Senior Presentation Topic Modelling Notes

* Stance classification: for, against, or neutral
* Author’s stance in categories: social, environmental, or economic (financial)
* Better ways to group based on author’s opinion?
* Topic analysis algorithms – suite of algorithms to discover and annotate large archives of documents with topical information.
  + Utilized unsupervised learning.
  + Analysis of tweets
* Latent dirichlet allocation:
  + Discover hidden structure in words.
  + List of related words.
  + Scikit-learn and Gensim.
* Removed stop words, irrelevant words, and focused on Tweet text itself.
* Results were unclear – could not identify distinct topics
  + Gensim results better than Scikit-Learn
* Other methods of topic analysis?
  + Word2Vec – unsupervised algorithm.
    - Clustering of words that are related.
    - Clusters that can point to a topic.
    - Results were not great.
  + Word cloud
    - Based on each tagged Tweet.
    - In support
      * Financial, work, jobs, project partner, etc.
      * Seems to relate to TBL
    - Against
      * Climate, reef, protect, stop destroying environment, etc.
      * Seems to relate to TBL
    - Neutral
      * Don’t really point to anything interesting.
* TBL still most accurate method of categorization.
* LDA document too short, informal, and lack context.
  + Requires long documents in formal style.
  + Tweets have informal structures.
* Word2Vec Tweets are using too similar of words.
  + Tweets aren’t unique.
  + Similar words in for, against, or neutral – hence difficult to have unique cluster groups.

Senior Presentation SLO Stance Analysis Notes

* Stance analysis for SLO.
* Used around 10k of the 650k+ in the dataset.
* Test set – 200 examples.
* Stance versus sentiment.
  + Positive sentiment but against, for example.
* Used Bernoulli Naïve Bayes.
* Used Linear SVC.
* Used MLP – multi layer perceptron neural network.
  + Densely connected layers.
  + Hidden layers.
* Tokenized the Tweets – removed hashtags, vectorized categorical data, etc.
* Utilized grid searches.
* Top scores from the smaller networks.
  + Occam’s Razor – simplest solution is usually the right or best solution.
* Neural networks worked better – captures non-linearity.
* Look at sequences of characters instead of individual words to account for spelling errors or non-words.
* Possible improvements:
  + Explore combinations of networks.
  + Used bag-of-words and character n-grams.
  + Input layer of 18k – improve upon it using word2vec, etc.
  + How could apply to Calvin College’s Twitter data?
  + Custom-built neural networks from CSIRO – use those on the data.
  + Extend training/test sets to include more data.
  + Incorporated many companies into one dataset.
    - Instead, split into individual companies and their associated Tweets.
* Randomized permutations of training sets for each classifier.
* Testing data was hand-labeled.
  + 3 people labeled what they thought the stance were.
  + Compared classification between each 3 to find common agreement.