Topic Modeling - A Survey of Three Leading Unsupervised Methods

Austin Pennington

December 9, 2018

SYNOPSIS

The extraction of semantic topics from documents is of tremendous interest across a variety of fields, ranging from academic research, to intelligence, to financial services. Drawing from fundamental research on how meaning is derived from documents, numerous techniques have been adapted from various fields and several topic extraction algorithms have been created in recent years to address this interest. Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and Correlated Topic Modeling (CTM) employ three different approaches to the problem. We evaluate these three approaches across a common corpus by comparing the characteristics of their extracted topics. We do not attempt to evaluate techniques around the derivation of the optimal number of topics assumed to be present across the corpus, but, rather, use a starting point assumption (based on expert evaluation) number of topics and then see how all three techniques perform. The corpus is a set of US regulatory filings (SEC Form 8-K) representing material financial information filed on a required timely basis, communicating a certain set of "events" that must be reported to the public. The semi-structured nature of the corpus, along with the pre-defined range of events that any of the documents may contain, make this a particularly interesting study relevant to the legal and financial services fields. Our initial results show a clear similarity and consistency across all three techniques, with a slightly more semantically intuitive set of results from the CTM approach, as we had initially suspected may be the case. The NMF approach yielded the least distinct result set, with closer to uniform overall topic distribution compared to the other two methodologies. One clear outcome of this research is that more work is needed to develop new methodologies that quantitatively compare topic models in order to more consistently assess the quality of models generated by various algorithms, without resorting to expert opinion.

INTRODUCTION

Topic extraction is of critical importance to realize the potential of deploying computer algorithms across large sets of human-readable documents in order to achieve several benefits. One potential benefit of accurately determining meaning is the extraction of topics to automate and augment many basic aspects of research analytics. The ability to reliably deploy a set of algorithms that can ingest and categorize documents and sections of documents could form a "pre-processing" step to highlight particularly relevant and interesting content for deeper review and research by experts. This would expand both the depth and scope of coverage of a human expert. The approach taken in this research is to start with a corpus of similar documents that are in a semi-structured format, drawn from a large and dynamic source of information critical to the capital markets from a regulatory, legal, and financial perspective.

The US Securities and Exchange Commission (SEC) requires all companies who issue debt or equity into the public capital markets (registrants) to file a "current" report in what is called a form 8-K (and its derivatives). The purpose of the 8-K filings is to quickly disseminate material information impacting the valuation of public securities (stocks and bonds) to the entire investing public. We start with the determination that Form 8-K information will always fall into one of five overarching categories; financial condition updates, updates relevant to specific regulatory conditions regarding the companies' registered securities, new exhibits related to financial condition or regulatory updates, updates related to company legal conditions, and updates related to governance conditions including changes to the board of directors or executives of the firm.

The trigger for an 8-K filing is the occurrence of one of 28 events that are specifically defined in the rules, plus the disclosure of associated new "exhibits" which are attachments to the filings, plus "other" events, which are specifically defined in the rules. Even with this "other" event category, we know that all 30 potential filing triggers into one of the five overarching categories described already, since companies have no incentive to file anything beyond the minimum and over-disclosing on unrelated events opens up a firm to significant legal and regulatory risk. Thus, we predetermine a ground-truth set of 5 topics as inputs to our topic modeling algorithms (without defining what they are) as an input for our algorithms, along with the corpus itself.

We drew upon prior research in three ways:

(1) Research regarding semantic representation,

(2) Research into topic extraction techniques, from which we selected LDA, NMF, and CTM, and

(3) Research regarding prior work done on the Form 8-K corpus.

There are a handful of seminal research papers regarding each of the extraction techniques in the context of topic extraction from documents. For our review, we focused on the early and later works of David Blei and his collaborators (Blei, 2003; Blei, 2005; and Blei 2012), as well as scholarly work produced by Georgia Tech (Da Kuang, 2015). These papers are highly cited and serve as foundational ideas upon which many implementations and research efforts have been based.

THEORY

Semantic representation approaches mainly attempt to deploy statistical inference methods in such a way that words are associated into groups with similar "gist" (Griffiths, 2006), and these groups form topics in which meaning is disambiguated by making assumptions based on association with other words in the same document. These groups can be approached in one of three ways (Griffiths, 2006):

(1) Topic graphs,

(2) Topic clusters, or

(3) Topic models.

It is this basis, the idea that language can be represented as a set of probabilistic associations which can predict meaning, that allows us to experiment with extracting meaning from documents using computer algorithms. Topic graphs require some way of training (some form of human supervision) an algorithm in order to construct reliable graphs associating words with other words and applying weights between such words to derive context and meaning. Applications of clustering and topic modeling are promising for unsupervised applications, like ours. The three methods chosen for this research fall into the clustering (NMF) and topic modeling approach (LDA, CTM).

Nonnegative Matrix Factorization (NMF)

Clustering methodologies are a natural fit for grouping certain words together within or across documents. NMF is a dimensionality reduction technique that splits a term-document matrix into two lower dimension matrices specified by the number of topics presumed present.

$$TDM(m \times n) \sim [W(m \times k)][H(k \times n)] \quad (1)$$

Reduced dimension matrices W and H are determined by solving an optimization problem, in this project defined with the Kullback-Leibler (KL) divergence (Gajoux, 2018). The term-document matrix is an m-by-n dimension matrix, and the two reduced dimension matrices are sized using k as the number of topics presumed present in the corpus. Calculating W and H, we have to minimize a loss function that approximates the similarity of the two matrices. One approach is to minimize Frobenius distance, the distance between two matrices. This is performed by differencing two matrices, square the resulting matrix, taking the trace of that squared distance matrix, and taking the square root. This is the matrix equivalent of the euclidean distance. The approach we used in this research is to minimize KL divergence. This is a probabilistic approach that measures how different one probability distribution is from a reference probability distribution.

Latent Dirichlet Allocation (LDA)

A modification of early models, Latent Semantic Analysis and its successor, Probabilistic Latent Semantic Analysis, LDA is considered a superior probabilistic topic modeling algorithm (Blei, 2003). The assumption that over all language, words have certain probabilities of falling into any number of topics. Topics are independent and words are stochastically grouped into the overall set of possible topics following a Dirichlet distribution. For LDA, there are three main assumptions:

1. Bag of words - the order of words does not matter
2. Bag of documents - the order of documents does not matter
3. Fixed, known number of topics - meaning that topics do not change over time

To estimate the parameters for the Dirichlet distribution, variational expectation maximization (VEM) procedure has been used (Blei, 2003), which also optimizes on KL divergence.

Correlated Topic Modeling (CTM)

Building on LDA, adding on the association that topics can have with other topics, CTM was introduced to improve the ability of topic models to account for the likelihood that the occurrence of a topic gives additional information about the likelihood of other topics also being included in a particular document or corpus (Blei, 2005). Rather than assign topics to a Dirichlet distribution, in CTM topics are considered to follow a logistic normal distribution. The parameters of CTM are drawn from a multivariate normal; the covariance matrix of the CTM generation process allows for modeling correlation between topics in a way that is not possible given the independence assumption of the LDA model, imposed by the Dirichlet distribution.

The other main assumptions of LDA hold for CTM, and this makes the comparison of the two approaches particularly interesting on a corpora that has a technical, narrow range of ground-truth topics present. The approach we followed builds on the overall formulation of the LDA model. Parameter estimation and optimization is performed similarly to the LDA algorithm described above. Since the corpus is presumed to have five ground truth topics, the presence we expect that the topics should be correlated and that the main focus will be to see if the algorithm produces more distinct and more descriptive topic models than LDA in this project.

PROCEDURE

To achieve greater consistency in our approach, each of the three techniques was implemented in R using algorithms specific to the three different analyses. Before applying the algorithms, however, a great deal of pre-processing must be performed on the documents before the Corpus can be consistently analyzed.

The first step to preparing the documents involved removing all the HTML header lines and document tags. Next, a random sample of 2,000 documents was pulled from the population of 200,000 documents. This sample formed the corpus.

Then, across the entire corpus, data cleansing work involved converting all letters to lower case and removing several noisy characteristics including: all numbers, stop words, white space, and punctuation. The final step consisted of stemming the remaining words and then taking this cleansed corpus to begin the various topic modeling approaches.

The next critical step to implement the NMF algorithm is to create the Term-Document matrix from the corpus. That becomes the input along with the specified number of topics presumed present. For this project, the Rpackage:NMF was deployed. The Kullback-Leibler (KL) optimization (factorization) method was used, with a random seeding method to start the algorithm.

For LDA and CTM, the algorithm starts with a Document-Term matrix from the corpus. That becomes the input along with the specified number of topics presumed present. For this project, the Rpackage:topicmodels was deployed. For both LDA and CTM, parameter estimation used the VEM method described for LDA, above.

RESULTS

Based on the final topic models for each of the three methods, Beta weights are used to filter the top fifteen words (highest Beta) associated with each of the five topics. Since each word can be associated with more than one topic, the probabilities allow finding a word across several topics. However, the top fifteen ranking should be distinct enough to easily distinguish between topics. This ranking was then reviewed to infer topic alignment back to the original five presumed topics of the Corpus. The results of this mapping are shown in Table 1. The relative occurrence of each of the topics was also assigned a value of high, medium, and low, based on the ranking of the topics for each methodology. It is clear that the patterns of LDA and CTM are very similar, and the NMF pattern is much closer to an even distribution of topics across the corpus.

Topic Description	NMF	LDA	CTM
Financial Condition Disclosure	Topic 2	Topic 1	Topic 3
Relative Occurrence	Medium	High	High
Regulatory Update Disclosure	Topic 1	Topic 4	Topic 1
Relative Occurrence	Medium	Low	Low
Financial & Regulatory Exhibits	Topic 3	Topic 5	Topic 4
Relative Occurrence	Medium	High	High
Legal Condition Disclosure	Topic 4	Topic 2	Topic 2
Relative Occurrence	High	Low	Low
Governance/Board/Executive Disclosure	Topic 5	Topic 3	Topic 5
Relative Occurrence	Medium	Medium	Medium

Table 1 - Cross Method Topic Comparison

For each of three methods, we review beta weights as a primary way of explaining the makeup of each of the topic models (see figures 1, 2, and 3 below).

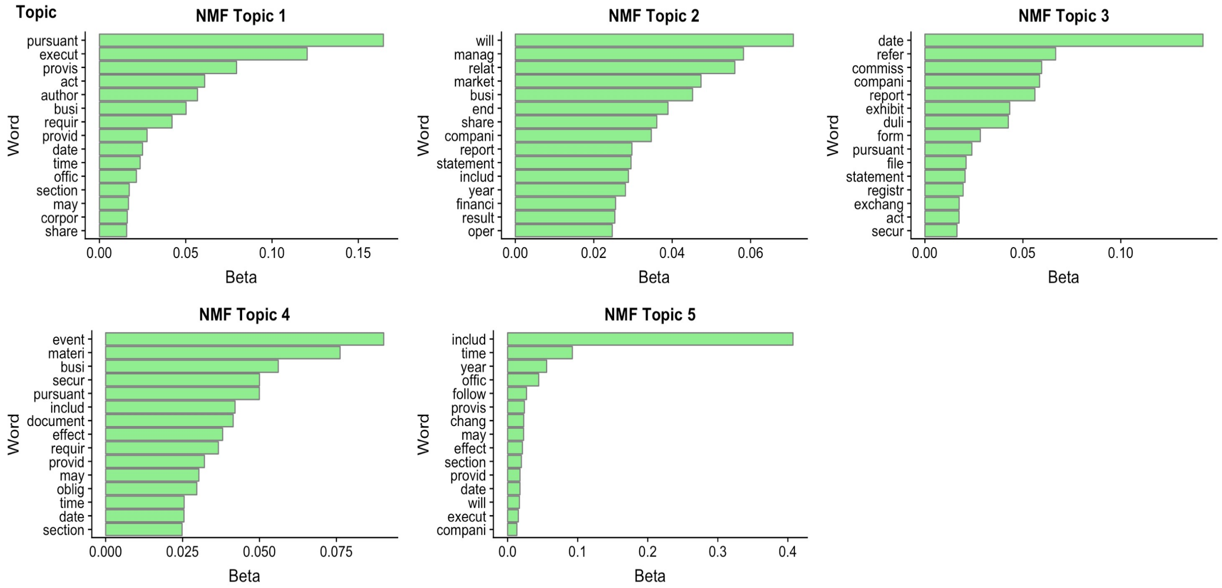

Figure 1 - NMF Top 15 Beta

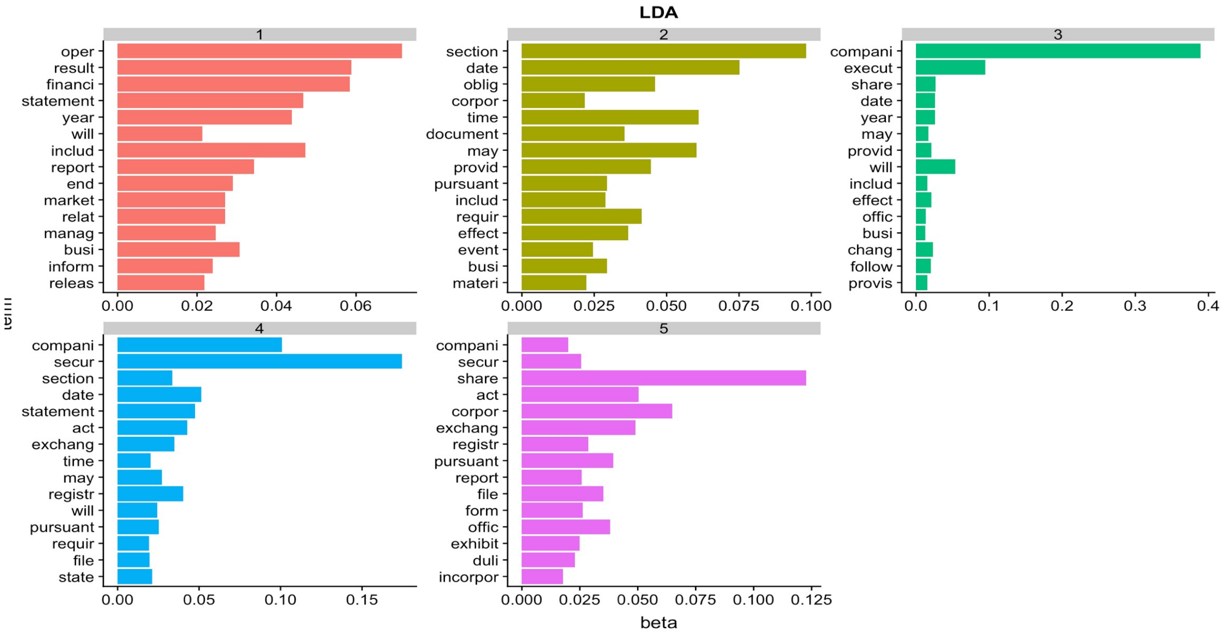

Figure 2 - LDA Top 15 Beta

Beta weights represent the probability of a topic including a particular word across the entire corpus. Each technique produced and organized topics with some consistency across methods, however the distribution for NMF (qualitatively described in Table 1) was distinctly different.

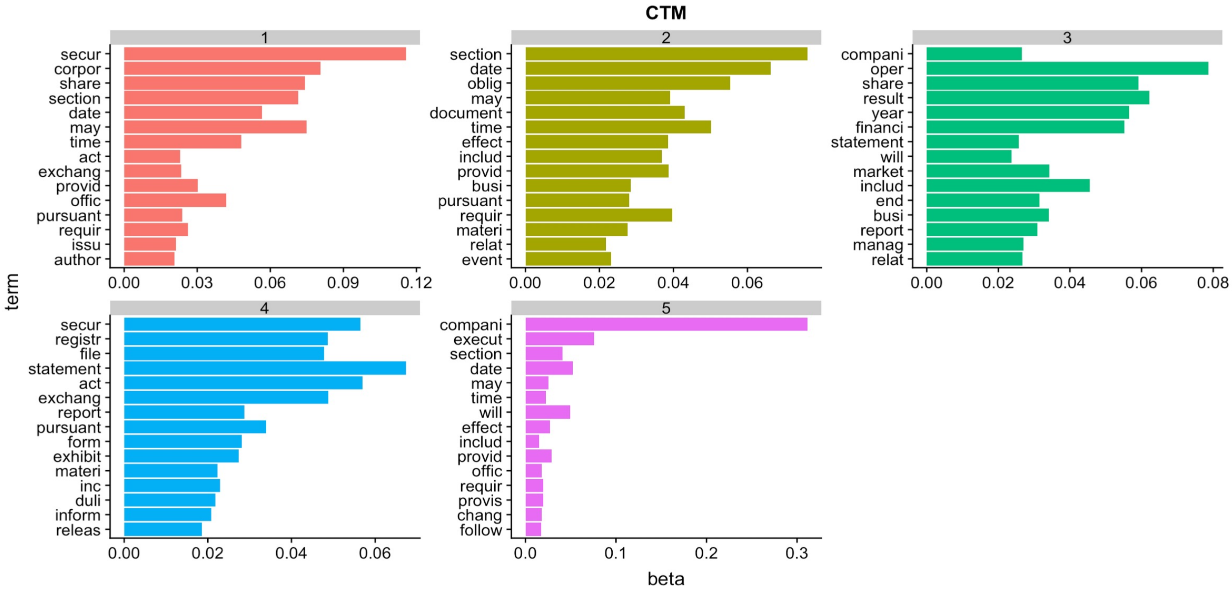

Figure 3 - CTM Top 15 Beta

Gamma Distribution comparison of NMF, LDA, and CTM

Finally we compare paired sets of topic distribution histograms and gamma histograms for each of the three methods as another view into the distinctions between methods and their corresponding topic models (see figures 4, 5, and 6 below). This is primarily interesting in relation to the relative comparison on LDA and its derivative, CTM.

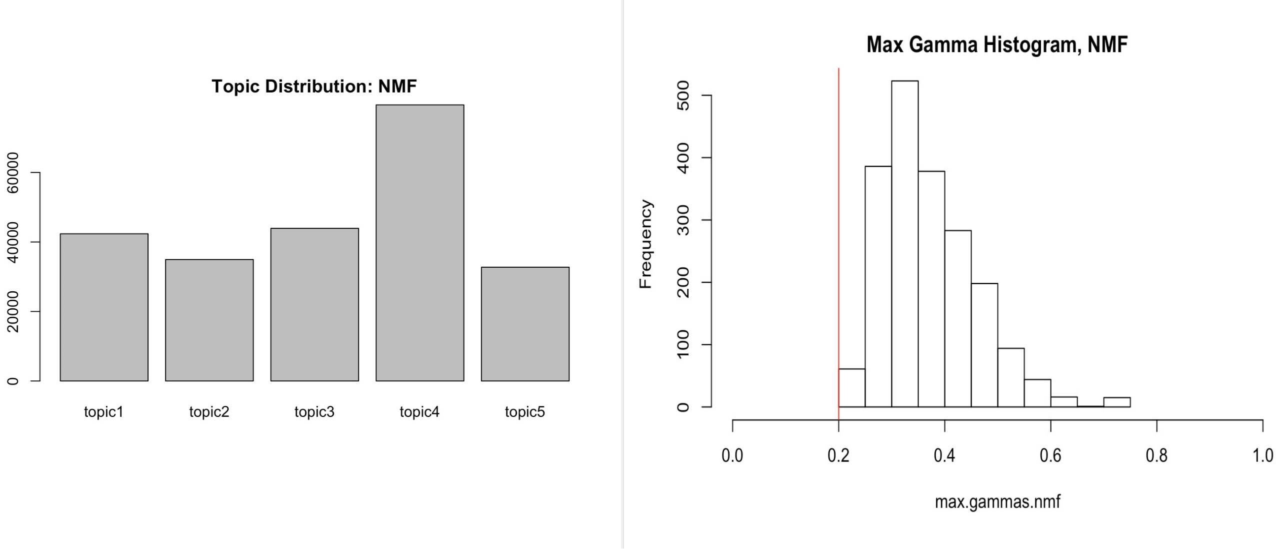

Figure 4 - NMF Relative Topic Distribution (Left) and Maximum Gamma Histogram (Right)

The results of the NMF method, while they do show a similar grouping of topics following a review of the beta rankings, diverge from both the LDA and CTM results. The main difference is that NMF produces a more near uniformly random topic distribution as shown in Figure 4. The maximum gamma distribution is more nearly random, with a grouping near 0.3 and almost no topics showing a maximum gamma of greater than 0.6 over the corpus. Topic distribution shows a maximum for topic 4 (Legal Condition Disclosure). Referring back to Table 1, this is the lowest ranked topic for both LDA and CTM, and it is also one of the lower expected topics based on subjective review of the Form 8-K requirements.

As shown is Figures 5 and 6, LDA and CTM have peak gammas of just around 0.6, and have topic distributions showing two peaks and three lower occurrence topics. CTM has the most distinct topic distribution, and it matches presumed occurrence patterns for the 8-K filings well. The most prevalent expected topics in 8-K filings are financial disclosures and their respective exhibits, consistent with the topic distribution shown in Figure 6.

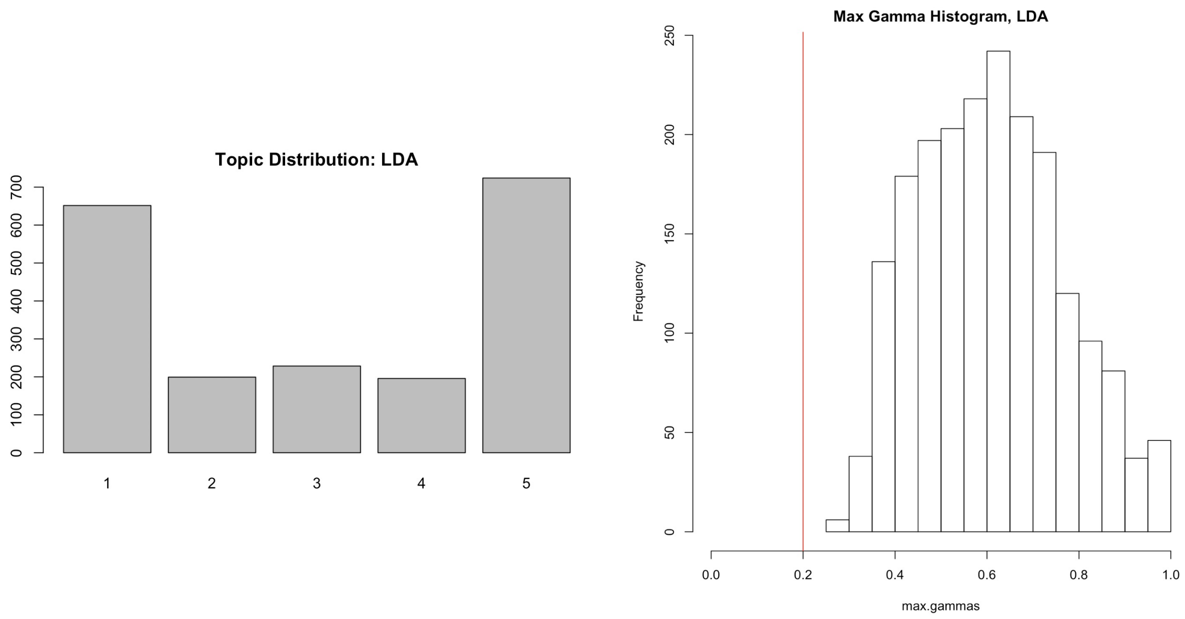

Figure 5 - LDA Relative Topic Distribution (Left) and Maximum Gamma Histogram (Right)

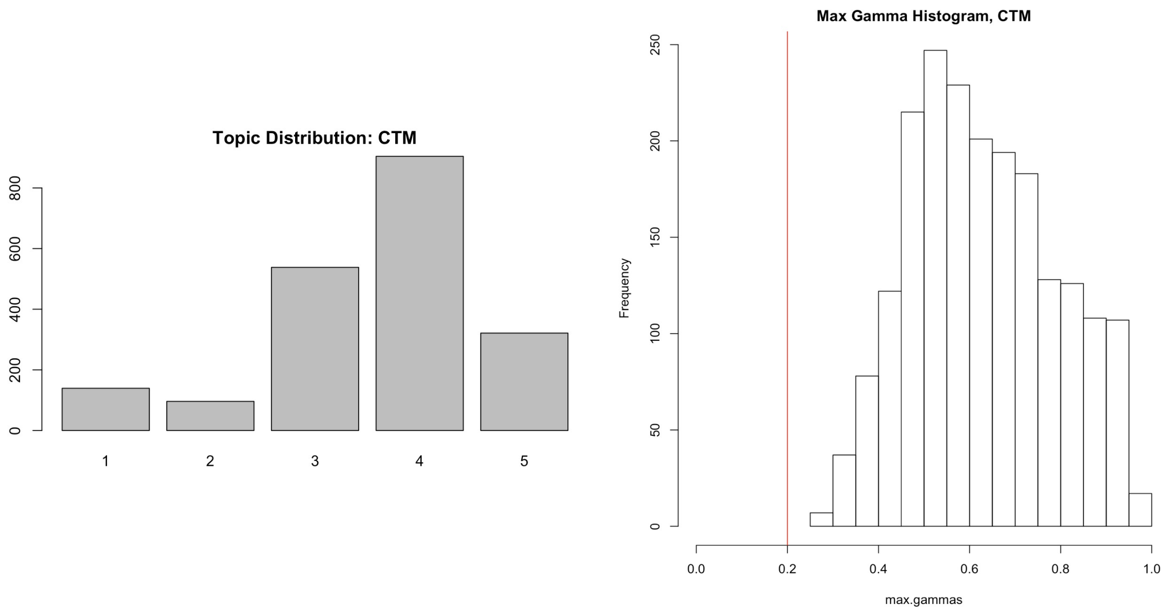

Figure 6 - CTM Relative Topic Distribution (Left) and Maximum Gamma Histogram (Right)

Conclusion and Further Research

In our research we focused on comparing statistics related to high frequency keywords to represent the semantic meaning within each topic, and to illustrate distinctions across topics. This approach is weakened by significant keyword noise requiring expert interpretation by users, relying on some degree of a priori knowledge around the context and expected types of topics from the corpus. The distinction was clear between NMF and the other two methods (LDA and CTM). The topic distribution pattern of the CTM approach best matched the expected results of the topic modeling exercise.

One important area for continued research is to develop ways to measure effectiveness and quality of topic extraction across any algorithm. New quantitative comparisons are needed to better determine the quality of the topic models produced by these algorithms.   It would also be interesting to perform supervised classification using SEC labels, and compare both the accuracy of the unsupervised and supervised topic modeling as well as analyzing possibilities of redundancy and low-information labels within the SEC classification system.

References:

Griffiths et al, “Topics in Semantic Representation”, 2006, http://web.mit.edu/cocosci/Papers/topics15.pdf

Kuang et al, “Nonnegative Matrix Factorization for interactive topic modeling and document clustering”, 2012, https://www.cc.gatech.edu/~hpark/papers/nmf_book_chapter.pdf

Blei and Lafferty, “Correlated Topic Models”, 2005, http://people.ee.duke.edu/~lcarin/Blei2005CTM.pdf

Blei et al, “Latent Dirichlet Allocation”, 2002, http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Blei, “Probabilistic Topic Models”, 2012, http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf

SEC, “Form 8-K Fast Answers”, 2018, https://www.sec.gov/fast-answers/answersform8khtm.html

Lee et al, “On the Importance of Text Analysis for Stock Price Prediction”, 2014, https://nlp.stanford.edu/pubs/lrec2014-stock.pdf

Goldstein and Wu, “Disclosure Timing, Information Asymmetry, and Stock Returns: Evidence from 8-K Filing Texts”, https://www.semanticscholar.org/paper/Disclosure-Timing-%2C-Information-Asymmetry-%2C-and-%3A-Di-Wu/c44c1193518d177d1bb83f77ec124dd2ac80b131

Gajoux, “An Introduction to NMF Package”, 2018, https://cran.r-project.org/web/packages/NMF/vignettes/NMF-vignette.pdf

Grun and Hornik, “Package ‘topicmodels’”, 2018, https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf