Topic Modeling in Financial Documents

Patrick Grafe
Department of Computer Science
Stanford University
pgrafe@stanford.edu

ABSTRACT

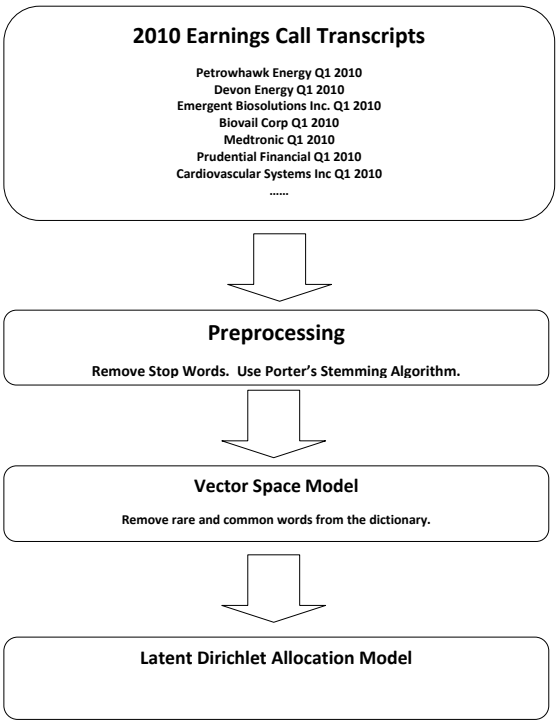
This paper describes the application of topic modeling techniques to quarterly earnings call transcripts of publicly traded companies. Earnings call transcripts represent an interesting case for analysis because the document is relatively unstructured and potentially more informative than 10K and 10Q disclosures due to the question and answer session consisting of unprepared statements. This paper addresses the clustering of these documents as well as the segmentation of individual documents into clusters for products and industries the company is active in. The goal is for each transcript to be assigned to some number of topics, and the specific segments of the transcript which address a given topic to be specified as well. Thus, not only will the documents be classified as covering some set of topics, but the documents themselves will be partitioned into different sub-topics. I will discuss progress I made in achieving these goals as well as challenges and issues which remain. This work could prove useful in financial document summarization as well as improving search and display of documents and information relevant to a user's search and interests. Furtheremore, applying NLP and machine learning concepts to financial document analysis is increasingly being used by trading firms and hedge funds to gain competitive advantage.

Keywords

Latent Dirichlet Allocation, Clustering

1. INTRODUCTION

The goal of this project is to effectively discover common topics among a large data set of earnings call transcripts of publicly traded companies. Each transcript will be assigned to some number of topics, and the specific segments of the transcript which address a given topic will hopefully be specified as well. Thus, not only will the documents be classified as covering some set of topics, but the documents themselves will be partitioned into different sub-topics. This work would be useful in improving the search of financial documents and serving users appropriate documents including those that, while not about a specific company, may be highly relevant due to a shared industry or product. The following sections discuss all major aspects of this task including: data gathering, preprocessing, segmentation, clustering, and analysis.



2. DATA PREPROCESSING

The data set was gathered by scraping 3800 earnings call transcripts from seekingalpha.com. These transcripts represent approximately one year's worth of financial data. The transcripts were first stripped of HTML markup while preserving all other punctuation, character case, and stop words. The data was then tokenized using the TreeBank Word Tokenizer provided by the Natural Language Toolkit (nltk) python package. Commonly used words in the English language, known as stop words, are also subsequently removed in order to reduce the number of features as well as to prevent clustering from being affected by such content-free

Table 1: Topics

Oil and Natural Gas	Biotech	Real Estate	Media and Networking	Misc
gas	patient	occupants	network	gas
drill	trial	tenant	brand	scrap
rig	clinic	hotel	software	mario
barrel	fda	revpar	wireless	russo
acreage	dose	music	tv	glenrock
haynesville	cancer	noi	video	gasoline

words. Finally, to further reduce and improve the feature set, the tokens were stemmed using nltk's implementation of Porter's stemming algorithm so that words with common roots would now be counted as the same.

2.1 Vector Space Model

Initially, I represented the data set using unigram, bigram, and trigram counts; however, after confronting severe memory limitations, I opted to use a vector space model using simply unigrams which would likely be similarly effective and considerably more memory efficient. Thus each word encountered in the documents are stored in a dictionary, and the ID of those words are only included in a given bag of words model if the document itself contains that word. The remainder of my work uses this bag of words vector space model. I considered making use of bigrams in my models; however, from prior knowledge about the relative ineffectiveness of bigrams in sentiment analysis I decided not to do so.

3. CLUSTERING USING LDA

Originally, I intended to make use of k-means to cluster the data; however, the limitations of k-means especially with regard to its inability to maintain multiple topic distributions for each document led me to try Latent Dirichlet Allocation instead. Initial analysis using LDA proved unsuccessful with this data set because the clusters formed along the lines of common words found in a majority of the documents with topic salience matrices lead by words such as "million", "income", "growth", etc. I removed stop words in hopes of avoiding such a situation; however, it became clear I also needed to remove words common to my data set specifically. Thus in all future analysis, words that appeared in a very large percentage of the documents (or in the next section, segments) were removed. Similarly words found in very few documents were also removed as well. The thresholds used were arbitrary, removing words that appear in more than 50 percent of the documents as well as those that occurred in fewer than 2 percent of the documents. These thresholds represent significant tuning parameters for the clustering algorithm which I will discuss later in my analysis. Some results using this analysis are shown in Table 1.

4. DOCUMENT SEGMENTATION

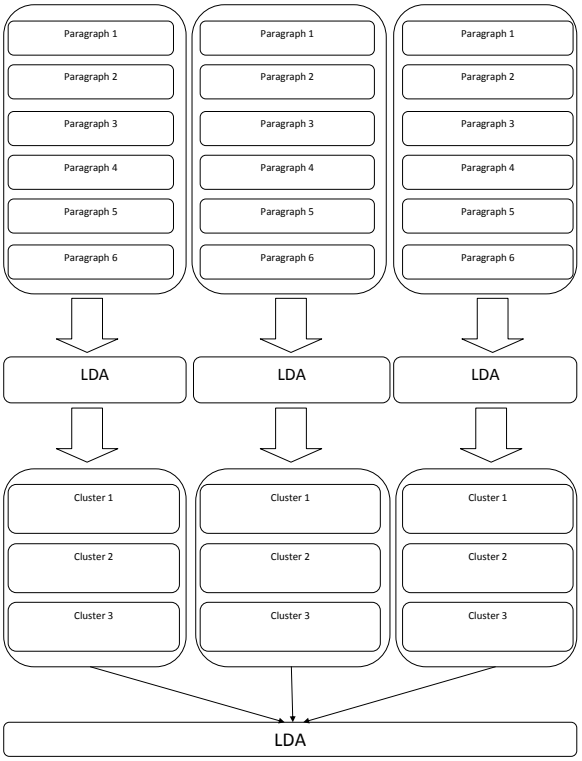
To deal appropriately with companies involved in multiple industries, I used an approach described by Tagarelli and Karypis [1] which involved splitting each individual document into paragraphs and then proceeding to cluster the paragraphs into segments within the document. These clusters are then used in lieu of the complete documents when performing LDA across the entire data set. This allows a company such as Halliburton which works in multiple industries (oil and natural gas, construction, military work) to be correctly grouped into each of these three industries more effectively.

The first step was to split each document into its constituent paragraphs. For earning's call transcripts, there were many "paragraphs" which consist of simple greetings, introductions, and relatively content-free one line questions. Some of the one line paragraphs represent questions, which should properly be grouped with their associated answers, but for simplicity, I chose to remove these paragraphs from my analysis. Thus any paragraphs with fewer than 100 characters were removed.

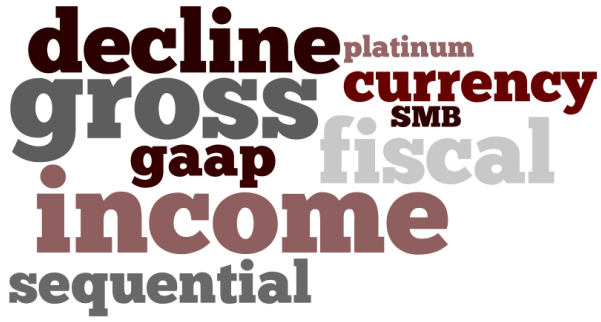
5. INTRA-DOCUMENT CLUSTERING

Given the fact that most companies operate in a limited number of industries, I performed clustering in this step with K = 3. The results were somewhat disappointing for this intra-document clustering. Unlike with my previous analyses, words such as "million", "income", and "revenue" were not removed because a majority of the segments made no mention of them. As a result, the intra-document clustering resulted in one or more clusters of very basic accounting terminology common to all corporations. On the positive side, for many companies there were one or two correctly identified clusters clustering on industry specific terms. Overall, for this data set, I do not believe this method holds much promise because a review of the companies involved shows that most do not have a clear presence in multiple industries and thus clustering within those documents is futile.

6. CLUSTERING DOCUMENT SEGMENTS



Once the clustering within each document was complete, these segment clusters for the entire data set were further clustered using values of K = 5 and K = 10. Given that the majority of companies operate in one primary industry, no significant changes were observed when clustering segments as compared to documents. The only significant change was that since each document always contained a cluster on accounting and finance terms, there was a clearly defined cluster on these terms in the final set of clusters. The following tag clouds are two representative examples of the results seen during this analysis, one is a cluster of networking companies and the other is the cluster on accounting concepts (graphics courtesy of wordle.net).



7. TUNING PARAMETERS

Since my dataset is currently unlabeled, choosing K as well as some other previously mentioned tuning parameters was a dificult and imprecise task. The choice of K depends on how ambitious I wish to be with clustering the data. Clustering for K = 5 worked surprisingly well with four very well clustered industries (biotech, energy, real estate, and technology) and one miscellaneous category. For K = 10, the resulting clusters were considerably less homogenous than for K = 5. There were more incorrect classifications and some incorrect mixing of some categories; however, with these drawbacks also came some improvements. The new clusters covered more narrow industries. Most prominently, there were now two clusters related to real estate, one of which focused on hotels and rental properties. Another cluster was found which was mostly representative of the semiconductor industry. Choosing values of K greater than 10 was not as successful. Most clusters then fell under the category of miscellaneous, while some clearly defined clusters such as oil and natural gas remained.

One of the problems seen at values of K greater than 10, was that clusters would form around proper names and places, such as "john," "needham," and "mario." Another problem seen was clustering around words which aren't indicative of any particular industry, product, or place. I attempted to deal with these problems by tweaking the thresholds I had previously used to remove words common to most documents and words that are very rare, but unfortunately doing so only made things worse, even for K = 5.

8. CONCLUSIONS

The results obtained in this project have been mixed at best. It is clear that these techniques hold potential to correctly group companies into their respective industries and special areas. The energy, biotech, real estate, and technology industries are very clearly defined and grouped; however, for every successful clustering, there's a disappointment. For all values of K analyzed, including K = 5, there was always at least one cluster that can at best be described as miscellaneous. This may be acceptable at times, for example, it's probably not possible to create a reasonable way to split the data set into a mere 5 categories. Unfortunately, since there's no automated way to determine which of the clusters should be classified as miscellaneous, this makes the usefulness of these techniques questionable. These techniques also have some potential to cluster well on an intra-document level as well, but again the same caveats apply. Furthermore, there's a lot of noise in the data when examining at the segment level. Fortunately, for intra-document segment clustering, if the words common among all documents (ie accounting jargon, boilerplate, etc) are removed as a preprocessing step, then very different and likely more meaningful results will be found. Overall, despite my belief that this would be a useful data set to analyze, it has proven to be more difficult than expected.

9. FUTURE WORK

In the future, I will need some specific metric to compare the quality of clustering obtained using different values of K. For my current analysis, I experimented with the following values of K: 5, 10, 20, 30. Unfortunately, since I am working with an unlabeled data set and with an ambiguous objective, it's difficult to quantify success and what is the optimal value of K. In the future, I will need to create a small labeled test set, which will allow me to quantify my success and provide a metric, namely perplexity and classification error, to determine the optimal value of K.

Additionally, for values of K larger than 10, there is an interesting phenomenon where proper names appear to be of great importance to topic discovery. This may help improve topic modeling for the training set because some names are indeed commonly seen only in a specific industry. For example, one investment capital corporation, Needham Capital Partners, invests heavily in semiconductor and other high tech corporations. Consequently, Needham Capital representatives attended and asked questions at teleconferences for such companies. As a result, "needham" was often highly indicative of a semiconductor or high-tech grouping. This may cause problems in generalization however, especially for more common proper names such as "john."

For this work to be useful it will be important to be able to determine which clusters are reasonable and appropriate and which consist of all that's left over just thrown into one. For the case of intra-document clustering, this seems

possible to do either in an intelligent manner or in a brute force manner, ie by eliminating a cluster automatically based on the prominence of accounting terminology in its term salience matrix.

Finally, I originally intended to try out the Correlated Topic Model proposed by Blei, et al [2]. This topic model extends LDA with the observation that topics are not independent, thus allowing correlations. Due to time constraints, I was unable to implement this, but I feel it may be useful in the future.

10. REFERENCES

[1] A. Tagarelli, G. Karypis. A Segment-based Approach To Clustering Multi-Topic Documents. 2008.

[2] D. Blei, J. Lafferty. Correlated Topic Models.