MSDS 697 - Group Project

1. **Topic/Title** - Reddit Clustering and Sentiment Analysis of World News and Events
2. **Team Members** - Joren Libunao, Anirav Jain, Daniel Tinoco, Zemin Cai, Theo (Byunghyun) Kim
3. **Data Sources (URL)**:
    a. Reddit API: https://www.reddit.com/dev/api (can interact with this better using PRAW, or Python Reddit API Wrapper: https://praw.readthedocs.io/en/stable/#getting-started)
    b. RapidAPI (Reddit): https://rapidapi.com/socialminer/api/reddit34
4. **ML Objectives**:
    a. Objective #1: Data collection, preprocessing and feature engineering using packages like Spacy and Regex
    b. Objective #2: Clustering topics and stories by similarity across various Subreddits
    c. Objective #3: Build a word cloud using topic segregated data
    d. Objective #4: Sentiment analysis of the topics and organizing into positive, negative, and controversial lists
    e. Objective #5: Build a model that can create a summary of news and changes of sentiments for a given period of time in natural language
5. **Timeline**:
    a. Week of Feb 5th to Feb 11th: Familiarize ourselves with the Reddit API and PRAW and NLTK, Spacy, Regex, and other NLP libraries etc.
    b. Week of Feb 12th to Feb 18th: Do a basic summary of posts and/or comments within the past month for the subreddits you're assigned. Begin constructing a data pipeline.
    c. Week of Feb 19th to Feb 25th: Finish constructing data pipeline and validate the data output.
    d. Week of Feb 26th to Mar 4th: Begin working on objectives #2-5.
    e. Week of Mar 5th to Mar 10th: Finish objectives #2-5 and refine the output, and write the final report.
6. **Final goal**:
    a. Clustering, sentiment analysis of trending/popular stories across various subreddits (see subreddits list on next page). Build a word cloud to identify popular topics.
    b. At least Level 2. Can push to Level 3 or 4 if the project progresses naturally.

Team meeting notes:

- Sentiment analysis of trending/popular stories on Reddit:
  - Generate trending topics and cluster into lists representing positive, negative, neutral, and/or controversial (Reddit posts that are upvoted and downvoted a lot) public sentiment
  - Can also compare different versions of the same story on different subreddits (i.e. conservative vs liberal)
  - Subreddits: r/news, r/technology, r/conservative, r/politics, r/worldnews, r/environment, r/economics, r/Health, r/Coronavirus, r/science, r/entertainment