Natural Language Processing of Economic News Articles

Project Fletcher - Weeks 7 and 8 of the Metis Data Science Bootcamp

Steven Bierer November 19, 2018

Introduction

For the fourth project, the data science class was encouraged to apply unsupervised clustering techniques on a natural language processing data set. I chose to examine newspaper articles and headlines about economic news, available from the "Data For Everyone" repository. News paper stories, and to a more acute degree headlines, must convey specific information to a reader in a limited amount of printed space. This makes newspapers a compelling medium from which to analyze textual language for meaning and sentiment. This particular data set consists of 8000 articles and accompanying headlines, from a variety of sources published between 1951-2014, along with impressions from surveyed readers regarding 1) whether the article is relevant to the United State's economy and 2) the tone of relevant articles, judged on a 1-9 scale with 1 being the most negative and 9 being the most positive.

Approach

The overall workflow of this project was as follows: 1) clean the document text in preparation for tokenization; 2) tokenize the text (i.e. parse into discrete words or groups of words); 3) create a dictionary and counts of all words; 4) generate underlying topics, representing probabilities of words that tend to group together in the documents; 5) determine the mixture of topics that best describes each document, effectively reducing the feature space from hundreds of words per document to a handful of topics; 6) train and test a classification model to predict the yes/no relevance indication; 7) train and test a regression model to predict the positivity score. The documents were tokenized into unigrams with the exception of some named entities (particularly known organizations like "The New York Stock Exchange") that were composed of multiple words.

It's worth noting that I originally intended to separately model the headline text in parallel with the article text. However, I discovered that many headlines were dominated by the name of the source (e.g. "The Wall Street Journal") or the name of the article's newspaper section (e.g. "The Morning Report"). Additionally, given the difficulty in forming robust predictions from the articles, I saw little reason to continue with the headlines, whose word corpus was much smaller.

The following tools and sources were used in the analysis:

> Data source: Data For Everyone (http://www.figure-eight.com./data-for-everyone/).
>
> Tools: 1) Python and the standard libraries *numpy*, *pandas*, and *matplotlib* for routine data manipulation and graphical display; the regular expression module, *re*, for finding string patterns. 2) Python modules *spacy, gensim*, and *nltk* for NLP analysis and *statsmodels* for classification and regression.
>
> Concepts learned and applied: Natural language processing (including tokenization, lemmatization, and named entity recognition), topic modeling (latent dirichlet allocation), unsupervised classification, and sparse matrix representation. Additionally, a prior oversize-file problem on the version-control cloud system Git was resolved and, for the first time, the code was developed on a temporary branch of the project's Git repository (merged at project completion with the master branch).

Token Creation

After cleaning the data of unrelated text, tokenization was performed using the *spaCy* module. This consisted of several steps. First, each document was divided into single-word strings (unigrams) in lowercase. Unigrams not contained in a list

of "stop words" (based on the the *NLTK* English language corpus) *and* matching an accepted part-of-speech were lemmatized into their root words. The parts of speech used in this analysis were nouns, verbs, adjectives, and adverbs. Each document was also searched for named entities that were organizations, geographical locations, monetary values, and percentage values. This level of analysis was included because such information may help distinguish economic from non-economic articles and stories about the U.S. from stories about other countries (or non-specific to geography). Percentages and monetary values were simply added as tokens called "PERCENT" and "MONEY" respectively, while the full names of organizations and places were changed to uppercase (such that named entities/categories can be distinguished from the first set of tokens).

Topic Modeling

Prediction Modeling

Conclusions