UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ (UTFPR)

Ciência de Dados 2 (CASI009)

Atividade 03 - Projeto

- **Tema:**
  - Título: Análise da comunicação das prefeituras do Brasil e a relação com o aumento da COVID-19
  - Descrição: O presente trabalho pretende extrair dados dos perfis das prefeituras do Brasil na rede social Twitter, focando nas capitais e comparar com o número de casos por município, analisando se a comunicação feita pelas prefeituras têm alguma relação no aumento de casos da COVID-19.
- **Equipe:**
  - Nome: *In Data We Trust*
  - Integrantes:
    - Jackson Cardoso (Mestrado Computação, UTFPR)
    - Osmary Camila Bortoncello Glober (Mestrado Computação, UTFPR)
- **Perguntas de pesquisa:**
  1. Prefeituras que fazem mais tweets sobre a COVID-19 tem um número menor de casos confirmados/óbitos?
  2. Existe uma diferença de mortalidade/casos confirmados de COVID-19 em cidades que se comunicam mais no Twitter do que aquelas que não se comunicam?
  3. Quais os tópicos mais utilizados pelas prefeituras nos tweets quando a pandemia teve o maior número de casos confirmados/óbitos?
  4. Quais os tópicos mais utilizados pelas prefeituras nos tweets quando houve a diminuição dos casos confirmados/óbitos?
  5. Qual a evolução dos tópicos dos tweets das prefeituras mensalmente e estes tópicos se relacionam com uma alta/baixa de número de casos confirmados de COVID-19?
  6. Prefeituras que têm um maior engajamento no Twitter teriam menor número de casos confirmados/óbitos?
- **Hipóteses:**
  - Prefeituras que fazem mais tweets sobre a COVID-19 tem número menor de casos confirmados/óbitos?
    - Hipótese: Quando as prefeituras utilizam o twitter como uma ferramenta de comunicação efetiva contra COVID-19 compartilhando medidas de prevenção e informações sobre a doença, isto leva a diminuição dos casos confirmados/mortes.
  - Existe uma diferença de mortalidade/casos confirmados de COVID-19 em cidades que se comunicam mais no Twitter do que aquelas que não se comunicam?
    - Hipótese: As prefeituras que possuem uma taxa de mortalidade/casos confirmados utilizam o twitter de maneira eficiente na prevenção de casos confirmados/óbitos.

- Quais os tópicos mais utilizados pelas prefeituras nos tweets quando a pandemia teve o maior número de casos confirmados/óbitos?
  - Hipótese: Os seguintes termos foram mais utilizados durante o pico da pandemia: álcool em gel, máscara, prevenção, vacina, distanciamento social, lockdown.
- Quais os tópicos mais utilizados pelas prefeituras nos tweets quando houve a diminuição dos casos confirmados/óbitos?
  - Hipótese: Houve permanência das prefeituras nos tweets de prevenção e informação sobre a COVID mesmo após a queda dos casos confirmados/óbitos.
- Qual a evolução dos tópicos dos tweets das prefeituras mensalmente e estes tópicos se relacionam com uma alta/baixa de número de casos confirmados de COVID-19?
  - Hipótese: prefeituras com mais tópicos relacionados com a COVID-19 tiveram um número de casos confirmados/óbitos menor do que aquelas que não falam sobre covid mes a mes.
- Prefeituras que têm um maior engajamento no Twitter teriam menor número de casos confirmados/óbitos?
  - Hipótese: Sim, teriam.

- **Dados e modelos:**

a) Dados:

As variáveis da pesquisa são: data da criação do tweet, texto do tweet, quantidade de like do tweet, quantidade de retweet do tweet, quantidade de reply do tweet, número de casos confirmados de COVID-19, quantidade de óbitos por COVID-19, percentual de letalidade de COVID-19; número de casos confirmados de COVID-19 por 100 mil habitantes, quantidade de óbitos de COVID-19 por 100 mil habitantes, população estimada do município, município, UF, região brasileira, tipo do municipio: capital/cidade.

Os dados principais são oriundos da rede social Twitter, com limite de 3200 tweets por perfil da prefeitura, do tweet mais novo para o antigo, resultando em 72 mil tweets após a aplicação de limpeza de dados.

A pesquisa foi embasada nos dados dos boletins de COVID-19, oriundos do portal Brasil.io, focando nos 100 municípios com maior número acumulado de óbitos, sendo que apenas 90 prefeituras possuem perfil na rede social Twitter.

Para a pesquisa também foram obtidos dados auxiliares por meio do portal do IBGE (Instituto Brasileiro de Geografia e Estatística) sobre a população residente estimada de cada município brasileiro.

b) Modelos:

- Implementação de LSI na análise temporal de tópicos nos tweets relacionados ou não à COVID-19;
- Implementação do classificador de polaridades para análise de sentimentos dos tweets;
- Implementação de análise dos N-gramas dos tweets.

- **Trabalhos correlatos:**

CAVALCANTE, Marcella; SILVA, Maria H.; GAMELEIRA, Pedro H. **UTILIZAÇÃO DA INTELIGÊNCIA ARTIFICIAL EM BIG DATA NA MEDICINA DURANTE A PANDEMIA DE COVID-19.** Semana de Pesquisa do Centro Universitário Tiradentes, 2020. Disponível em: < https://eventos.set.edu.br/al_sempesq/article/view/13786>. Acesso em: 25 abr. 2021.

FILHO, Francisco R. da Silva; COUTINHO, Emanuel F. **Uma Análise de Tweets sobre o Coronavírus.** Revista Sistemas e Mídias Digitais, 2020. Disponível em: < https://revistasmd.virtual.ufc.br/arquivos/volume-5/numero-1/rsmd-v5-n1-3.pdf>. Acesso em 27 abr. 2021.

MELO, Thiago de; FIGUEIREDO, Carlos M.S. **A first public dataset from Brazilian twitter and News on COVID-19 in Portuguese.** Elsevier, 2020. Disponível em: < https://www.sciencedirect.com/science/article/pii/S2352340920310738>. Acesso em: 25 abr. 2021.

ORTIZ, Junia; BROTAS, Antonio M.P.; MASSARANI, Luisa. **Ciência e COVID-19 no Brasil: a repercussão das decisões da OMS no Twitter.** Revista Latinoamericana de Comunicación, 2020. Disponível em: <https://revistachasqui.org/index.php/chasqui/issue/download/Chasqui%20145/155#page=47 6>. Acesso em: 25 abr. 2021.

XAVIER, Clarissa C.; SOUZA, Marlo. **Extração e Classificação de Dados Semânticos do Twitter.** Anais do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web: Minicursos, 2020. Disponível em: <https://sol.sbc.org.br/livros/index.php/sbc/catalog/download/8/16/55-1?inline=1>. Acesso em: 28 abr. 2021.

- **Cronograma:**

Para a elaboração do cronograma seguiu-se o ciclo de vida para um projeto de Ciência de Dados, conforme a Figura 1.

Figura 1 - Ciclo de vida de um projeto de Data Science

Desse modo, tem-se abaixo o cronograma de atividades a serem desenvolvidas e as respectivas datas de entrega a serem cumpridas:

1. Escolha do Tema, a ser realizada por todos os membros até 26/03/2021;
2. Análise exploratória, a ser realizada por todos os membros até 14/04/2021;
3. Projeto, a ser realizado por todos os membros até 28/04/2021:
   3.1 Análise de N-gramas dos tweets até 02/05/2021;
   3.1 Análise de polaridades dos tweets até 05/05/2021;
   3.2 Análise temporal dos tópicos dos tweets até 09/05/2021.
4. Pesquisa, a ser realizada por todos os membros até 12/05/2021;
5. Resultados finais, a serem realizados por todos os membros até 19/05/2021.