December 16: Question and Data Set, Elevator Pitch

What is the question you hope to answer? What data are you planning to use to answer that question? What do you know about the data so far? Why did you choose this topic?

- I would like to use NLP to classify content off of Wikipedia to understand what each page talks about, and then match it with its page views and scale it with all their articles to understand what topics perform well and are interesting.
- Wikipedia Data Set with 4 characteristics per page.
- I know that the data set is 800GB and it seems pretty messy with page titles such as “Barack_Obama,_Sr. and Barack_Obama_%22HOPE%22_poster”.
- I chose this topic because it falls underneath SEO and NLP, both topics which I’m very interested in.