Final Project Progress Report for CS 410, Fall 2022

Project Github: https://github.com/rylandtikes/cs410-course-project

Team Name - Mountain Group
Binod Pandey bpand3@illinois.edu
Charles Stolz (Captain) cstolz2@illinois.edu
Michael Gambino mgambin2@illinois.edu

1) Which tasks have been completed?

We have created the Reddit application to extract comments and headlines from the API. The application was written in Python and has been tested locally during our prototyping phase. Using this application we have created several datasets and uploaded them to Amazon S3. Our final datasets will be labeled and will reside in S3 so other application services we create can read them. In regards to data visualization, we have agreed upon the front end design and the underlying studies that will be answered by our data-driven figures. Once established, we have set up the necessary infrastructure on Heroku with Plotly Dash.

Work Breakdown Structure:
Complete but need to automate - Create a tool to pull and transform the data (Python)
Testing - Create tool for Sentiment Analysis (Python)
Testing - Create an app to display the results (Python with Plotly Dash and Heroku)

2) Which tasks are pending?

We are currently testing our Sentiment Analysis tool which uses the NLTK toolkit vader lexicon. We have 6 datasets labeled, ranging from extremely negative to extremely positive. The work to automate new datasets daily and to create a tool to display the data visualization are the current tasks being worked on. We are also researching sentiment classification from the emotions aspect.

3) Are you facing any challenges?

Currently we are making good progress on our project and are not facing any major challenges. The team is learning about Sentiment Analysis, news headlines and comments are more challenging than other datasets such as product reviews. One challenge we are facing is many of the top level Reddit comments are low quality and difficult to classify, headlines are higher quality so much of our analysis will be focussed on the headlines but we still plan to analyze both.