AC 221 Final Project Proposal
Siyuan (Tom) Zhang

Title

Digital Moods: Exploring the Connection between Twitter Activity and Mental Health

Objective

Investigate the relationship between Twitter usage and mental health by analyzing the sentiment and content of tweets. Identify trends and correlations between certain types of content and mental well-being.

Motivation

I have previously tried Twitter's API for another project, and I really wanted to use it again while I still have access. I also love natural language processing (NLP), and Twitter has a very large amount of text data to perform experiments on. Further, mental health is an important topic I care deeply about. There is plenty of research done regarding social media and mental health. Here, I would like to conduct my own analysis, with an emphasis on NLP techniques and implementing an entire pipeline from scratch.

Methodology

1. **Data collection**: Write a data collection script that interacts with Twitter's API. Run it on Harvard's HPC (the Cannon Cluster) to scrape all tweets from a specific sample of Twitter users and time period, say all US users from 2018 to 2019.
2. **Data processing**: Clean and preprocess the collected tweet data.
3. **Sentiment analysis**: Use NLP techniques to analyze the sentiment of the tweets. Identify overall sentiment trends and variations between user groups.
4. **Content analysis**: Build a topic model. Identify common themes and patterns in tweet content related to mental well-being.
5. **Statistical analysis**: Examine correlations and potential causal relationships between tweet content, sentiment, and mental health indicators.

Scope

Right now I am only dealing with a very specific sample of users. This could certainly be adjusted (either expand or reduce) based on the speed of the actual collection process. The script should run no longer than a week. The sentiment analysis and topic modeling part can be very open ended, implementation-wise. I will experiment with different models and compare their performances. In my analyses, the effects of user groups will also be examined. In particular, the granularity of how groups are defined can also be experimented on (e.g., socioeconomic status, geographic location, education). This can be achieved by joining external datasets such as the US Census Data.