CS410 Project Progress Report

1) Which tasks have been completed?

Currently, the completed tasks include learning how to use praw, the reddit API, as well as scraping the reddit threads for data and cleaning the data. The original plan was to scrape and use tweets from Twitter, but due to some challenges (mentioned under q3), I switched over to the reddit platform instead.

2) Which tasks are pending?

Pending tasks include data cleaning, performing sentiment analysis and testing.

3) Are you facing any challenges?

One of the challenges I faced was due to language. As I was scraping tweets regarding an international competition featuring a number of non-English speaking nations, understandably, a large number of tweets were in foreign languages. While there are a number of translation apps out there (Google Translate, etc.) from experience, they are not very accurate. While removing all the 'foreign' language tweets from my data could be an option, instead, I opted to swap to scraping data from Reddit instead. While there are users across the world on that platform, the platform tend to feature mostly English discussions.