto, identificação de motivos linguísticos, ou a transformação do texto em formato de dados estruturado para derivar dados quantitativos que possam ser analisados futuramente (AGGARWAL, 2015).

Uma das principais justificativas a favor do uso de mineração de textos é a sobrecarga de informação devido ao grande volume de textos. Pode-se elencar entre as dificuldades para os humanos relacionadas a essa condição: profissionais se manterem atualizados com toda a literatura existente, encontrar informação precisa e relevante, sintetizar informação de fontes diversas e descobrir novos conhecimentos.

O dataset CORD-19 representa a mais extensa coleção de literatura sobre coronavírus, legível por máquina, disponível para mineração de dados até o momento (KAGGLE.COM, 2020). Isto representa uma oportunidade de aplicar abordagens de mineração de texto e dados para encontrar respostas a perguntas e conectar informações sobre esse conteúdo em apoio aos esforços contínuos de resposta ao COVID-19. Há uma crescente urgência para essas abordagens devido ao rápido aumento da literatura sobre o coronavírus, dificultando o acompanhamento da comunidade médica. O site Kaggle.com (2020), que promove competições/desafios em aprendizado de máquina, tem elencado algumas questões científicas importantes para mineração de texto no dataset CORD-19, extraídas de tópicos de pesquisa do SCIED (Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats) do NASEM (National Academies of Sciences, Engineering, and Medicine) nos EUA e da Organização Mundial da Saúde (OMS) para COVID-19.

A amplitude de trabalhos pertinentes ao assunto da pandemia atual que se encontram presentes no dataset, fazem necessário a exploraração e agrupamento de items de relevância ao que se quer entender sobre o tópico de interesse. Esta exploração se dá por forma de pesquisas, que podem ser de forma aberta como uma pergunta (e.g. *Qual o efeito da COVID em mulheres grávidas?*), ou também de forma mais especifica, como uma consulta (e.g. *"COVID" gravidez [efeito | impacto | sequelas]*), e possibilitar a realização de ambas formas de pesquisa é o que funda

o desenvolvimento de uma ferramenta e a escrita desta tese de conclusão.

3 Objetivos

3.1 Objetivo Geral

A partir de esforços anteriores do estudo e mineração dos textos presentes no dataset CORD-19 propõe desenvolver uma ferramenta de indexação, pesquisa e visualização fazendo uso de mineração de textos e aprendizagem de máquina com o objetivo de fornecer uma aplicação que facilita a pesquisa de textos acadêmicos, de interesse de ambos o público em geral e também da comunidade médica, seguindo critérios da OMS.

3.2 Objetivos Específicos

Dentre os principais objetivos específicos destacam-se:

- Pesquisa e revisão bibliográfica sobre técnicas gerais de mineração de textos e aprendizagem de máquina
- Estudar as características/informações gerais da base CORD-19
- Pesquisa e seleção de questões científicas relevantes referentes à COVID19 e o SARS-CoV-2
- Implementação de técnicas identificadas como adequadas à mineração de textos relacionadas às questões selecionadas
- Interpretação/avaliação dos resultados obtidos
- Construção da aplicação que utiliza das informações processadas (mineradas) para pesquisa e recomendação de artigos relacionados
- Redação de artigo científico para que o aluno possa desenvolver a capacidade de escrever de maneira científica sobre o próprio processo de pesquisa e os resultados obtidos

4 Plano de Trabalho e Cronograma de Execução

As atividades a serem desenvolvidas são:

1. Revisão bibliográfica sobre técnicas gerais de mineração de texto e aprendizagem de máquina e respectivas ferramentas open source (Python, Java, Scikit-Learn, R, Rapidminer Text Mining, Data Science Toolkit e KNIME);

2. Estudo das características gerais da base CORD-19 e questões científicas relevantes referentes à COVID-19 e o SARS-CoV-2;

3. Implementação da ferramenta proposta para mineração de texto na base CORD-19;

4. Análise dos resultados obtidos pelos algoritmos de mineração de textos e aprendizado de máquina usados.

5. Elaboração e apresentação no tutorial de TCC (Trabalho de conclusão de cuso);

6. Construção da aplicação para pesquisa e recomendação de artigos relacionados;

7. Redação do artigo descrevendo o processo de pesquisa e do resultado obtido;

8. Redação da monografia para submissão a banca;

Na Tabela 1 é apresentado o cronograma de atividades a ser seguido.

	Período									
Atividades	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	jul	ago
1 - Revisão Bibliográfica	•	•	•	•	•	•				
2 - Estudo da base CORD-19			•	•	•					
3 - Implementação da ferramenta			•	•	•	•				
4 - Análise dos resultados					•	•				
5 - Elaboração e apresentação no tutorial de TCC					•	•	•	•		
6 - Construção da aplicação					•	•	•	•		
7 - Redação do artigo					•	•	•	•		
8 - Redação da monografia						•	•	•	•	

Tabela 1: Cronograma das Atividades

5 Material e Método

Para desenvolvimento deste projeto serão usados os materiais:

- Computador Desktop 16 GB de RAM, 2 TB HDD, 500 GB de SSD, AMD 3600, Placa de vídeo Geforce 1660;

- Linguagem de programação javascript, python e suas bibliotecas científicas (Numpy, Pandas, Scikit-Learn);
- Ferramentas open source voltadas especificamente para mineração de textos (Aika, Rapidminer Text Mining, Data Science Toolkit e KNIME);
- Quanto à metodologia de desenvolvimento de software será adotado modelo iterativo e incremental (PRESSMAN; MAXIM, 2016).

A pesquisa bibliográfica será desenvolvida com fundamento em (LAKATOS; MARCONI, 2001) nas seguintes etapas:

1. Levantamento bibliográfico preliminar;
2. Busca das fontes;
3. Leitura do material;
4. Redação do texto;

Sendo 1 listado na síntese bibliográfica deste documento. 2 será feito e documentado posteriormente na redação do texto. Todas estas etapas incorporarão a atividade 1 citada no cronograma 1.

Os dados que serão utilizados durante o processo de mineração são os encontrados no dataset do Kaggle (KAGGLE.COM, 2020), nele contém a literatura médica sobre a doença e vírus o qual a comunidade médica possui interesse de estudo nos dias atuais.

O público que este texto está focado é o de Ciência da Computação pois as técnicas serão discutidas em termos de eficiência algoritmica para eficácia de classificação, já que o objeto a ser manipulado exige entendimento especializado. A aplicação que será desenvolvida já é para o uso da comunidade médica sendo levado em consideração a facilitação de pesquisa e exploração dos tópicos incorporados pelos textos acadêmicos.

A manipulação do texto será feita da seguinte forma: seus abstracts incorporados ao corpo do texto, mantendo uma refêrencia da sua origem para uso posterior; serão limpos e organizados os registros; e serão aplicadas as ferramentas estudadas no período da elaboração da revisão.

O resultado será analisado em termos de coerência dos clusters a partir dos tópicos incorporados nos textos.

6 Critérios de Avaliação

A partir da utilização de modelos de transformação de frases para vetores de alta dimensionalidade terá como produto um extenso indíce de valores a ser percorrido toda vez que se for feita alguma consulta. em contraparte, se a pesquisa for feita em forma de consulta, terá de ser percorrido o conteúdo da corpora a fim de encontrar os textos acadêmicos que satisfazem a consulta.

Sabendo disso, uma métrica apropriada para a ferramenta é a velocidade de extração dos items de maior proximidade com a consulta feita.

7 Referências

AGGARWAL, C. C. *Data mining: the textbook*. Suíça: Springer, 2015. Citado na página 3.

BRAMER, M. *Principles of data mining*. Londres: Springer, 2007. v. 180. Citado na página 3.

BRASIL, M. erio da S. *O que é COVID-19*. 2020. Disponível em: <https://coronavirus.saude.gov.br/sobre-a-doenca#o-que-e-covid>. Citado na página 2.

DINOV, I. D. *Data science and predictive analytics: Biomedical and health applications using R*. Suíça: Springer, 2018. Citado na página 3.

GERON, A. *Mão à Obra Aprendizado de Máquina com Scikit-Learn & TensorFlow – Conceitos, Ferramentas e Técnicas para Construção de Sistemas Inteligentes*. Rio de Janeiro: Alta Books, 2019. Citado na página 3.

KAGGLE.COM. *COVID-19 Open Research Dataset Challenge (CORD-19) An AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House*. 2020. Disponível em: <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks>. Citado 2 vezes nas páginas 3 e 6.

KAUR, A.; CHOPRA, D. Comparison of text mining tools. In: IEEE. *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. India, 2016. p. 186–192. Citado na página 3.

LAKATOS, E. M.; MARCONI, M. de A. *Metodologia do trabalho científico: procedimentos básicos, pesquisa bibliográfica, projeto e relatório, publicações e trabalhos científicos*. [S.l.: s.n.], 2001. Citado na página 6.

PRESSMAN, R.; MAXIM, B. *Engenharia de Software*. 8. ed. Brasil: McGraw Hill, 2016. Citado na página 6.

8 Síntese Bibliográfica

IGUAL, L.; SEGUÍ, S. *Introduction to Data Science*. Suíça: Springer. 2017 Nenhuma citação no texto.

SKIENA, S S. *The data science design manual*. Suíça: Springer. 2017 Nenhuma citação no texto.

ALPAYDIN, E. *Introduction to machine learning*. 2. ed. London, England: MIT Press. 2020 Nenhuma citação no texto.