Big Data Project Steps

**Overview**

* product features mining
  + how to extract a single label?
    - topic modelling? then keyword extraction?
* sentiment analysis by product features (split by sentences)
* identify which product features are important for high ratings and low ratings
  + will probably just do a weighted average of frequency and sentiment by number of stars
    - sum of all sentiment \* 1 across all instances
    - this would give the "frequency" of that aspect
    - then divide by denominator
  + better versions but less explainable
    - https://wing.comp.nus.edu.sg/~antho/P/P11/P11-1150.pdf
    - <http://www.projectsgoal.com/download_projects/data-mining/data-mining-projects-GDM00106.pdf>

**Steps**

* ~~Download Amazon Review dataset 150m+ observations~~
* ~~Subset to two categories and work with it in this state~~
* ~~First step is to extract product features for each product~~
  + ~~Use spacy to parse sentences~~
  + ~~In each review, extract noun chunks~~
  + ~~Extract just the noun and adjective~~
  + ~~Record sentences that have these noun chunks~~
* ~~I need to work with the product descriptions as well~~ 
  + ~~Will help with stuff like camera or battery life for phones~~
* Now I need to extract common feature names
  + Group by product category
  + Run clustering to group aspects
    - Identify optimal set of clusters dynamically using average silhouette or gap statistic
  + For each cluster take the word with the highest frequency or the noun phrase in the center
    - Or is there some heuristic that I can use to combine these?
  + This will be our actual label
* Sentiment analysis
  + Use spacy to run a sentiment analysis
  + Get sentiment scores by aspect (hopefully between 0 and 1)
* Calculate feature scores
  + Two methods:
    - Sum the sentiment scores for each feature by product
    - Divide by total sum of sentiment scores by product
    - For each product this gives a basic ranking of aspect importance
  + PageRank by category
    - Create graphs by product category
      * Categories as refined as possible
    - Within each graph compute the page rank for each feature
* Visualize

**Extract Features**

* I need to keep track somewhere the linguistic rules that I am following
* Future steps
  + Separate descriptors (e.g., “great”, “sturdy”) from noun chunks (“battery life”, “customer service”)
  + Make number of clusters dependent on observations? Or how else to measure variety?
    - We could introduce topic models by category, but we would need the product description