**Term Project – Project Proposal**

**The Big Idea**

The main idea of this project is to present metadata and other information of a popular social media platform (i.e. Reddit) by collecting information and presenting it in a visually aesthetic manner. This project will explore data mining/collection, provide the methodology and rational used to analyze that data, and present that data in an easily digestible manner. Most likely this project will use text and sentiment analysis to find the average Redditors reaction to certain topics, although additional possibilities are still being explored. The minimal viable product will have data collected in a set time period and show the analysis of that data alongside visuals on a website. A stretch goal would be to allow users to input specific subreddit/dates and have the analysis be performed through a server. The visualizations would also be more dynamic and interactive. For example, we may have a graph showing the sentiment score for all posts in a 24 hour period. Hovering your mouse/tapping on a point will provide a link to the original comment.

Alternatively I could collect data from twitter over a period of time about a specific topic or on a specific figure, and perform similar analysis.

**Learning Goals**

The main goal of this project is to develop my abilities in three areas: Data mining, data analysis, and data visualization using Python. A higher-level goal is to develop a mindset necessary for data analysis.

**Implementation Plan**

I will take advantage of several Python modules to aid me in the journey. Some potential modules:

* PRAW – for getting Reddit data
  + Tweepy – If the Reddit data is difficult to use I’ll just gather data from Twitter.
* Pandas – For manipulating the dataset
* NumPy – Only if I find numerical data I want to work with
* Nltk – Useful for text processing. Can preprocess text and also do sentiment analysis.
* Wordcloud – Might be interesting to use to explore text data.
* Potential packages for interactive visualizations:
  + Mpld3 – If I focus on small sets of data
  + Pygal – Plus I won’t need to worry about JavaScript
  + Bokeh – Additional graph options
  + Holoviews – Let’s me build data structures conducive to visualizations.
  + Plotly – Large range of graph types, can use SQL, can build custom controls.

The first goal is to be able to extract the type of data I want. I will use a module to gather information from either reddit or twitter.

The next goal is parsing that data – there will be numerical data that may need to be cleaned. If I extend the scope of this project to text analysis, I will need to do text preprocessing.

Once the data is cleaned and prepped, it will undergo analysis. I also intend to create some visualizations that will help in the analysis. Any interesting visualizations will be further expanded on for the next step.

Once the data is collected and cleaned, the information will be presented online. I intend to create a simple webpage/website (probably just on heroku or github pages) to display the information.

If these steps are finished and there is time left in the semester, I plan on making the visualizations more interactive and visually appealing. Furthermore, depending on the complexity of the first three steps, I may be able to automate the entire process so that users could input a specific subreddit/time frame and have the analysis be done real time. I will probably need some sort of backend/server to do this, so this is highly unlikely.

I will need to do additional research on the specific libraries and test them out to see if I will be able to learn to use them within the next 8 weeks.

**Project Schedule**

* Week 1: Become comfortable with data scraper, finalize where to get data from.
* Week 2: Clean and parse data.
* Week 3-5: Explore and analyze data. Learn to utilize necessary library. Continue to gather data if necessary.
* Week 6-8: Develop visuals based on analysis. Build simple website. If ahead on schedule, follow the reach goals listed above.

**Collaboration Plan**

I plan to keep the professor updated on my progress, and ask him any questions if I encounter any major problems. As I am working alone there really isn’t much to state here.

**Risks**

There are few risks I need to consider. First, I may find the data I collect is difficult to work with. I may need to redefine the scopes of the project, or collect data from another source. Second, analysis of the data may take longer than expected. I may need to limit my findings to only a few things to compensate. Third, I may have difficulty getting the visualizations I want. As the visualizations will be a major component of this project, as well as a huge personal learning goal, I will need to dedicate a significant amount of time to this portion. Finally, I need to ensure that I am not being too ambitious and spending too much time on one aspect of the project. A MVP is superior to having nothing.

**Additional Course Content**

Looking at the syllabus, the class sessions on text-analysis and data-analytics using Numpy and Pandas would be relevant to my project. Web scrapping would be useful if it happened earlier but since it is near the end of the semester it won’t help too much.