Metis Week 7 - Project Fletcher Proposal

Unsupervised Clustering of News Headlines

Steven Bierer, November 5, 2018

Data Resources and Tools

This project emphasizes natural language processing and unsupervised classification of text. I have chosen to analyze news headlines and articles about the U.S. economy, obtained from the website "Data For Everyone" (https://www.figure-eight.com/data-for-everyone/). The data consists of headlines and short snippets from thousands of articles, plus survey responses about whether a participant believes an article gave an indication, negative or positive, about the health of the U.S. economy. Additional news headlines and/or articles will be obtained from alternative sources, as needed, whether to bolster the training algorithm or to serve as additional (and potentially more challenging) data for testing the algorithm.

Analysis will be performed in Python using the standard modules *pandas*, *numpy*, and *skikit-learn*. Classification will be implemented using topic modeling algorithms from the model *gensim*. The modules *matplotlib* and *seaborn* will be used for graphical display. The NoSQL document database manager *MongoDB* will be used to store the large amounts of data. Other tools will be added as needed.

Clustering Approach

I plan to take the following steps with the data, though depending on outcomes this workflow will likely be altered. I will focus on headline text, but may add article text if needed for model robustness.

1) Import the data into a MongoDB collection on a remote AWS server.
2) Connect to the remote server via ssh using Jupyter Notebook.
3) Remove stop words (common words, like "the" and "is", with little meaning) from the training and test data.
4) Split the data into training (80%) and test (20%) sets.

5) Apply Latent Dirichlet Allocation (LDA), a statistical framework commonly used to generate models about word occurrences in text documents, to extract a discrete number of topics from the headlines in the training data. The number of topics may have to be adjusted.
6) Run the model on the test data and see if the model captures the topical essence of select headlines.

Prediction Modeling

Based on the unsupervised topic clustering of the headlines, I will train a classification model on whether or not the headline relates to the U.S. economy, as determined subjectively by survey respondents. The performance of a number of classifiers will be compared for this, including logistic regression, support vector machine, and decision trees with boosting. Using the best classifier, I will make predictions on the reserved test data set. I will also train and test a similar classifier model on whether economic headlines indicate negative or positive health of the U.S. economy.

Expectations

As natural language processing is a completely new field to me, I am not certain what to expect. I'm not sure, for instance, how much data will be needed for the LDA algorithm to converge to a meaningful model. However, I can use larger data sets from news media (e.g. millions of headlines from the Australian Broadcasting Corporation) or scrape headline data from the web, though this would sacrifice having labeled data with regard to the economic content of the articles.