Universidade do Estado do Amazonas
Escola Superior de Tecnologia
Pós-graduação Lato Sensu em Ciência de Dados
Programação para Ciência de Dados

12 de setembro de 2020

Defesa: 10/10/2020 *Trabalho Prático*

Especificação do Trabalho

1 Contextualização

O surto do COVID-19 na China foi noticiado em dezembro de 2019 [8]. A Organização Mundial de Saúde (OMS) declarou estado de emergência devido ao rápido espalhamento do vírus no mundo. Na América Latina, o Brasil é o país mais afetado pela doença. De acordo com o relatório pela OMS [10], houve o registro de 347,398 casos de infectados e de 22,013 mortes no Brasil.

Devido à rápida propagação da doença no mundo, as plataformas de mídias sociais como Twitter, Facebook e Instagram tornaram-se locais onde ocorre uma intensa e contínua troca de informações entre órgãos governamentais, profissionais da área de saúde e o público em geral. Um representativo número de estudos científicos têm mostrado que as mídias sociais podem desempenhar um papel importante como fonte de dados para análise de crises e também para entender atitudes e comportamentos das pessoas durante uma pandemia [6, 5, 3].

Com o objetivo de auxiliar o monitoramento da saúde pública e também para dar suporte a tomada de decisão de profissionais, diversos sistemas de monitoramento são desenvolvidos para classificar grandes quantidades de dados oriundos das mídias sociais. Estes dados podem ser empregados para identificar rapidamente os pensamentos, atitudes, sentimentos e tópicos que ocupam as mentes das pessoas em relação à pandemia do COVID-19 [1]. A análise sistemática desses dados pode ajudar os governantes, profissionais da saúde e o público em geral a identificar questões que mais lhes interessam e tratá-las de maneira mais apropriada.

Dentre as plataformas de mídias sociais, o Twitter é uma das mais populares. De acordo com [4], existe aproximadamente 200 milhões de usuários registrados nesta plataforma e que publicam mais de 500 milhões de tuítes diariamente. Portanto, pode-se aproveitar desse alto volume e troca frequente de informações para se conhecer as dúvidas sobre determinadas doenças. Como exemplo de importância desta plataforma em situações de crise, a atual pandemia de COVID-19 foi primeiro comunicada para a populacão na China através do site Weibo, que é o correspondente ao Twitter na China, antes mesmo do pronunciamento oficial das autoridades locais. Recentemente, existe um grande interesse de pesquisadores investigando o uso do Twitter para pesquisas relacionadas à saúde pública [1, 6, 9, 7].

Considerando o contexto descrito acima, o objetivo das equipes consiste em analisar as mensagens trocadas por usuários do Twitter sobre a COVID-19, com o intuito de realizar uma análise exploratória e visualização de dados, passando também pelas etapas de limpeza e organização. Mais especificamente, as equipes devem explorar a análise nas perguntas (questões) dos usuários, pois arguimos que seja um tipo de mensagem apropriado para se compreender as principais dúvidas das pessoas sobre a atual pandemia. Os dados já foram coletados e serão disponibilizados para os alunos.

As equipes serão compostas de 2 alunos e devem trabalhar de maneira organizada e sistemática para produzir os seguintes artefatos:

- **Relatório Técnico**. Um notebook com o texto e o código-fonte apresentado na ferramenta Google COLAB. Este texto deve ser construído na forma de parágrafos, utilizando um encadeamento lógico, com figuras e tabelas que a equipe considerar apropriadas para esclarecer os questionamentos avaliativos. O código deve estar organizado em blocos (células) de maneira a facilitar o entendimento.
- **Repositório no GitHub**. De caráter público (privado durante o desenvolvimento, público na ocasião da entrega), incluindo todos os integrantes da equipe e contendo as evidências de código e de progresso, registrando o trabalho dos integrantes;

2 Detalhamento das Atividades

A seguir, será apresentada a especificação do trabalho e detalhadas as principais atividades.

2.1 Base de Dados

A base de dados a ser utilizada pelas equipes encontra-se disponível no site do professor[1]. Esta base de dados contém cerca de 1.6 milhão de tuítes coletados durante o primeiro semestre de 2020. Maiores detalhes sobre a base de dados podem ser encontradas em [2].

2.2 Visão Geral dos Dados

O primeiro passo que as equipes devem efetuar após a importação da base de dados consiste em identificar as perguntas (questões) publicadas pelos usuários nas mensagens de Twitter. A mensagem completa está na coluna `texto` da base de dados. Esta coleção com a identificação das perguntas irá gerar um novo *dataset* chamado de "*DuvidasDB*". Em seguida, as equipes devem preparar um relatório para apresentar uma visão geral dos dados:

1. Devem apresentar um resumo (sumário) com as estatísticas dos dados originais, ou seja, sem qualquer pré-processamento. A apresentação deste tipo de informação é relevante para que

[1] `http://tiagodemelo.info/datasets/dados-curso-completo.csv.tar.gz`

outras pessoas possam ter uma visão geral dos dados. As estatísticas podem ser apresentadas através de tabelas e/ou gráficos.

2. As mensagens foram pré-processadas para que as perguntas (questões) fossem identificadas. Essa coleção de perguntas corresponde ao *dataset* "*DuvidasDB*". A seguir, a equipe deve apresentar as estatísticas sobre esses dados.

2.3 Temas Discutidos

Deve-se fazer uma análise dos temas discutidos nas perguntas que formam o *dataset DuvidasDB*. As perguntas dos usuários podem envolver diversos tópicos. Por exemplo, a pergunta "*Qual o risco do coronavírus chegar no carnaval no Brasil?*" é uma pergunta específica sobre a doença, enquanto que a pergunta "*Posso usar chá de alho para o tratamento da COVID-19?*" é sobre um possível tratamento para doença.

As equipes devem fazer uma análise sobre os temas que são debatidos nas perguntas dos postadas pelos usuários. Faça uma análise se as perguntas são relativas aos seguintes temas:

a) Doença. Quando a pergunta é relativa à doença. Deve-se observar que a doença é identificada por vários nomes. Exemplo: coronavírus, corona, COVID-19, etc.

b) Medicamento. Quando a pergunta é sobre o uso de determinado medicamento no tratamento da doença.

c) Organizações. Quando a pergunta é relativa a uma determinada entidade ou organização. Emissora de TV, Ministério da Saúde ou empresas são exemplos de organizações.

d) Pessoas. Quando a pergunta é sobre determinada pessoa. Por exemplo, a pergunta pode ser sobre a atuação que determinado político ou pessoa famosa teve durante esse período de pandemia.

2.4 Visão Temporal

Deve-se fazer uma análise temporal das perguntas que formam o *dataset DuvidasDB*. Pode-se considerar o intervalo temporal de dias, semanas ou meses. A escolha do intervalo de tempo ficará a cargo das equipes. Exemplos de análise temporal: a) houve um aumento no número de perguntas ao longo do tempo? b) houve uma mudança no perfil das perguntas ao longo do tempo?

2.5 Visão Geográfica

Deve-se fazer uma análise geográfica (espacial) das perguntas que formam o *dataset DuvidasDB*. Existem algumas colunas no *dataset* que trazem a informação das localizações como, por exemplo, o país, estado e cidade. Em alguns tuítes é possível ainda identificar as coordenadas geográficas de latitude e longitude. Exemplo de análise geográfica: a) os usuários de regiões diferentes fazem

perguntas com diferentes focos? Por exemplo, será que os usuários de uma região perguntam mais sobre a doença ou sobre o tratamento? Essa análise ainda pode ser realizada em diversos níveis de área (cidade, estado ou região). Além de apresentar a distribuição das dúvidas dos usuários por região, a equipe deverá fazer uma análise dessa distribuição. Por exemplo, apresentar as razões (ou hipóteses) da ocorrência dessa distribuição.

2.6 Tecnologias e Sugestões

Para a realização desta tarefa, é obrigatório o uso da linguagem de programação Python 3 (e superiores) e das bibliotecas Pandas e NumPy. Para exibição dos gráficos, as equipes podem usar as bibliotecas MatPlotLib[2] ou Searborn[3]. Outras bibliotecas complementares também podem ser usadas, especialmente na análise exploratória.

3 Critérios de Avaliação

Os critérios de avaliação leverão em conta a organização do repositório, a qualidade do código produzido, a completude das tarefas solicitadas, a documentação, a quantidade textual do relatório em termos de utilização da norma culta, coesão, coerência, o respeito aos prazos e a colaboração da equipe na elaboração do projeto.

A apresentação do trabalho será ainda avaliada pelos demais colegas da turma. Cada aluno deverá avaliar e dar uma nota aos trabalhos dos demais colegas. A média aritmética dessas notas irá compor 20% da nota final e a nota professor irá compor 60% da nota final. Os 20% restantes serão avaliados pela participação do aluno em sala de aula (conforme descrito no plano de ensino).

4 Datas

O trabalho deverá ser enviado para o professor até o dia 09 de outubro através do Google Classroom. A defesa (apresentação) do trabalho será no dia 10 de outubro.

Referências

[1] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah. Top concerns of tweeters during the covid-19 pandemic: infoveillance study. *Journal of medical Internet research*, 22(4):e19016, 2020.

[2] T. de Melo and C. M. Figueiredo. A first public dataset from brazilian twitter and news on covid-19 in portuguese. *Data in Brief*, page 106179, 2020.

[2] https://matplotlib.org
[3] https://seaborn.pydata.org

[3] H. Du, L. Nguyen, Z. Yang, H. Abu-Gellban, X. Zhou, W. Xing, G. Cao, and F. Jin. Twitter vs news: Concern analysis of the 2018 california wildfire event. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 207–212. IEEE, 2019.

[4] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep. Classification of tweets data based on polarity using improved rbf kernel of svm. *International Journal of Information Technology*, pages 1–16, 2020.

[5] E. H.-J. Kim, Y. K. Jeong, Y. Kim, K. Y. Kang, and M. Song. Topic-based content and sentiment analysis of ebola virus on twitter and in the news. *Journal of Information Science*, 42(6):763–781, 2016.

[6] L. Li, Q. Zhang, X. Wang, J. Zhang, T. Wang, T.-L. Gao, W. Duan, K. K.-f. Tsoi, and F.-Y. Wang. Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. *IEEE Transactions on Computational Social Systems*, 7(2):556–562, 2020.

[7] Q. Liu, Z. Zheng, J. Zheng, Q. Chen, G. Liu, S. Chen, B. Chu, H. Zhu, B. Akinwunmi, J. Huang, et al. Health communication through news media during the early stage of the covid-19 outbreak in china: Digital topic modeling approach. *Journal of medical Internet research*, 22(4):e19118, 2020.

[8] M. Malta, A. W. Rimoin, and S. A. Strathdee. The coronavirus 2019-ncov epidemic: Is hindsight 20/20? *EClinicalMedicine*, 20, 2020.

[9] C. Ordun, S. Purushotham, and E. Raff. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*, 2020.

[10] W. H. Organization et al. Coronavirus disease 2019 (covid-19): situation report, 126. 2020.