Using Unsupervised Learning to Determine Stock Market Sectors

Atish Sawant, atishs@stanford.edu

Introduction and Motivation

Every company in the S&P 500 submits a report to the SEC detailing their performance for the quarter along with market risks. The index is further broken down into 11 market sectors. I wished to see if the market sector delineation was borne out by the reports, and the quality of the sectors created. If quality sectors can be created through text-mining of the SEC reports then it implies that news headlines can be classified into clusters, and can help manage returns and risk.

Data

The data was collected from the SEC repository of company filings--
https://www.sec.gov/dera/data/

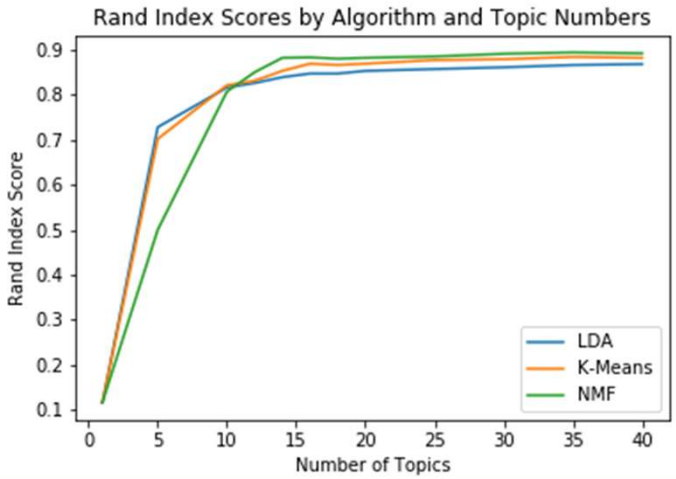
The data included all of the text in every report submitted to the SEC by every company in the first and second quarters of 2017. I limited the data to just the S&P 500, and limited the minimum length of each text segment. Many text segments were title headers, or footers and were less than 20 characters. Each document was actually a full document, since each instance of text in a report was classified separately. All of the partial instances of text were reassembled into one large document with one label prior to modelling. Special characters and numbers were also removed.

Features

As a text mining project, the feature set includes all of the words in the documents. As part of the experimentation process the words had to show up in at least two documents, and the upper bound was varied to identify the optimal model. Words showing up in 90% of the documents likely have no explanatory power, but words showing up in 30% may have significant meaning. Therefore the upper bound was tweaked as a hyperparameter.

For NMF and K-Means Clustering, TFIDF was performed, while for LDA, word counts were used. To test for clustering cohesiveness, Rand Index was used with the ground truth based on current sector identity. Sub-sectors were not used, but could be an area for further research.

Results



Models

Three primary models were used for the analysis.

1. K-Means Clustering

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

K-Means served as an initial baseline, and close distances between the vectors of TFIDF interpreted as closeness of meaning

2. Non-negative Matrix Factorization(NMF)

$$\min_{W \geq 0, H \geq 0} \|A - WH\|_F^2.$$

NMF reduced the dimensionality of the TFIDF vectors for each document down to a preset number of assumed components which could be viewed as clusters.

3. Latent Dirichlet Allocation (LDA)

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

LDA assumes each document is a mixture of topics, and each topic has a word probability associated with it.

Discussion

NMF was the best performing of the algorithms on this task as shown by the highest Rand Index score. This is somewhat surprising, as the expectation would have been for LDA to outperform given the additional freedom it has in picking a mix of topics. Part of the reason for underperformance in LDA may have been because it was picking up types of reports rather than types of industries. Controlling for business stage and cycle is something that would warrant further research.

All 3 algorithms saw a quick uptick in scores right up to the 10-12 topic range which makes sense given that the actual number of sectors is 11. However, I did expect that some of the current sectors could be broken down further and the Rand Index scores back this up, as they rise a little bit more up until 15 before largely plateauing. Airlines and heavy machinery companies are both in the Industrials group while one could argue that they are different enough to warrant their own section.

All the clusters were not created equally however. Often times, there were one or two clusters that had a jumble of groups present while the others were relatively pure. These jumbled groups may have to do with the type of report or company stage that they were in.

Further Research

Further areas for research involve using industry subsectors to see if a more granular ground truth can also result in more accurate clusters.

One pernicious problem is that NMF and LDA assume a vector of topics, but the real space may be a matrix. These reports are market based, so a struggling retail company may have many similarities to a struggling industrial company. Accounting for market performance may help solve this problem, and result in a stronger ability to tag news headlines to sectors.