Project Proposal

1. The Big Idea

In our divided world today, accusations of media bias sully the political process and stifle intellectual progress through a perpetual focus on eradicating any emotion behind admonition. Many on both sides of the political aisle see the other side's news organizations as overly emotional and negative. We will put these accusations to the test. For this project we will be building a Python program that scrapes the web for news articles about Donald Trump from several discrete time periods, performs a vader sentiment analysis on them, and graphs the mean sentiment on three axes: positive, negative, and neutral. We will repeat this process for Barack Obama and then for other significant people on the political spectrum. We will then sort and visualize the data by media outlet, calculate the centroid of each one, and from then apply a k-nearest neighbors algorithm from which we hope to be able to classify an article from an unknown news source into its correct place of origin.

2. Learning Goals

Ben: To become proficient at data mining, data manipulation, and the basics of machine learning in Python.

Michelle: To create conclusions from large sets of data and visualize them through various forms of ML methods.

John Edwin: To learn the various techniques that exist in data mining and machine learning and explore all existent packages and libraries to that end.

3. Implementation Plan

1. Strategy: To identify individual members strengths, assigning them to lead the part of the project they are most excited about completing.
2. Research: Research news sources and compile a list of sources we wish to use to generate our data sets.
3. Implement:
   A. To build a scraper that extracts articles from various sources, parses the data, and vectorizes them.
   B. To run various forms of analysis through the methods of ML
   C. Visualize the data in 3D space
   D. Adapt, modify, add additional resources
   E. Visualize classification on a GUI

4. Project Schedule

Week 1 - **We are here!** Refine our project scope, develop team norms, and strategize the development of project.

Week 2 - Collection: Web scraper is built, x amount of news articles are pulled, parsed, and ready for manipulation. Team assesses strength of data collected, decides to continue forward or pivot to other topic.

Week 3 - Manipulation: The Big Kahuna-- this week is dedicated to creating and using the data collected to apply ML methods to begin to draw conclusions. Two methods should be employed and visualized

Week 4 - Documentation: Website is made and lessons learned, desired outcomes, and other hypothesis are documented.

Manipulation, cont.: Time is set aside in case week three presents obstacles that stand in the way of completion

Week 5 - Refinement: Begin working on additional project deliverables, such as poster and website is updated.
Week 6 - Refactoring: Project should be completed, this last week is dedicated to increasing optimization, speed, and refining the output.

5. Collaboration Plan

At the beginning of this project we plan to meet fairly often (two to three times per week outside of class) in order to develop a strong understanding of our collective strengths, and to put our heads together in order to develop a program architecture. For the heavy lifting portions of this assignment, we plan to work independently, still checking in once or twice per week in person outside of class to get a fresh look at any bugs or work through any roadblocks we might have encountered individually. We intend to use an agile development model, all working on different parts of the project at once, building an extensible architecture that will hold fast in the unlikely event of an unforeseen impassable obstacle.

6. Risks

Elementary understanding of project complexities
Rough time estimation
Conflicting priorities

7. Additional Course Content

Scrapy for web pages, basic operations with pandas, numpy, matplotlib, and scikit-learn