Risk Identification: What is in the 10-K?

February 15, 2019

Abstract

The Sarbanes-Oxley Act of 2002 requires all SEC-registered firms to provide a Risk Factor Assessment in their Annual Report known as the 10K. This paper uses a structural topic model (STM) to assess the types of topics contained in the 10Ks and to see how they change over SIC code sectors and over time. While topic models are useful in many respects, they only look at the words of the document and do not account for other information about the document's author, time, type of business, or location. An STM will allow us to refine our topic selection based upon a number of covariates external to the document. The STM approach also permits use of independent variables to add information to the evolution of certain topics such as information security or litigation.

JEL Codes: **G14, C55**

Keywords: **Enterprise Risk Management, Structural Topic Models, Risk Factors, Text Analysis**

1 Introduction

Risk disclosure has been an important, but imperfect, part of the annual reporting process for firms trading on public exchanges. The annual report, known as the 10K, and the general information disclosure requirements mandated by the Securities Acts their Rules was supposed to provide material information to investors. Even before the Enron and Worldcom scandals interested parties such as the American Accounting Association and the FASB were discussing how to implement risk reporting in the financial statements (Schrand and Elliott (1998)). The SEC did act to increase risk reporting over time in certain areas in the management discussion and analysis called Item 1 in the 10K and in the initial registration statement for an IP0, but it wasn't until recently a more formalized requirement was put in place.

Section 1A which covers "Risk Factors" was added to the SEC annual 10K filing as part of the Sarbanes-Oxley Act. Firms must list their Risk Factors in this section of the annual statement, but it is not clear what is actually required in the statement. The idea behind information disclosure is to provide useful information to shareholders, but how is this accomplished?. The SEC makes three statements about the Risk Factors. The first is found in a web document entitled "How to Read a 10K" and the SEC tells readers the risk factors

> include information about the most significant risks that apply to the company or to its securities. Companies generally list the risk factors in order of their importance. In practice, this section focuses on the risks themselves, not how the company addresses those risks. Some risks may be true for the entire economy, some may apply only to the company's industry sector or geographic region, and some may be unique to the company. SEC (2019b)

Second, the SEC provides directions to the filing companies directing them to

> [s]et forth, under the caption "Risk Factors", where appropriate, the risk factors described in Item 503(c) of Regulation S-K ... applicable to the registrant. Provide any discussion of risk factors in plain English in accordance with Rule 421(d) of the Securities Act of 1933 ...). SEC (2019a)

Finally, the SEC has official guidance contained in Rule 503(c) which states:

> [t]his discussion must be concise and organized logically. Do not present risks that could apply to any issuer or any offering. Explain how the risk affects the issuer or the securities being offered. Set forth each risk factor under a subcaption that adequately describes the risk. The risk factor discussion must immediately follow the summary section.[1]

Because of this rather terse guidance in the documents above, there are likely a number of options management could take in reporting its risk factors. First, as Campbell et al. (2014) suggest, managers could simply disclose all possible risks irrespective of their materiality. Another option might be to describe risk so broadly that they convey no meaning to a potential investor. In the above directions they SEC warns filers to avoid generic descriptions of risk factors applying to companies generally. Casual evidence suggests management believes these instructions means that it should attempt to explain all of the risks that are part of the firm's enterprise risk discussions.[2] Campbell et al. (2014)) find evidence that this is true and that firms do not provide boilerplate risk disclosures. However, the Investor Responsibility Research Center IRRC Institute (2016) disagrees and that the companies do not describe their risks with enough specificity and talk about common risk to all companies rather generically.

While the results of these types of studies are interesting, this paper takes a different tack. Risks evolve and affect future firms in the same business as well as in other sectors in the future. it is evident that the risk disclosures do seem to be related to some financial information (Li (2008)). However, these risks are not static and may have a different affect across sectors and years. In addition, it may be that firms, on average, do describe their risk adequately, but some firms do spend more time or effort and devote a greater amount of space in the disclosure discussing certain risks. These factors suggest a different approach may provide more information about the effect of risk discussions on the firm. Annual statements from the SEC are used to discover topics within the text using a structural topic model (STM) approach. (Roberts et al. (2016)) The STM uses machine learning techniques and general linear models to discover latent topics within a set of documents while using independent variables to assist in the predictive ability of the model. From this model, we obtain a probability of a topic being in a document which we can then use to see how

[1] 229 CFR 503(c) (2018).

[2] Discussions with about seven Fortune 500 enterprise level risk directors suggest that material risks discussed by the board are included in the Risk Factor statement.

this probability changes over time and across sectors. In addition, we can see how the vocabulary for a topic changes over time. What is different about this approach is that we can use covariates such as the SIC Code or year predictors of the types of topics each firm will discuss.

As a preview of the very preliminary results, we find that while there are numerous topics that may be contained in a firm's risk factor statement, there does not seem to be much variation among broad sectors. However, there is some variation over time.

The paper is organized as follows. In the next section we present a review of the literature which is followed by a description of the data. We then present an overview of the structural topic model and report preliminary results.

2 Literature Review

Unstructured data, like that found in a firm's annual report, is a tremendous source of information about companies' performance, strategy or risks. Specifically, there is a growing interest in looking at SEC documents and eliciting information from the text to see how it relates to other things in the 10K such as financial results. for example, Li (2008) studies the relationship between corporate earnings and the readability of 10-K filings. Higher complexity 10-Ks as measured by the Fog readability index Gunning (1969) and the length of the 10-K lead to lower earnings.

Other researchers look at sentiment. Sentiment is the difference between positive words that reflect some notion of a positive outlook and negative words which reflect some notion of a negative outlook. The underlying words of positive and negativism come from various dictionaries. Some are psychologically based (Feldman et al. (2010)) and others like Loughran and Mcdonald (2011), develop dictionaries related to finance, accounting, or management of firms.[3] Baker et al. (2016) do a similar analysis looking at policy uncertainty using words found in newspapers. They created a dictionary of words related to policy uncertainty–monetary policy, or trade to develop an economy wide-uncertainty index. Ganguly (2018) looks at the contents and sentiments in the 10Ks to predict litigation risk for class action lawsuits.

Huang and Li (2011) are likely the first to specially look at the Section 1A. They use a multi-label text classification algorithm to identify 25 risks from the Risk Factor section. This process is

[3]See e.g., Loughran and McDonald Financial Sentiment Dictionaries http://www3.nd.edu/ mcdonald/Word_Lists.html.

a k-nearest neighbor classification system which finds words that attempt to minimize the distance between other data features. Campbell et al. (2014) also look at the Risk Factor section. They find that firms disclose more risk if they are facing more risk and tend to provide a meaningful discussion of the risks. Rawte et al. (2018) look at changes in Item 1A from year to year to focus on predicting bank failures. They use a neural network approach and accurately predict over 98 of bank failures in their sample.

3 Data

3.1 Risk Factors Section 1A

We have data from 2011 to 2017 from the Item1 A. Risk Factors section of the annual 10K.[4] Figure 1 and Figure 2 shows the number of companies by year and across 1 digit SIC codes. The figure shows the number of 10-Ks in our analysis each year as well as the length of the document in terms of sentences. We can see that the length of the section is increasing over time. From Table 2 we also see the distribution of firms in our data is predominately in manufacturing, followed by Financial, Insurance and Real Estate (FIRE).

Table 1: Summary of Section 1A Filings by Year and Sentence Length

Year	Count of Firms	Mean	SE	Min	Max
2011	4293	206.31	147.86	3.00	1271.00
2012	4284	217.39	156.18	3.00	1391.00
2013	4235	234.22	167.85	3.00	1595.00
2014	3327	255.05	180.98	3.00	1695.00
2015	3942	261.27	185.77	3.00	1769.00
2016	3617	269.81	190.79	3.00	1945.00
2017	3370	283.36	204.57	3.00	2137.00

3.2 Meta Data

In addition, to the text from the filings, we also have meta data. These data are information about the document itself. This would include the SIC code of the company that originated the 10k, the date it was issued, the fiscal year the 10K covered. The first part of the 10K itself

[4]See Appendix A for a summary of the process for obtaining and cleaning the data. We have data from 2004-2010 also, but more work needs to be finished on cleaning this data.

Table 2: Summary Of 10Ks by SIC Sector

Sector	Count of Firms
Missing Code	692
Agriculture, Forestry, Fishing	95
Construction	677
FIRE	13411
Manufacturing	21868
Mining	2918
Retail Trade	3250
Services	9154
Trans, Comms, Utilities	5383
Unclassified	31
Wholesale Trade	1590

contains this information. In addition, we calculated various readability scores for the Section 1a text. Finally, we matched data from Compustat such as total assets and ebita to the filings. We classified each company by its SIC sector which loosely translates to its one digit SIC code. We later take advantage of more refined sectors, but for the initial discussion, we focus on the sector. We use the Flesch.Kinciad measure because it is interpreted as the grade level necessary for comprehension of a document. For example, a level of 17.2 represents a first-year college vocabulary. We do see some extremes with the minimum reflecting a second-grade education and the maximum implies an education requiring 36 years in school. Other readability measures are included for future work.

Table 3: Summary Statistics for Meta Data

Statistic	N	Mean	St. Dev.	Min	Max
Filed_year	27,068	2,013.838	1.998	2,011	2,017
SIC Code	27,068	4,759.309	1,973.608	100	9,995
EBIT	27,068	496.416	3,077	−25,913	130,622
Total Assets ($Mil)	27,068	10,644.	92,817.	0.001	3,287
N Sentences	27,068	244.502	177.616	3.000	2,137
Flesch.PSK	27,068	8.443	0.716	4.497	12.065
Flesch.Kinciad	27,068	17.246	2.966	2.329	36.590
FOG	27,068	21.030	3.372	2.694	41.359
FOG.PSK	27,068	9.912	1.861	1.156	27.354

3.3 Sentiment Analysis

Sentiment analysis is another tool researches use to assess the content of 10Ks (see e.g. **?** and Loughran and Mcdonald (2014)). We show a couple of examples to be used as contrasts to our STM approach below. First, one obtains a dictionary of words. Ideally, this dictionary is designed for the purpose at hand. **?** shows that the traditional sentiment dictionaries are not appropriate for looking at financial documents and they develop a more appropriate one. The first set of dictionaries are from Baker et al. (2016) an consist of words list devoted to various types of economic policy uncertainty (EPU). We also use the the economics and finance positive and negative words lists by Loughran and McDonald (denoted by the prefix LM). **?** also have two other dictionaries based on uncertainty and litigation also denoted by the prefix LM.

5 shows the dictionaries' word counts applied to the FIRE Sector of the Risk Factor Data. This is just an example of the type of anlaysis that could be done using dictionaries. Essentially, we count the words in a document belonging to a specific dictionary. The word count is just regressed against time and sector fixed effects. We could regress these word counts again financial indiactors for firms as do many of the papers cited above. In this example in 5 we see that the sectors and year fixed effects are nearly all significant. It is important to note that our model described below will generate dictionaries for various topics contained in the data through an unsupervised machine learning technique. This is in contrast to curated dictionaries. A benefit to the later model is that one need not specify a word list as the data will provide that list using a machine learning approach.

4 Methods

4.1 STM

Topic models help researchers with the very large datasets of text that are often unstructured (Roberts et al. (2016)). The underlying method uses Latent Dirichlet Allocation or LDA (Blei et; al. (2003) and Blei and Lafferty (2009)) In this paper we use a structured topic model approach (Roberts et al. (2014)) which uses LDA. Chaney and Blei (2012) and Kuhn (2018) outline the LDA process as collecting documents and words in the documents from a generative model. This model

Table 4: Summary Statistics for Dictionaries

Dictionary	Median	Mean	S.E.	Sum	Min	Max
EPU Entitlement Progs	124	160.90	172.53	2157807	0	234
EPU Financial Regulation*	77	99.78	110.35	1338206	0	126
EPU Fiscal Policy	10692	12,240.27	12392.18	164154276	0	0
EPU Health Care	32819	51,540.45	64676.51	691208942	0	268
EPU Labor	0	0.02	0.29	202	0	65
EPU Legal	1	6.85	15.75	91837	0	51
EPU Monetary Policy*	0	0.00	0.00	0	0	15
EPU Nat'l Security	0	3.40	13.83	45584	0	21
EPU Regulation	0	0.85	1.81	11459	0	193
EPU Taxes	0	1.18	7.56	15814	0	267
EPU Trade Policy	3	7.39	12.87	99091	0	15
LM Litigation	0	1.89	5.38	25382	0	833780
LM Negative	0	0.03	0.33	366	0	1983
LM Positive	3	5.42	7.45	72738	0	1312
LM Uncertainty	0	0.51	1.78	6883	0	215919

has observed and unobserved components. Each document (d $\in D$) is generated by the following process:

(1) N_d is the number of words $\sim$ Poisson

(2) The parameter θ_d is the topic proportions within a document and $\sim$ Dirichlet(α)

(3) Each word is associated with a topic. The topic of the word $Z_{d,n} \sim$ multinomial(θ_d)

(4) Each topic k has a model parameter β_k representing word proportions in the topic.

(5) Words $W_{d,n} \sim$ multinomial($W_{d,n}/Z_{d,n}$,β_k)

Undr LDA each document is is split into individual words. The words are randomly assigned to a topic. Using a Bayesian sampling technique, the sampler interactively updates the topic assignment given the topic assignment of all other words. The process continues until a maximum likelihood is reached.

Each document will have the same set of topics with words having a probability β_k of being in each topic. The topics are derived from the word combinations in the set of 10-Ks, but each document exhibits those topics with different proportions (θ_d).

Table 5: Sentiment Analysis Results for 3 Dictionaries for Firms in the Fire Sector

	Dependent variable:			
	HU_SCORE	LM_SCORE	LM_UNCERT	LM_LITIG
	(1)	(2)	(3)	(4)
2012	0.170***	−0.078	671.583	1,684.758
	(0.063)	(3.536)	(567.730)	(3,109.345)
2013	0.109*	6.974**	−874.526	−6,034.890*
	(0.063)	(3.539)	(568.204)	(3,111.938)
2014	0.185***	10.388***	−566.694	−3,818.702
	(0.067)	(3.741)	(600.789)	(3,290.400)
2015	0.149**	14.714***	−1,987.771***	−13,076.310***
	(0.064)	(3.586)	(575.890)	(3,154.037)
2016	0.256***	21.344***	−2,287.742***	−15,964.050***
	(0.065)	(3.646)	(585.531)	(3,206.836)
2017	0.225***	29.466***	−3,199.138***	−23,560.660***
	(0.066)	(3.713)	(596.239)	(3,265.482)
Holding & Other Investment Offices	0.297***	−17.244***	3,310.656***	12,314.790***
	(0.043)	(2.395)	(384.595)	(2,106.349)
Insurance Agents, Brokers, & Service	0.504***	−29.143***	5,981.658***	30,085.320***
	(0.159)	(8.930)	(1,434.017)	(7,853.826)
Insurance Carriers	0.512***	−4.646	2,329.053***	5,869.482**
	(0.059)	(3.317)	(532.658)	(2,917.264)
Nondepository Institutions	0.214**	−22.280***	3,887.783***	20,186.150***
	(0.088)	(4.947)	(794.446)	(4,351.021)
Real Estate	0.253***	−29.265***	4,690.097***	20,885.350***
	(0.080)	(4.510)	(724.223)	(3,966.423)
Security & Commodity Brokers	0.262***	7.088*	−495.668	−6,321.357*
	(0.071)	(4.004)	(642.901)	(3,521.037)
Constant	1.365***	−86.955***	15,349.670***	71,943.280***
	(0.048)	(2.709)	(434.947)	(2,382.117)
Observations	7,053	7,053	7,053	7,053
R^2	0.018	0.028	0.026	0.024
Adjusted R^2	0.016	0.027	0.024	0.023
Residual Std. Error (df = 7040)	1.470	82.585	13,261.220	72,629.060
F Statistic (df = 12; 7040)	10.663***	17.153***	15.592***	14.605***

Note: *p<0.1; **p<0.05; ***p<0.01 .

HU Score is #Positive words - #Negative Words from Liu and Hu(2004), LM Score is #Pos-#Neg, LM_Uncertain and LM_Litig are from LM's website at https://sraf.nd.edu/textual-analysis/resources/#LM

In a toy example shown in Figure 1 we see three firms writing 10-Ks. Two are in the oil business and one is in the beverage business. There is a distribution of the topic oil price risk over all of the risk factor statements. Some documents will have a good deal of discussion about the topic "oil price risk" (a high θ_d), and others will have little or no discussion of the topic (a low θ_d). Finally, for each word in the document (say, oil, water, pipeline) it is assigned a probability of belonging to a topic.

The first firm in our example has a 10-K with three topics: oil risk, environmental risk, and some other risk common to call companies. For now, this common topic is just something like "all investments are risky." The second firm has a topic about oil price risk, a new topic called retailing, and the common risk topic. The third firm has water risk, retailing risk, and the common risk. Each of the firms has a different proportion of their 10-K taken up with a different set of topics. In terms of the probability of a word belonging to each topic, the word oil can potentially belong to any topic. However, we see in the example that oil belongs to oil risk with a probability 0.98, retailing with a probability of 0.01, water risk with the probability 0.001, and pipes with a probability of 0.009. The residual topic is the common boilerplate statements that risk is everywhere. In this example, the word oil has a zero probability of being in common risk topic.

This model described above is the probabilistic LDA topic model as there is some probability distribution for both topics in a document and words belonging to a topic. What makes this a structural topic model is the use of additional variables that provide information about the firm writing the document or the time period the document was written. Roberts et al. (2016) provide descriptions of the method in detail, but it essentially changes two assumptions. Under LDA the $\theta_d \sim Dirichlet(\alpha)$), but using an STM the θ_d is now drawn from a Log Normal(V) where the V is a vector of document specific data. In our case this is the meta data described in Table 3. In addition, the β_k were assumed to be common across the set of 10-Ks under LDA. In the STM, however, a multinomial logit is used for word distributions where the word's probability of being in a topic is based on the topic, the document specific covariates (V), and interactions between the covariates and the topic. Kim (2018).

Four major benefits exist for employing STM rather than estimating a topic model under LDA and then regressing independent variables against topic proportions θ_d or the probability a word is in a topic β_k. The first benefit is that the STM accounts for correlations among the topics and

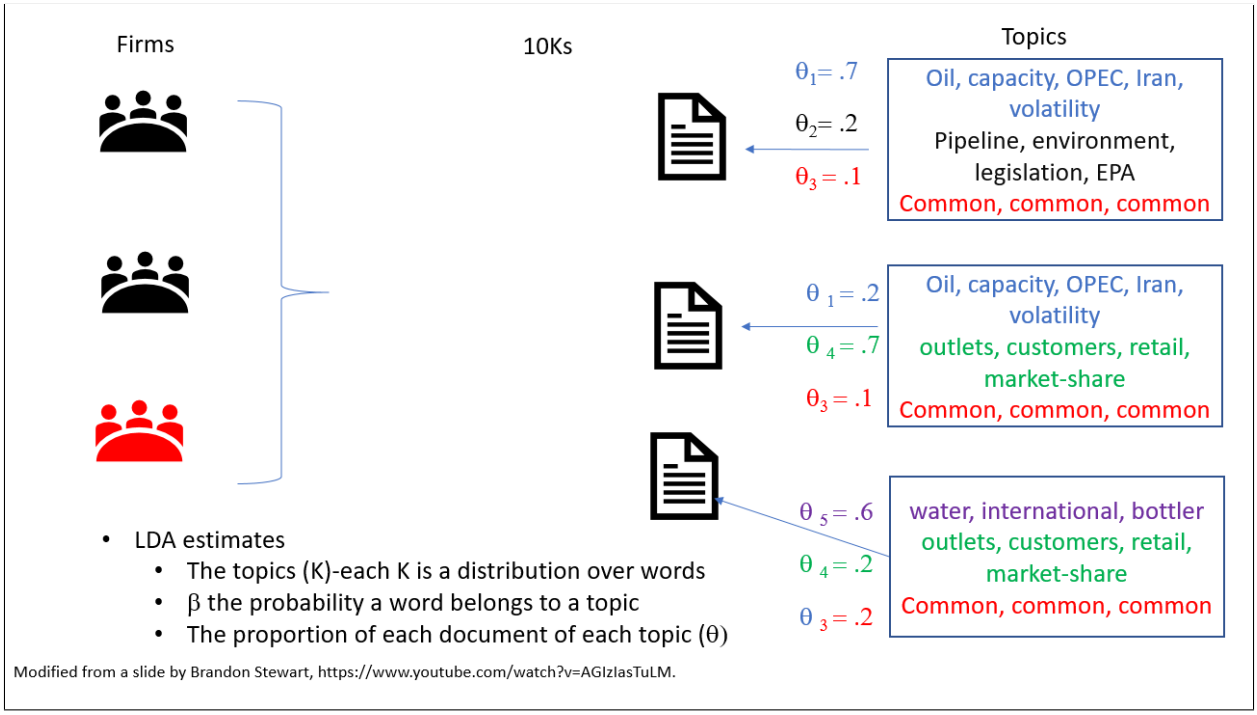

Figure 1: Topic Models

LDA does not. Secondly, the STM allows for topic vocabulary to vary by covariates. Thirdly, LDA can miss covariates relationships within the data. Finally, STM includes measurement uncertainty in the estimation of the latent topics. Reich et al. (2015)

STM techniques have been used in education to gather information about student evaluations in MOOCs (Reich et al. (2015), detecting fraud in medical prescriptions (Zafari and Ekin (2018)), and media bias (Kim (2018)).[5] The purpose of this paper is to examine the topics in the 10-Ks to see how they vary with important information about the company. A second purpose is to find firms that differ from their cohort (SIC, Year) to see if they have a different risk exposure. This paper will eventually test the effects of being an outlier on future financial metrics, but at present, it provides the topic extraction and initial analysis of the topics contained in the 10-Ks. [6]

The STM provides a set of interpretable topics and a representation of the proportion of each document in the set of 10-Ks according to those topics $\theta_1, \ldots, \theta_D$. Many researchers have used

[5]For a list of additional papers using STM see Published Applications section of the STM package's home page: https://www.structuraltopicmodel.com/.

[6]There is an issue with using topic models in that the number of topics is chosen by the user. Some researchers such as Huang and Li (2011) have used 25 topics in their analysis, but one can test for an optimal number of topics using methods like Griffiths et al. (2004) found in the ldatuning package in R.[7] Using their techniques it appears that there are 85 topics in the data.

this as a topic predicting process, but like Roberts et al. (2016) we use this to analyze the current set of 10-Ks.

Appendix B shows the list of the topics generated by the model. The naming of these topics was undertaken "by hand" in the sense that two different output measures of word membership were used to derive the topic name. The first is the probability (β) of a word belonging to the topic. Using a list of the topic 50 words in order of probability I deduced what the topic was based on these word probabilities. The second, method uses the FREX score. The FREX score is based on a notion of both the frequency of appearance in a topic and the exclusivity of a word in the sense that words are likely to only be in that particular topic **?**.[8]

5 Results

First, we present the top 20 topics ordered by prevalence or the predicted θ_i. 2 shows the plot of the top 20 topics and the words associated with the topic. These words the highest β for the topic. Another way to look at this is to look at the distribution of words by each topic. 3 shows the histogram distribution over documents of the first 21 topics as an example of the distributions of topic θ in the corpus. One thing to note is that if there is a spike at 0 and 1, we would be seeing a topic that strongly classifies documents. We do not see that in this corpus at all. Mostly we have spikes at close to zero as many topic have a low probability of occurrence. Some topics are rarely mentioned (low θ) such as Energy or Operations. Alternatively, some like accounting misstatements have a distribution that suggests some documents are more likely to have this topic. Looking through the set of 85 topics, I have found 15 with distributions with a more pronounced distribution of θ. These is shown in 4.

Of these 16 topics shown in 4, I then choose six for illustrative purposes to show the results of the prevalence regressions. These regressions examine the effect of SIC code sector, year, total assets and a reading score on the topic prevalence θ per document i. The prevalance regression is estimated as a multivariate logit.

$$\theta_i = \alpha + \beta * ReadingScore_i + \phi * TotalAssets + \sum \beta_s * SIC_s + \sum \delta_t Year_t + \epsilon_i \qquad (1)$$

[8] A more formal process for naming topics will be developed to remove the idiosyncratic possibility of a single individual deriving the topic names

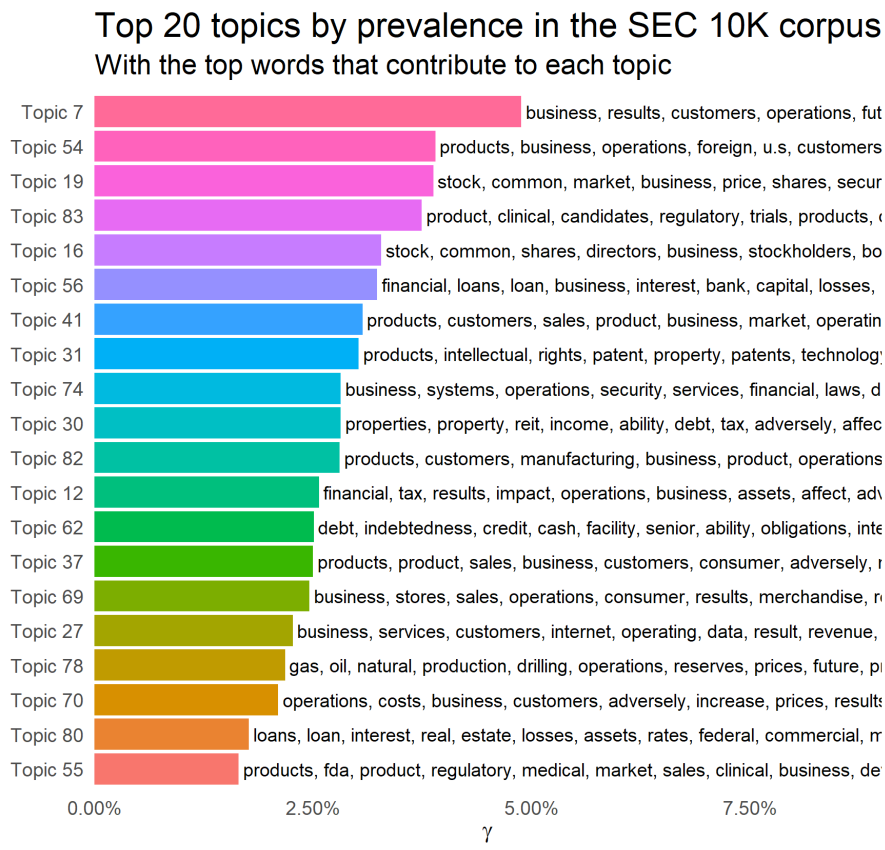

Figure 2: Topic 20 Topics by Mean (γ)

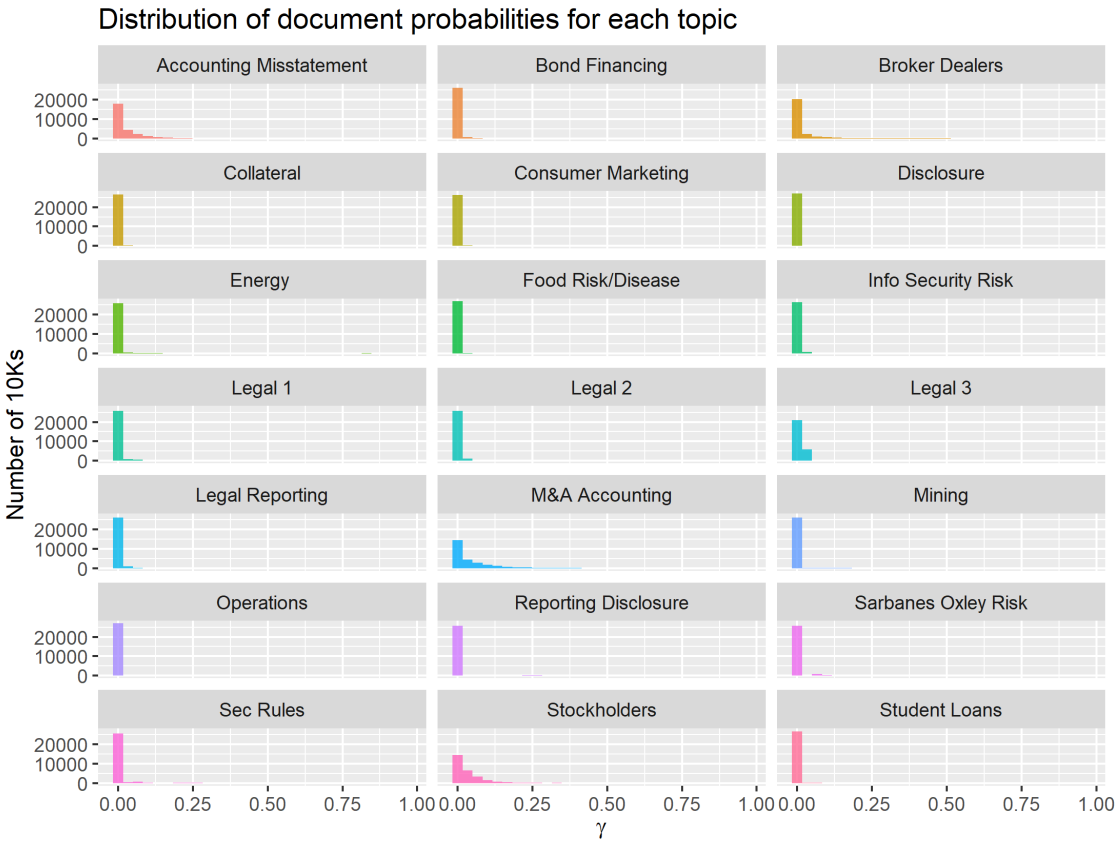

Figure 3: Topic Probabilities for Topics 1-21

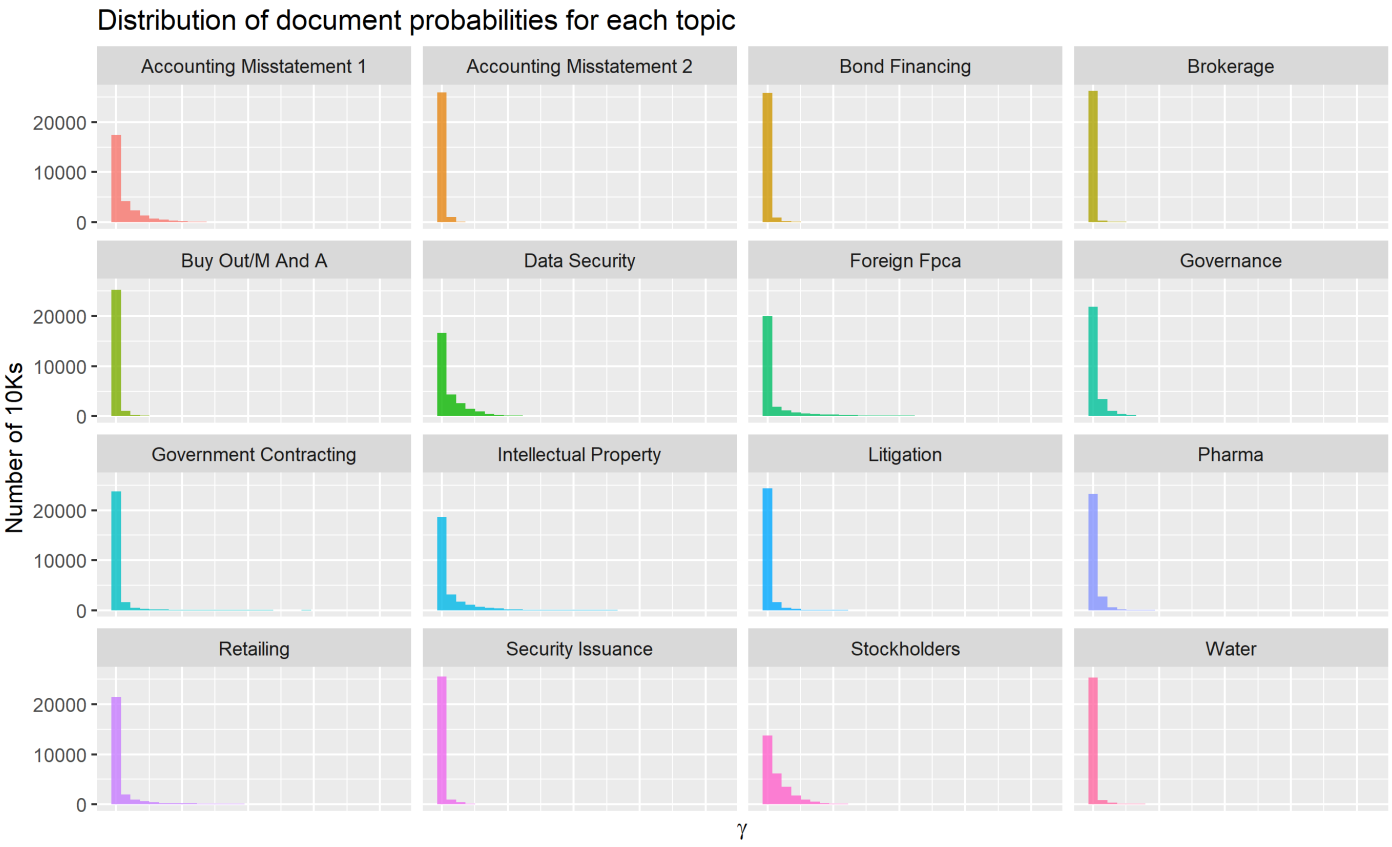

Figure 4: Topic Probabilities for Topics of Interest

The six topics chosen for this initial analysis include Accounting Misstatements #1, (6), Stockholder issues (7), Litigation (8), Governance (9), and Intellectual Property (10). What is striking about the results so far is that while some years have significant effects, none of the sectors seem to have influence on the prevalence. In addition, the size of the firm in terms of total assets is not related to topic prevalence, but the reading score is almost uniformly related to prevalence.

Table 6: Topic 12: Accounting Misstatement

Variable	Beta estimate	Std Err.	t-stat	p value
Intercept	-0.041	0.0059	-6.9051	0.0000
2012	-0.0021	0.0017	-1.251	0.2109
2013	-0.002	0.0016	-1.246	0.2128
2014	-0.0038	0.0017	-2.2119	0.0270
2015	-0.0044	0.0016	-2.709	0.0068
2016	-0.0039	0.0016	-2.4247	0.0153
2017	-0.0036	0.0016	-2.3103	0.0209
Agriculture, Forestry, Fishing	-0.0033	0.0125	-0.2605	0.7945
Construction	-0.0045	0.0057	-0.7806	0.4350
FIRE	-0.0048	0.004	-1.1944	0.2323
Manufacturing	-0.0031	0.0039	-0.7886	0.4303
Mining	-0.0042	0.0042	-0.9857	0.3243
Retail Trade	-0.0031	0.0042	-0.7331	0.4635
Services	-0.0049	0.0041	-1.1892	0.2344
Trans, Comms, Utilities	-0.0034	0.0042	-0.8088	0.4186
Unclassified	-0.0046	0.0162	-0.2846	0.7759
Wholesale Trade	-0.0019	0.0047	-0.3924	0.6948
cusip_tot_asset	0.0000	0.0000	0.3745	0.7080
Flesch.PSK	0.0083	0.0006	15.1113	0.0000

6 Summary:

This is a preliminary investigation of the Item 1A Risk Factor using a structure topic model approach. There are a number of possible contributions that may be made. First, from an ERM perspective it would be helpful to use the universe of data on how firm's discuss risk to determine if another company with a risk exposure sees the risk exposure and discusses that risk in its 10-K. The second potential benefit is to determine whether meta data increase our understanding of risk topics and how they evolve over time. I use only a subset of the meta data available for simplicity. However, I have access to more refined SIC code levels, and the state of incorporation which may be related to certain governance issues. In addition, there is additional financial information that

Table 7: Topic 16: Stockholders

Variable	Beta estimate	Std Err.	t-stat	p value
Intercept	0.0128	0.0063	2.0281	0.0426
2012	0.0006	0.0018	0.3229	0.7468
2013	0.0003	0.0017	0.1762	0.8602
2014	0.0030	0.0017	1.7595	0.0785
2015	0.0028	0.0019	1.4772	0.1396
2016	0.0027	0.0018	1.4930	0.1354
2017	0.0014	0.0017	0.8253	0.4092
Agriculture, Forestry, Fishing	0.0050	0.0119	0.4225	0.6727
Construction	0.0023	0.0057	0.4034	0.6866
FIRE	0.0005	0.0039	0.1404	0.8883
Manufacturing	0.0006	0.0039	0.1649	0.8690
Mining	0.0009	0.0042	0.2043	0.8381
Retail Trade	0.0000	0.0041	0.0032	0.9974
Services	0.0007	0.0040	0.1828	0.8550
Trans, Comms, Utilities	-0.0006	0.0040	-0.1409	0.8879
Unclassified	0.0125	0.0195	0.6429	0.5203
Wholesale Trade	-0.0013	0.0049	-0.2601	0.7948
cusip_tot_asset	0.0000	0.0000	0.0586	0.9533
Flesch.PSK	0.0015	0.0006	2.6878	0.0072

can be used to provide additional information (such as R&D expenditures or litigation reserves). One can also then see if risk topics rather than curated dictionaries are more closely related to financial performance. What is interesting about this, is that while a curated dictionary is like an expert system, a topic model can see links that people may not observe and may prove to be a better way of classifying risk topics.

Status of the project.

Data from 2004 to 2010 need to be cleaned. When this is finished the data set will be doubled in size and will have a panel from 2005 to 2017. Additional financial performance data needs to be merged into the data. I have obtained a number of balance sheet and income statement items. Also I plan to merge information about the number of analysts following the company to determine whether risk clarity is influenced by the number of experts reading the 10-K. I also captured the Litigation Risk Section of he 10-K and will likely analyze that in conjunction with Section 1A.

The work will be mostly finished by May as I am just waiting on the data cleaning process in order to continue the project.

Table 8: Topic 22: Litigation

Variable	Beta estimate	Std Err.	t-stat	p value
Intercept	0.0223	0.0048	4.661	0.0000
2012	0.0004	0.0012	0.3165	0.7516
2013	0	0.0013	0.0248	0.9802
2014	-0.0001	0.0014	-0.0651	0.9481
2015	0.0015	0.0014	1.0848	0.2780
2016	0.002	0.0015	1.3708	0.1704
2017	0.0021	0.0013	1.6414	0.1007
Agriculture, Forestry, Fishing	-0.0005	0.0088	-0.0611	0.9513
Construction	-0.0011	0.0042	-0.2613	0.7939
FIRE	0.0007	0.0031	0.2276	0.8200
Manufacturing	-0.001	0.0031	-0.3077	0.7583
Mining	-0.0009	0.0035	-0.2731	0.7848
Retail Trade	-0.0002	0.0035	-0.0634	0.9494
Services	0.0008	0.0033	0.2475	0.8045
Trans, Comms, Utilities	-0.0008	0.0032	-0.2534	0.7999
Unclassified	-0.0055	0.0115	-0.4784	0.6324
Wholesale Trade	-0.0006	0.0041	-0.1564	0.8757
cusip_tot_asset	0	0	-0.4045	0.6858
Flesch.PSK	-0.0016	0.0005	-3.466	0.0005

Table 9: Topic 24: Governance

Variable	Beta estimate	Std Err.	t-stat	p value
Intercept	-0.0173	0.0048	-3.5967	0.0003
2012	-0.0001	0.0014	-0.0507	0.9595
2013	0.0008	0.0013	0.6074	0.5436
2014	0.0034	0.0015	2.2783	0.0227
2015	-0.0008	0.0015	-0.5612	0.5747
2016	-0.0025	0.0015	-1.6751	0.0939
2017	-0.0025	0.0014	-1.7987	0.0721
Agriculture, Forestry, Fishing	0.0012	0.01	0.1249	0.9006
Construction	0.002	0.0055	0.3728	0.7093
FIRE	-0.0009	0.0032	-0.2727	0.7851
Manufacturing	-0.001	0.0031	-0.3294	0.7419
Mining	-0.0025	0.0032	-0.7857	0.4321
Retail Trade	-0.0024	0.0036	-0.6532	0.5137
Services	-0.002	0.0032	-0.6146	0.5389
Trans, Comms, Utilities	-0.0015	0.0033	-0.4584	0.6467
Unclassified	0.0073	0.0159	0.4619	0.6442
Wholesale Trade	-0.0001	0.0044	-0.0256	0.9796
cusip_tot_asset	0	0	-0.412	0.6804
Flesch.PSK	0.0038	0.0005	8.1711	0.0000

Table 10: Topic 31: Intellectual Property

Variable	Beta estimate	Std Err.	t-stat	p value
Intercept	-0.0518	0.0091	-5.6947	0.0000
2012	0.0010	0.0023	0.4324	0.6655
2013	0.0021	0.0021	1.0027	0.3160
2014	0.0016	0.0023	0.7108	0.4772
2015	0.0037	0.0023	1.5787	0.1144
2016	0.0026	0.0024	1.0869	0.2771
2017	0.0041	0.0026	1.5947	0.1108
Agriculture, Forestry, Fishing	-0.0033	0.0165	-0.2028	0.8393
Construction	0.0042	0.0081	0.5215	0.6020
FIRE	0.0018	0.0054	0.3367	0.7363
Manufacturing	0.0008	0.0056	0.1375	0.8906
Mining	0.0031	0.006	0.5249	0.5996
Retail Trade	0.0035	0.0062	0.5682	0.5699
Services	0.0019	0.0057	0.3269	0.7437
Trans, Comms, Utilities	-0.0005	0.0062	-0.0799	0.9363
Unclassified	0.0165	0.0294	0.5632	0.5733
Wholesale Trade	-0.0017	0.0071	-0.2411	0.8095
cusip_tot_asset	0.0000	0.0000	0.1671	0.8673
Flesch.PSK	0.0091	0.0008	11.5361	0.0000

Bibliography

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, "Measuring economic policy uncertainty," *Quarterly Journal of Economics*, 2016.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo, "quanteda: An R package for the quantitative analysis of textual data Sofware Review Depositary Archive," *Journal of Open Source Software*, 2018, *3* (30), 774.

Blei, D. M. and J. D. Lafferty, "Topic Models," in "Text Mining: Classification, Clustering, and Applications" 2009.

Campbell, John L., Hsinchun Chen, Dan S. Dhaliwal, Hsin min Lu, and Logan B. Steele, "The information content of mandatory risk factor disclosures in corporate filings," *Review of Accounting Studies*, 2014.

Chaney, Ajb and Dm Blei, "Visualizing Topic Models.," *Icwsm*, 2012.

Feldman, Ronen, Suresh Govindaraj, Joshua Livnat, and Benjamin Segal, "Management's tone change, post earnings announcement drift and accruals," *Review of Accounting Studies*, dec 2010, *15* (4), 915–953.

Ganguly, Arup, "TEXTUAL DISCLOSURE IN SEC FILINGS AND LITIGATION RISK." PhD dissertation sep 2018.

Griffiths, Thomas L, Mark Steyvers, by Blei, and Jordan Blei, "Finding scientific topics A first step in identifying the content of a document is determining which topics that document addresses. We describe a generative model for documents, introduced," Technical Report 2004.

Gunning, Robert, "The Fog Index After Twenty Years," *Journal of Business Communication*, jan 1969, *6* (2), 3–13.

Gunratan Lonare, Maintainer, "Edgar, R Software Package," 2017.

Huang, Ke-Wei and Zhuolun Li, "A multilabel text classification algorithm for labeling risk factors in SEC form 10-K," *ACM Transactions on Management Information Systems*, oct 2011, *2* (3), 1–19.

IRRC Institute, "The Corporate Risk Factor Disclosure Landscape," Technical Report January, JOHN L. WEINBERG CENTER FOR CORPORATE GOVERNANCE, University of Delaware, Newark, DE 2016.

Kim, Sung Eun, "Media Bias against Foreign Firms as a Veiled Trade Barrier: Evidence from Chinese Newspapers," *American Political Science Review*, nov 2018, *112* (04), 954–970.

Kuhn, Kenneth D., "Using structural topic modeling to identify latent topics and trends in aviation incident reports," *Transportation Research Part C: Emerging Technologies*, feb 2018, *87*, 105–122.

Li, Feng, "Annual report readability, current earnings, and earnings persistence," *Journal of Accounting and Economics*, aug 2008, *45* (2-3), 221–247.

Loughran, Tim and Bill Mcdonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance*, 2011.

_ **and** _ , "Measuring readability in financial disclosures," *Journal of Finance*, 2014.

Rawte, Vipula, Aparna Gupta, and Mohammed J. Zaki, "Analysis of year-over-year changes in Risk Factors Disclosure in 10-K filings," in "Proceedings of the Fourth International Workshop on Data Science for Macro-Modeling with Financial and Economic Datasets - DSMM'18" ACM Press New York, New York, USA 2018, pp. 1–4.

Reich, Justin, Dustin Tingley, Jetson Leder-Luis, Margaret E. Roberts, Brandon Stewart, and Brandon Stewart, "Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses," *Journal of Learning Analytics*, nov 2015, *2* (1), 156–184.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi, "A Model of Text for Experimentation in the Social Sciences," *Journal of the American Statistical Association*, 2016.

_ , _ , **Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand**, "Structural Topic Models for Open-Ended Survey Responses," *American Journal of Political Science*, oct 2014, *58* (4), 1064–1082.

Schrand, Catherine M. and John A. Elliott, "Risk and Financial Reporting: A Summary of the Discussion at the 1997 AAA/F...: EBSCOhost," *Accounting Horizons.*, 1998, *12* (3), 271–282.

SEC, "SEC.gov — General Instructions," 2019.

_ , "SEC.gov — How to Read a 10-K," 2019.

Zafari, Babak and Tahir Ekin, "Topic modelling for medical prescription fraud and abuse detection," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, dec 2018, p. rssc.12332.

7 Appendix A

(1) Part I

- Download SEC Master file using Edgar package from R (Gunratan Lonare (2017))
- Use SEC Masterfile to access 10-Ks
- Download 10-Ks (2004-2017) is about 259GB compressed

(2) Part II

- I extracted the Item 1A Risk Factors section of the 10-K by selecting the text between the two end points that start with "Item 1A. Risk Factors" and ends with "Item 1B Unresolved Comments to Staff." The problem here is that some files have table of contents with these headings, so care har to be taken to obtain the table of contents and the actual section of text needed. Then the table of contents information was stripped. This is not a perfect process, however, one of the topics is really about the table of contents. This shows how capable the topic model is at determining the topics found in the text.
- I removed unwanted text. Each file has a similar structure and the files become more uniform over time. However, many have extraneous characters and HTML codes. I removed common stopwords using r smart stopwords. I also removed the following words:

  remove <-c("x.item","staff","comments","unresolved", "xitem", "1a", "1b", "Item","1A", "1B", "item","x47","a","s","e","k","risk factor", "risk factors").
- After cleaning the text I merged meta data with text. The meta data I obtained from the first page of the 10K, compustat (ebita and total assets), and by using packages such as Quanteda (Benoit et al. (2018)) to obtain information about the document such as sentiment, size of document, and dictionary word counts.
- Using Quandeta I created an stm object for the structured topic model package.
- I then estimated structural topic model using the STM package developed by Roberts et al. (2016).
- Clean code will be provided when the project is finished.

8 Appendix B Topics and Reference Number

Topic Number	Topic	Topic Number	Topic
1	Legal 1	46	SEC Filings
2	Student Loans	47	Water
3	Food Risk/Disease	48	Pharm Retail
4	Info Security Risk	49	Regulation
5	Collateral	50	Transportation
6	Reporting Disclosure	51	Boiler Plate
7	M&A Accounting	52	Banking Operations
8	Operations	53	Banking Capital Standards
9	Bond Financing	54	Foreign FCPA
10	Legal 2	55	Med Tech
11	SarbanesOxley Risk	56	Consumer Banking
12	Accounting Misstatement	57	Consumer Credit
13	Mining	58	Buy Out/M and A
14	Disclosure	59	Telecom
15	Legal 3	60	Farm Chemo
16	Stockholders	61	Gaming
17	Consumer Marketing	62	Bond Refinancing
18	Legal Reporting	63	Advertising Effectiveness
19	Broker Dealers	64	Smart Glass Company Specific
20	SEC Rules	65	Financing
21	Energy	66	Cruise Hospitality
22	Litigation	67	Supplements
23	Real Estate	68	Automobile
24	Governance	69	Merchandising
25	Government Contracting	70	Metal Industries
26	Solar Tech	71	Bank Capital
27	Internet	72	M and A
28	Pharma	73	Energy Lps
29	Financial Crises	74	Data Security
30	REITS	75	Chemical
31	Intellectual Property	76	Outsourcing
32	Reinsurance	77	Airlines
33	Boilerplate	78	Fracturing
34	Mortgage Lending	79	Banks
35	Medical Provider	80	Saving Banks
36	Security Issuance	81	Homebuilding
37	Retailing	82	Semi Conductors
38	Accounting Misstatement	83	Pharma R&D
39	Overseas Investment	84	Pharma Regulation
40	Fast Food	85	Sat TV
41	Internet Technology		
42	Energy Production		
43	China		
44	Agriculture		
45	Brokerage		