Coronavirus Tweets Analysis-Project Proposal

Xufei Li

Question/need:

- What is the framing question of your analysis, or the purpose of the model/system you plan to build?

  Using Topic modeling to categorize top topics that people discuss the most for each country for each time slot, then apply sentiment analysis and clustering for our preprocessed data to separate posts then give them labels.

- Who benefits from exploring this question or building this model/system?

  Governments may benefit from this analysis as they know what the priority issues are over the time. Then they can take action on it.

Data Description:

- What dataset(s) do you plan to use, and how will you obtain the data?

  Coronavirus tweets NLP data & get most recent tweets from twitter API for one topic(eg: vaccination in US)

- What is an individual sample/unit of analysis in this project? What characteristics/features do you expect to work with?

  Each tweet(document).

- If modeling, what will you predict as your target?

  Unsupervised ML, no target. (I will drop the label from my dataset)

Tools:

- How do you intend to meet the tools requirement of the project?

  Python packages for data manipulation, EDA, modeling, visualization.

- Are you planning in advance to need or use additional tools beyond those required?

  Maybe Tableau for visualization

MVP Goal:

- What would a minimum viable product (MVP) look like for this project?

  Be able to find the top topics each country discusses the most. Clustering data to different groups then gives them labels - “positive”/ “negative”.