Spring 1-1-2019

Talking About Regulation in 10-K Annual Reports; Uniformity in a Naive Sample.

Katerina Ruth Mills
University of Colorado at Boulder, katerina.r.mills@gmail.com

Follow this and additional works at: https://scholar.colorado.edu/fnce_gradetds
Part of the Finance Commons

Recommended Citation
https://scholar.colorado.edu/fnce_gradetds/14

This Thesis is brought to you for free and open access by Finance at CU Scholar. It has been accepted for inclusion in Finance Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
TALKING ABOUT REGULATION IN 10-K ANNUAL REPORTS;

UNIFORMITY IN A NAIVE SAMPLE.

by

KATERINA RUTH MILLS, ESQ.,
B.A., University of California Los Angeles, 2011
J.D., University of Denver, 2013
M.A., University of Denver, 2014
B.A., University of Denver, 2015

A Thesis submitted to
the Faculty of the Leeds School of Business
University of Colorado, Boulder in partial fulfillment
of the requirements for the Degree

Master of Science

Department of Finance

2019
This thesis entitled:
Talking about Regulation in 10-K Annual Reports; uniformity in a naive sample.
Written by Katerina Ruth Mills
has been approved by the Department of Finance

______________________________
Professor Edward Van Wesep, Committee Chair

______________________________
Professor Diego Garcia, Committee Member

______________________________
Professor Tony Cookson, Committee Member

Date: __________________________

The final copy of this thesis has been examined by the above signatories, and we find that both the content and form meet acceptable presentation standards of scholarly work in the above-mentioned discipline.
Abstract

This project seeks to highlight the difficulties of employing required annual reports in analyses attempting to tie left-hand side outcomes, whether past or future, through the use of Natural Language Processing techniques to analyze firm discussion of regulators, regulation, laws, and other regulatory regimes in the context of required 10-K annual disclosures under U.S. public company reporting regimes governed by the U.S. Securities and Exchange Commission (SEC). While Natural Language Processing (NLP) techniques have gained popularity in turning text into data facilitating a multitude of varied new analytical projects in the fields of academic corporate research, using NLP mechanisms particularly for regulation-oriented corporate speech analysis presents relative uniformity across filers and industries which this small project seeks to highlight as a possible burden for NLP usage in this particular extension of the legal and financial reporting analyses. The small sample, simple word lists, and naive comparison employed here seeks to highlight, through simple methodology, that even an industry-by-industry analysis method may be less than meaningful in addressing regulatory, industry standard, and legal practice nuances defining regulatory discussion due, at least in part, to uniform reporting summaries (relative uniformity) across firms, not only those participating in a single industry, but across SEC annual filers generally. This project will conclude with a brief summary of possible causal factors in the highlighted report uniformities while refraining from any implication that regulatory or legal comparison is or should be among the comparative factors for which design and response uniformity guidelines for 10-K reports are generated or sought by the SEC for use by potential or current investors as an indicative factor of firm performance or future performance.

All you really need to know for the moment is that the universe is a lot more complicated than you might think, even if you start from a position of thinking it’s pretty damn complicated in the first place.

Douglas Adams
Acknowledgments

Special and unending thanks to Professors Edward D. Van Wesep, Diego Garcia, and J. Anthony (Tony) Cookson whose patience, guidance, and expertise have been and continue to be invaluable in constructing and finishing this project, through the past two years of my graduate study, and, importantly, in my choice and ability to defend this small thesis. Special personal thanks to Stefanie Ungphakorn, without whom the last two years would not have been surmountable.

Additional thanks go to those who are responsible for maintaining the Securities and Exchange Commission and reporting specific regulatory databases, summaries, and filing practice handbooks employed to finish this project. Without their efforts and provision, navigating the complex system of regulations and regimes over the entities and firms reporting their 10-K filings each year would be all but impossible.

Thanks, no matter how special or unending, could never be enough to truly thank my mother and father for their constant support, guidance, and unending patience. And finally, particular thanks to my wonderful spouse, Michael, whose unconditional support and understanding are constant and without limit.

“You know,” said Arthur, “it’s at times like this, when I’m trapped in a Vogon airlock with a man from Betelgeuse, and about to die of asphyxiation in deep space that I really wish I’d listened to what my mother told me when I was young.”

“Why, what did she tell you?”

“I don’t know, I didn’t listen.”

Douglas Adams
"I thought," he said, "that if the world was going to end we were meant to lie down or put a paper bag over our head or something."
"If you like, yes," said Ford.
"Will that help?" asked the barman.
"No," said Ford and gave him a friendly smile.
Douglas Adams
LIST OF TABLES AND FIGURES

All Filers Data Table ........................................................................................................ 16

All Filers Figures

Total 10-K Reports Filed ..................................................................................................... 17
Average Total Paragraphs per Filings .................................................................................. 17
Average Regulatory Paragraphs per Filings ........................................................................ 18
Average Regulatory Words Flagged per Filings ................................................................. 18
Average Regulatory Paragraphs Total Length per Filing All years .................................. 19
Average Regulatory Paragraphs Total Length per Filing 1998-2009 ............................. 19
Average Total Number of Blame Words per Filing ........................................................... 20

SIC Comparison Figures

Total Filings ......................................................................................................................... 22
Average Paragraphs ............................................................................................................ 22
Average Regulatory Paragraphs ......................................................................................... 23
Average Regulatory Words Flagged .................................................................................. 23
Average Regulatory Paragraph String Length .................................................................. 24
Average Blame Words Counted in Regulatory Paragraphs ............................................. 24

“A learning experience is one of those things that says, 'You know that thing you just did? Don't do that.'”

Douglas Adams
INTRODUCTION

10-K disclosures have been the focus of numerous and varied Natural Language Processing (NLP) endeavors inside finance and accounting literature since NLP tools, as mechanisms for turning words into data, infiltrated the academic corporate research community. This small project seeks to highlight and describe an embedded framework of report uniformity which may require special attention or additional specific controls or disclaimers for the promulgation of NLP techniques in the analysis of regulatory speech in these uniform corporate disclosure documents to provide useful insight into left-hand side outcomes, whether industry or filer specific. This small thesis is an initial step in opening the door to a detailed discourse on the informativeness or lack thereof, of regulatory speech within these documents. Originally undertaken in the context of a larger research design seeking to show to opposite, mainly that a nuanced approach to regulatory speech was necessary due anticipated stark differences in the industry and filer differences in regulatory discussion due to applicable regulatory frameworks, this project seems instead to support the view that regulatory speech in these filings is highly uniform across industries even given the varying regulatory frameworks to which they are subject. Using a simple general regulatory word list paired with a list of words generally indicating responsibility or effect, the simple NLP methodology undertaken here highlights the issues associated with what may be boilerplate or response uniformity in analyzing firm regulatory discussions in the context of annual reporting.

This project analyses 10-K annual reports filed over a 14-year period, 1997 through 2010, employing a paragraph-by-paragraph analysis method to locate and grab paragraphs discussing general regulatory topics and analyze those paragraphs for the number of words indicating effect, implication, and/or responsibility. In short, two naive word lists were created in an effort to identify firm discussion of regulatory implications, burdens, or regulatory “blame” in an initial inspection of filer-wide uniformity in discussions of regulators, laws, and other regulatory regimes which effect, or may be included or discussed by firms as affecting firm productivity, regulatory and or tort risk, among other negative firm implications,
restrictions, or limitations. To this aim, general, non-industry non-regulation specific words were chosen with the intent to get a broad, naive and simple sample for a swift, and easy comparison between industries. General filer-wide summary statistics and temporal changes have been provided, with explanation, as well as a surface level industry analysis and comparison with accompanying explanatory detail for three selected industry subgroups. The comparative subgroups were chosen with specific reference to their regulatory regime differences, obligations, and differences in industry standard and practice. The conclusory bulk of this project suggests several structures and practices with regard to these regulatory report filings which may be at play in creating uniformity in sections devoted to regulatory discussion, boilerplate, or the explanation of risks included in these filings.

*Don’t Panic.*

*Douglas Adams, Hitchhiker’s Guide to the Galaxy*
SECTION I

Brief Literature Review

While there are many applications for and analyses of the words and data included in 10-K EDGAR filings, the study of regulation-oriented speech and disclosures is only a fringe use of the expanded interest in Natural Language Processing (NLP) techniques in corporate, legal, and business research communities. While the Corporate, Finance, Accounting research community have looked at 10-K filings and their varied data and text components as meaningful determinants of left-hand side outcomes for years, regulatory speech and legal regime-oriented speech has been predominantly absent from these analyses. Legal regime reactions and speech outside of corporate or reporting firm outcomes, while perhaps interesting to the legal research community and general Law and Economics studies of corporate disclosure, requirement, and speech; the focus of these research communities thus far has been on corporate political speech following Citizen’s United (2010) and the complicated theoretical and jurisprudential orientation towards the identity of corporate entities and the rights or obligations with regard to their shareholders or ultimate owners. Specific studies oriented at tying firm discourse on regulators and the legal frameworks to which they are subject within the required disclosure framework is absent in the major corporate and legal research communities. This literature summary begins summarizing the foundational works by Tim Loughran and Bill McDonald in applying NLP tools and techniques to financial reports and corporate arenas and summarizes the current focus of other applicable research regimes within the Finance, Accounting, General Business, and Legal research communities. Currently, there is not a robust foundation of applicable projects from which this project can claim to originate, instead the simple summary provided seeks to explicate possible issues with extending the analysis of legally oriented disclosure language in applications similar to those included in this literature summary.

From the corporate finance, accounting, and general business perspective Tim Loughran and Bill McDonald are the fathers of the application of Natural Language Processing (NLP) techniques in finance arenas, reporting, and analysis. Their projects include, but are hardly limited to; studies detailing the
readability of corporate disclosures using traditional semantic indexes, the mislabeling of positivity and negative sentiment in word list applications in financial reports, and several summaries and discourses on the appropriateness, limitations, and applicability of word list usage and NLP techniques in finance and accounting (Loughran and McDonald, 2009, 2011, 2014, 2015, 2016, 2017, and forthcoming). While neither of the authors would likely approve of the wordlist herein employed for encumberment or burdening as “Blame Words” due to their extensive explanations regarding the inappropriateness of using negative colloquial tone indicators in corporate disclosures, their specific works in the applicability of word lists and textual analysis generally and in the fields of finance and accounting provided the foundation for the structure of this research design (see, Loughran and McDonald, 2015 and 2016). Their applicable works focus on applying NLP techniques to 10-K filings through 1994 - 2007 and 2009, in detailing various implications in speech, rhetoric, and the applicability of plain English tonal indicators and their appropriateness with regard to financial behavior, market movement or status, and firm outcomes. In short, the authors laid the foundation for applying NLP methodologies and orientation to left-hand side outcomes in the context of corporate disclosures and financial reporting speech. In their 2016 summary, Textual Analysis in Accounting and Finance: A Survey, the authors detail useful guidelines and meaningful limitations on the usage of these NLP techniques in an effort to highlight the appropriate orientation of coming authors when using such flexible, and sometimes opaque techniques when making left-hand side suggestions about firm, market, or sentimental outcomes resulting from NLP, word list, word counting, and other NLP based data gathering methods (Loughran and McDonald, 2016). This project, while highlighting the lack of supplemental foundational orientations on law or legal regime specific applications of the Loughran and McDonald methodologies, seeks to provide another highlight suggesting that results oriented at tying left-hand side outcomes to legal or regulatory speech markers provided by NLP methods may require nuanced or supplemented support as Loughran and McDonald have themselves suggested regarding their more general applications.

Specific academic works citing, using, and guided by Loughran and McDonald’s fundamental principles and application methodologies in the Accounting, Finance, and General Business and
Organizational fields have generally focused on tying left-hand side outcomes in financial performance, market orientation, small investor behavior, and other corporate or topically specific data points (like management turn over, governance measures, organizational design, and specific litigation) to language included in corporate disclosures, including, but not limited to, Forms 10-K, 10-Q, and 8-K (a detailed list of SEC Forms and Form usage can be found on the SEC website, link included in the Bibliography section)(summaries included, Loughran and McDonald, 2016; Li, 2010; Fisher, et al. 2016; and Kearney and Liu, 2014). While many works related to or arguing similarly to Loughran and McDonald’s works analyzing textual tone (Jegadeesh 2013), text usefulness to readers (Frankel, et al., 2017, using NLP techniques on conference calls and predictions), and text readability (Li, 2008; also using the FOG index readability measure); others have detailed the textual changes in regulatory filings over the years in length, complexity, and detail (Dyer, et al. 2017; Cazier and Pfeiffer, 2016; among others), and still others have used NLP techniques to study; litigation risks and disclosure components (Nelson and Pritchard, 2007); risk factor disclosures (Campbell, et al., 2014); and financial statement fraud indicators (Goel and Gangolly, 2012). Tangential in argument to the initial call of this project, in seeking to explicate legal analyses surrounding annual reporting, several works have sought, with specific respect to the environmental reporting, standards, and disclosures, to detail the need for explicated and specific frameworks for such disclosures and reporting compliance, a similar argument posed in the concluding remarks to this project’s introductory sections (Peters and Romi, 2013; Eccles, et al. 2012). While one final vein of corporate research seeks to elucidate which topics, should be of interest to investors (Starks, 2009), an additional component pertinent to the selected introductory discussion. Such prior studies and methodologies, while instructing the use of word lists, NLP techniques, and orientation employed in this project, have yet to breach employing these techniques in analyzing regulation, legal regime specific speech.

Legal research pertinent to the subject of corporate discourse on laws, regulation, and corporate speech overall, is scant in the arena of required disclosures. Instead, legal research on corporate speech can be categorized into two separate topical categories; corporate political speech as it relates to the first amendment (pre and post Citizens United (2010)), and corporate speech as it related to preferable treatment
under the laws resulting from campaign contributions and corporate political activism. Corporate personhood and jurisprudential topics regarding corporate or firm rights to free speech can be arranged as those preceding the 2010 U.S. Supreme Court case *Citizens United (2010)* and those following or arguing its validity in relation to the rights of associated parties, with respect to shareholder, manager, director, and overall corporate entity speech freedoms (Redish and Wasserman, 1997-1998; Siebecker, 2006-2007; Winkler, 2006-2007; Bebchuk and Jackson, 2010; Tucker 201-2011; Melone, 2010-2011; ). While additional legal discourses, discuss the validity of restricting corporate speech or curtailing statement forms by the SEC (Page and Yang, 2005-2006), the overarching legitimacy of enforcing Regulation Fair Disclosure (Humes, 2006-2007), and other non-disclosure related reporting obligations and duties arising under U.S. Federal Securities Laws (Backer, 2003). However, firm speech discussing regulators, laws, or other regulatory regimes or risks in corporate disclosure documentation has remained untouched by prior legal research, a fact which motivated the initial design of this inquiry into firm legal and regulatory speech in form 10-K.

**10-K Reporting Requirement and Components**

Not all companies are required to submit annual reports to the Commission’s EDGAR filing database, but those which have undergone their Initial Public Offering (IPO) and have a volume of shareholders numerous enough to be considered “public” or “publicly traded” under the applicable Securities Acts, (defined in greater detail below) definitions are required, under the Acts, to submit annual and sub-annual reports to the Commission. These annual 10-K filings are overseen by the Division of Corporate Finance (DCF) which oversees the filings and registrations required under the Securities Act of 1933, defining the various securities instruments, their sale, and all relevant disclosures and sale reporting and retail, involved in securities promotion in the United States; the Securities Exchange Act of 1934 defining the transaction of such publicly, privately, and or off market traded securities, exchanges, over the counter retailers, and all initial offerors; and the Sarbanes-Oxley Act of 2002 and Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010 which include additional definitional, application, and
reporting guidelines to which these publicly traded firms must adhere. Generally speaking, DCF office seeks to ensure that investors are provided with the information, disclosures, and reports required to make informed decisions on publicly listed and traded companies, stocks, and other securities instruments offering and circulated in the U.S., additional information on the calling, duties, and powers assigned to the DCF office can be found on their webpage, but their divisional agency description reflects the below:

“In support of the Commission’s mission to protect investors, maintain fair, orderly, and efficient markets, and facilitate capital formation, the Division of Corporation Finance seeks to ensure that investors are provided with material information in order to make informed investment decisions, both when a company initially offers its securities to the public and on an ongoing basis as it continues to give information to the marketplace. The Division also provides interpretive assistance to companies with respect to SEC rules and forms and makes recommendations to the Commission regarding new rules and revisions to existing rules.” Taken from the DCF homepage, Last Modified: Jan. 31, 2017.

For the purposes of this project the key facets of the DCF and over all Commission reporting regime is the concerted focus on report uniformity. The Commission, through the DCF, its website, and the various Commission Accounting Offices provide a wide supplement of practices aids, reporting submission guides and guidelines, and other financial reporting and filer oriented structural guides and definitional riders and references. In short, the Commission provides a wide panoply of guides for filers in an effort to enforce relative uniformity in information submitted and report structures employed by firms issuing the required annual filings. What follows here, preceding a description of the data herein included, is a brief summary of the detailed informational guides, sample forms, and informational sections suggested by the Commission for all 10-K filing firms.

Informational Obligations

The annual 10-K report functions as the annual report, supplemental to the required annual report issued directly to shareholders, which is typically much more detailed, relaxed, and also typically written in the company’s tone and presented in concert with the required annual meeting where shareholders elect directors and handle other applicable shareholder and board functions. The 10-K form, unlike the annual

* Links and citations to relevant SEC materials, practice guides, and other material used in this requirement summary can be found in the Bibliography section at rear.
shareholder report, is formulaic and provides a scripted, but comprehensive overview of the company’s business, financial, and other material conditions, as prescribed by the reporting requirements binding all U.S. issuers of publicly traded securities under the 1934 Act. In the Commission’s provided sample form instructions, the “Rules as to Use of Form 10-K” specifically cite the applicable 1934 Act reporting requirements the Form if designed to meet and for what purposes and covering what required annual disclosures the Form is intended to cover under the Section 13 and or 15(d) obligations of the 1934 Act (15 U.S.C. 78m or 78o(d)). Which references obligations under the Act as follows;

“Every issuer of a security registered pursuant [to this Act title] shall file with the Commission, in accordance with such rules and regulations as the Commission may prescribe as necessary or appropriate for the proper protection of investors and to insure fair dealing in the security...such annual report (and such copies thereof), certified if required by the rules and regulations of the Commission by independent public accountants, and such quarterly reports (and such copies thereof), as the Commission may prescribe.” 15 U.S.C. 78m(a) and 15 U.S.C. 78m(a)(2) respectively. Periodic Citation, Cornell Law, https://www.law.cornell.edu/uscode/text/15/78m.

While the 10-Q and other non-annual filings, like the 8-K filing for material events, are more brief, and often take a slightly more free form often providing for the addition of appendage of supplemental forms and materials, the bulk of the 10-K is meant to provide a regulation based summary form for the annual wellbeing of reporting obligated firms in a form and structure which is designed, without flourish, to inform investors of informational parameters and through a form designed to meet the obligated disclosure goals and guidelines provided in the 1934 Act and adhering to the Commission’s calling to keep investors well informed.

Information Included

The Commission’s calling is not only to provide investors with meaningful or material information regarding the securities and companies into which they may invest, but to do so in a way that, when a consumer sees a portfolio of various different stock offerings, whether those already listed, or those os several IPO offerors coming to market, the investor can compare the various investment opportunities using a uniform scale or material measures form, no matter if the issuers are from very different industries. TO this end, the Commission has designed a sample 10-K Form, for use by filers to maintain relative uniformity
both in structure, but also in an effort to create comparability in the material corporate information, accounting reports, and financial statements included. While the Form appears fill in the blank, the actual 10-K filing is done free form and submitted by part through the online Commission repository EDGAR (Electronic Data Gathering, Analysis, and Retrieval system) provided through registered access with the Commission via their agency website. The Commission provides the following directives on completing the Form for filers using the Sample Form 10-K as a guide:

“This form is not to be used as a blank form to be filled in, but only as a guide in the preparation of the report on paper meeting the requirements of Rule 12b-12. Except as provided in General Instruction G, the answers to the items shall be prepared in the manner specified in Rule 12b-13.” [FORM 10-K ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934, General Instructions, https://www.sec.gov/about/forms/form10-k.pdf; Rule 12b references timeliness guidelines dependent on Filer Size as prescribed under the 1934 Act as amended]

The Form itself then details four subparts which included within-section subparts for applicable material information to be included, detailed, and explained. [See pages 8-11 on the Sample Form]. The sections lay out the filing formula, or required sectional inclusions deemed informationally material of necessary by the Commission, including information on company history, organizational structures, equity plans, subsidiaries, executive compensation, and relevant financial information and statements. A greater explication of the information sought in 10-K filings is detailed in the following section. To this end the Commission provides supplemental guidelines and practice handbooks to report authors specifying the specific labels and information components to be included in each section. These guidelines include both specific guides for the 10-K report as well as a full Financial Reporting Handbook, which provides equivalent detailed guidance on each information submission required or suggested by the Commission for all public or otherwise registered companies. These guides provide a highly formulaic and very detailed view of how and in what format and structure information for filers under the 1934 Act should compile, report, and display their information for investor consumption. The formulaic nature of the partitions and subsections is highlighted by the Commission’s investor guidance on “How to Read a 10-K,” provided on the Commission website, detailing the information included, expected, and required within each of the Commission’s defined sections.

The 10-K filing system has been designed, to this end, to ensure that investors have accessible uniformity in the information they are being presented with regard to each possible offered securities investment available to them in the U.S. market, meaning firms from different industries are funneled into a uniform filing system oriented at
providing a uniform and consistent presentation of material firm information. This intended uniformity, as will be
detailed in this short project below, is highlighted.

**Uniformity for Comparison**

While clear from some quotations highlighted above, the Commission requires, structures, and provides
practice and submission guidelines for 10-K filings so that their contents can be uniformly compared across firms,
industries, and operating characteristics. However, while the reporting guidelines speak to more than just creating
uniformity in report structure alone, their components imply that there is a specific subset of firm related information
appropriate or applicable to investor decision making and by extension firm performance as determined by the
Commission. The 10-K is divided into sections by topic and applicable headers for each discussion of required on all
10-K forms submitted. The 10-K is broken down into four sections, each “Part” having to do with a particular
granularity of overall firm performance. A brief title list of the Parts and subpart or “Items” is provided below,
regulatory or legally relevant Item text summaries have been included, courtesy of the Commission’s “How to Read
a 10-K” Investor guide. Summary text sections have been included and emphasis has been added to Item sections and
subtopics where companies tend to discuss legal issues, regulation, regulators, and topics including legal words picked
up by the NLP techniques used in this project.

**PART I**

**Item 1 - “Business”** requires a description of the company’s business, including its main
products and services, what subsidiaries it owns, and what markets it operates in. This
section may also include information about recent events, competition the company faces,
*regulations that apply to it*, labor issues, special operating costs, or seasonal factors. This
is a good place to start to understand how the company operates.

**Item 1A - “Risk Factors”** includes information about the most significant risks that apply
to the company or to its securities... 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.

Item 1B - “Unresolved Staff Comments”.

**Item 2 - “Properties”**.

**Item 3 - “Legal Proceedings”** requires the company to include information about
significant *pending lawsuits or other legal proceedings, other than ordinary litigation*.

**PART II**

Item 5 - “Market for Registrant’s Common Equity, Related Stockholder Matters and Issuer
Purchases of Equity Securities”

Item 6 - “Selected Financial Data”

Item 7 - “Management’s Discussion and Analysis of Financial Condition and Results of
Operations”

Item 7A - “Quantitative and Qualitative Disclosures about Market Risk”

Item 8 - “Financial Statements and Supplementary Data”

Item 9 - “Changes in and Disagreements with Accountants on Accounting and Financial
Disclosure”

Item 9A - “Controls and Procedures”
At first glance the first point of notice is that there are no sections, parts, or items dedicated solely to the discussion of applicable legal regimes, regulations, or other regulatory frameworks. Instead, topics regarding applicable legal regimes and regulatory implications arise as organic components of sections dedicated to general business discussions, risk factors, and pending or ongoing litigation. In short, if the Commission’s suggested sectional components and requested descriptions are taken as indicative of components meaningful to investor comparison, applicable legal regimes, regulators, and frameworks may be viewed as unimportant to overall corporate performance or the appropriateness or future viability of any investment in a reporting public company. More plainly, where discussion of legal concepts, regulators, or applicable legal regimes arises, it is not meant to be used as a comparative measure made important to investor purchase decisions, but instead an outgrowth of firm discussion of general business summary, overarching regulatory risk, or an instance of specific litigation. Hence, discussion of legal topics is not a topical discussion purposefully elicited for the purpose of making a firm-to-firm comparison and instead arises as a descriptive component included by choice without any specific uniform structural request from the Commission or any specific applicable guidelines for the inclusion of such legal regime components. Without applicable structures oriented at providing required uniformity in disclosure structure and components for specific sectional discussions of legal and regulatory issues, employing the language used by firms to discuss those topics presents structural issues this simple project seeks to highlight.

Initial Concluding Remarks

To conclude these introductory and literary background sections before moving on into the simple sample and data comparison included in this project, the summary of Commission specific informational component requests and structural 10-K components was included to highlight the lack of legal or regulatory discussion elements formally required by the Commission and hence the structural lack of comparative guidelines through which meaningful uniform comparison of firm speech on legal or regulatory topics can be conducted. Where differences in the language
used to discuss required disclosure components, like financial performance, executive compensation, or other major “Items” may lead to meaningful outcome or left-hand side outcomes, without specific calls to submit legal or regulatory discussion or meaningful bounds through which such discourse should be submitted creating meaningful marks of comparison, comparing firm outcomes based on such legal and regulatory speech should be undertaken carefully.

When initially undertaken, this project sought to draw comparative delineations between firms based on their discussion of regulators and applicable legal frameworks. However, in the initial naive sample that follows, demarcation or meaningful differences between firm speech on regulators and applicable legal frameworks is sparse and a comparison of three industry groups is used to highlight this interesting outcome. In hopes of motivated further research on how disclosure burdened firms discuss regulators, this project concludes with suggested reasons why firm regulatory speech in the general business summary, risk, and legal proceedings sections may be either too sparse or too similarly bland to provide meaningful comparison for left-hand side outcomes.

_The chances of finding out what's really going on in the universe are so remote, the only thing to do is hang the sense of it and keep yourself occupied._

_Douglas Adams_
SECTION II

Data Summary

This project covers 10-K forms filed between the years of 1997 and 2010, including 14 separate years of filings made to the Commission’s EDGAR repository totally more than 230,000 total 10-K’s. Each file was cleaned prior to the implementation of the NLP methodology detailed below using the beautiful soup function for files not already in raw text form, those necessitating a conversion from raw html formats. Each year saw roughly between 12,000 and 17,500 filings, with the length of filings, measured by number of total paragraphs roughly increasing through the sample and the total number of filings submitted each year slowly decreasing from 1997 through 2010. A summary of all filer data and a three-industry group comparison were undertaken to compare the word counts found and overall filer and file characteristics found within the sample. Each filing set was arranged by file year.

NLP Program

The Natural Language Processing (NLP) methodology used in this project employs a simple word counting mechanism and word identification through two small operating word lists. Not unlike Loughran and McDonald’s works on textual sentiment using positive and negative word lists and accumulating word counts for the various tagged words found in text, the program created for this project located regulatory or law oriented words form a small word list of general legal terms and once, located, then stored the paragraph of text in which the word was located to a running string vector and then, at the end of the text document searched that string of compiled regulatory paragraphs for “blame” oriented words indicating responsibility, burden, or “due to” effects again, counting the number of “blame” oriented words found within those regulation or law specific paragraphs. The counting mechanisms and coding was done through Python and using the Natural Language Tool Kit (NLTK) plug ins and commands. The word counting mechanism and paragraph counting designs in this project are very simple and the author hopes that these techniques and the paragraph-by-paragraph orientation at 10-K topical speech can be used in other projects by future researchers or those who, like the author, are still developing their programming skills.
The program has been included at rear in PDF form in the appendices for reference, but in brief summary of the programmatic components, a process has been included below as well as the two applicable word lists used in word counting. The program begins by initializing and downloading the necessary packages, script sets, and operating functions to compile the word lists, execute word counting, and create the data output file once the word counts and paragraph counts, and other characterizing data points have been extricated from the files. In the first step, the program creates two vectors of words for the word counting mechanism to use when parsing, line by line, through the text file by importing the “Regulatory Words” and the “Blame Words” from their respective csv files using the methodology and structures used by Loughran and McDonald in their word counting projects. Each list of words, provide below, is turned into a vector, through which the word counting mechanism will loop in checking each line and each word element of each line as the program reads through the document in step two. Each word is a raw csv file listing the words, by structure they are listed in all caps, but since this program operates without specifically calling to capitalized words, stemming, or removing stop words, the words were left un-stemmed, and each string of paragrammatical pulls taken once a regulatory word was located was taking in full, with only punctuation being removed.

<table>
<thead>
<tr>
<th>Regulatory Words:</th>
<th>ACT</th>
<th>Blame Words:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUTHORITIES</td>
<td>BAN</td>
</tr>
<tr>
<td></td>
<td>AUTHORITY</td>
<td>CAUSE</td>
</tr>
<tr>
<td></td>
<td>GOVERN</td>
<td>CONTROL</td>
</tr>
<tr>
<td></td>
<td>LAW</td>
<td>DUE TO</td>
</tr>
<tr>
<td></td>
<td>LEGAL</td>
<td>EFFECT</td>
</tr>
<tr>
<td></td>
<td>LEGISLAT</td>
<td>IMPACT</td>
</tr>
<tr>
<td></td>
<td>ORDINANCE</td>
<td>IMPOSE</td>
</tr>
<tr>
<td></td>
<td>POLICIES</td>
<td>LIMIT</td>
</tr>
<tr>
<td></td>
<td>POLICY</td>
<td>OBLIGAT</td>
</tr>
<tr>
<td></td>
<td>REGIME</td>
<td>PROHIBIT</td>
</tr>
<tr>
<td></td>
<td>REGULATION</td>
<td>REGULATE</td>
</tr>
<tr>
<td></td>
<td>REGULATOR</td>
<td>REQUIRE</td>
</tr>
<tr>
<td></td>
<td>RULE</td>
<td>RESTRICT</td>
</tr>
<tr>
<td></td>
<td>STANDARDS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TREATIES</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TREATY</td>
<td></td>
</tr>
</tbody>
</table>

Once the vectors have been imported into vector form, the program is then sent year-by-year into the grouped filings, opening each document one-by-one and reading each text file line-by-line. Initial lines
containing filer-firm identifying information like unique firm identifier (CIK, Central Index Key number), firm name in text form, filing time and date, and industry identifier (SIC, Standard Industrial Classification number) are extricated and retained for identifying inclusion in the later data line which is appended at the end of the file reading process. The program counts each paragraph using the new line characters denoting empty space in the text file, and when a regulatory word is found within a paragraph, the lines of text making up that paragraph are saved to a repository vector in string form, once the reading functions comes to the end of the text document for that 10-K file, the program begins with step three. In step three, the string vector of regulatory paragraphs, paragraphs which were flagged as containing regulatory words, once punctuation is removed, is then counted for “Blame Words” in the same looping word counting mechanism that located the regulatory words.

Finally, in step four, the key demographic and identifying information recorded within the program for that particular filings are converted into a single line of comma separated text, and appended to the running vector of output data lines for that filing year, and once the entire year’s filings have been read, the comma separated line are converted into a csv file of output data. A sample output data file is supplied at rear in the appendices. This simple procedure returns equally simple, but enlightening data as to the general inclusion of regulation and legal regime-oriented speech which has been separated into the data presentations below for both a summary of data across all filers and a sub-selection of industry participants based on three separated industry identified subgroups.

*I refuse to answer that question on the grounds that I don't know the answer.*

Douglas Adams
## SECTION III

### All Filers Data and Initial Summary Remarks

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Filings</th>
<th>Average Paragraphs</th>
<th>Average Reg. Paragraphs</th>
<th>Average Number Reg Words Flag'd</th>
<th>Average Reg Paragraphs Total Words</th>
<th>Average “Blame” words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>17,147</td>
<td>978.590016</td>
<td>12.2764332</td>
<td>14.04047355</td>
<td>3311.627865</td>
<td>13849.007</td>
</tr>
<tr>
<td>1998</td>
<td>17,283</td>
<td>1002.14951</td>
<td>13.0299716</td>
<td>14.9572412</td>
<td>3645.08604</td>
<td>15528.147</td>
</tr>
<tr>
<td>1999</td>
<td>17,273</td>
<td>989.625601</td>
<td>12.4566665</td>
<td>14.377294</td>
<td>3414.08481</td>
<td>12449.9</td>
</tr>
<tr>
<td>2000</td>
<td>18,135</td>
<td>1006.57541</td>
<td>12.5793769</td>
<td>14.4310449</td>
<td>3524.18875</td>
<td>12905.304</td>
</tr>
<tr>
<td>2001</td>
<td>18,317</td>
<td>1192.12191</td>
<td>12.2279303</td>
<td>14.1599607</td>
<td>3953.7813</td>
<td>17043.384</td>
</tr>
<tr>
<td>2002</td>
<td>17,825</td>
<td>2164.47961</td>
<td>11.8808976</td>
<td>13.8873492</td>
<td>4217.01532</td>
<td>16762.141</td>
</tr>
<tr>
<td>2003</td>
<td>16,981</td>
<td>3276.2395</td>
<td>12.546022</td>
<td>14.5142807</td>
<td>5549.9377</td>
<td>18912.36</td>
</tr>
<tr>
<td>2004</td>
<td>16,914</td>
<td>4046.30708</td>
<td>12.4619842</td>
<td>14.4949155</td>
<td>5728.47611</td>
<td>20745.635</td>
</tr>
<tr>
<td>2005</td>
<td>18,195</td>
<td>4844.98873</td>
<td>11.1621325</td>
<td>12.799945</td>
<td>4876.16697</td>
<td>15077.44</td>
</tr>
<tr>
<td>2006</td>
<td>17,154</td>
<td>5680.97196</td>
<td>10.9470677</td>
<td>12.3952431</td>
<td>5243.86365</td>
<td>13773.917</td>
</tr>
<tr>
<td>2007</td>
<td>16,504</td>
<td>7220.05284</td>
<td>10.4194135</td>
<td>11.7424261</td>
<td>5225.68317</td>
<td>15129.344</td>
</tr>
<tr>
<td>2008</td>
<td>16,543</td>
<td>7116.11243</td>
<td>10.3126398</td>
<td>11.4379496</td>
<td>4319.48709</td>
<td>14746.89</td>
</tr>
<tr>
<td>2009</td>
<td>14,226</td>
<td>8018.58864</td>
<td>10.3195557</td>
<td>11.3812737</td>
<td>4993.82602</td>
<td>12271.789</td>
</tr>
<tr>
<td>2010</td>
<td>12,908</td>
<td>8518.42268</td>
<td>10.5868454</td>
<td>11.8703905</td>
<td>12627.2261</td>
<td>19017.428</td>
</tr>
</tbody>
</table>

While the total number of reports filed follows a roughly decreasing pattern from 1997 – 2010, the average length as measured by the total number of paragraphs included in each filing steadily grows throughout the sample. Figures have been provided on the following page. While data was initially taken from 1997 – 2012, errors in the programmatic application of the NLP methodology led to a distinct change in data grabbing following 2010; this could be the result of applicable reporting guideline and regulatory changes relating to public filers updated by the Commission in 2010 changing the applicability of pertinent word usage and the appropriateness of the word grabbing mechanism used, however, analysis of these pattern responses is outside the purview of this small project and therefore was simplified and removed until a sample of post 2010 filings can be undertaken using a different grabbing technique. However, the general summary progression of regulatory and blame word counts supplied in the diagrams that follow provides an interesting and discourse motivating snapshot of the general filer regulatory discourse included in these filings for summary purposes.
Time is an illusion. Lunchtime doubly so.
Douglas Adams
SECTION IV

Industry Comparison

While the summary figures previously shown show an interesting overtime development over the average and aggregate filing characteristics for U.S. reporting public companies, the focus of this paper is instead on a three-industry comparison meant to illustrate the difficulties in using regulatory or legal regime oriented speech in the context of 10-K filings. The Commission classifies reporting companies into industry groups or segments using the SIC or Standard Industrial Classification system.

The Standard Industrial Classification Codes that appear in a company's disseminated EDGAR filings indicate the company's type of business. These codes are also used in the Division of Corporation Finance as a basis for assigning review responsibility for the company's filings. For example, a company whose business was Metal Mining (SIC 1000) would have its filings reviewed by staffers in A/D Office 9. SEC definitional description provided on the Commission’s webpage Division of Corporation Finance: Standard Industrial Classification (SIC) Code List.

Using the Commission's SIC classifications, three groups of report filers were selected for comparison, Petroleum and Natural Gas companies, Wholesale and Retail companies, and Financial and Investment Institutions, SIC codes: 1311 – 1389, 5000 – 5990, and 6021 – 6282 respectively. These selections were made to highlight the differences in the patterns of legal and regulatory speech and blame words found given the stark differences in the regulatory regimes, frameworks, and compliance standards applicable to firms in each of the three SIC groupings. Petroleum and Natural Gas companies must practice constant compliance with and adherence to strict and detailed state, federal, and international guidelines for products, implementation, and practices in an impact or specific control framework of regulators and regulations. Wholesale and Retail providers, on the other hand, face little structural control over operations and instead face the majority of their regulation in consumer suit, tort and tort reform, and other state and federal guidelines over market pricing and consumer or employee concerns. And the final category, Financial and Investment Institutions, faces a detailed system of investor sand consumer protection and market protection regimes detailing strict compliance, disclosure, and reporting guidelines, predominantly focused on entities like the FDIC, the SEC, Finra, and other report and disclosure oriented regulatory
regimes. The regulatory structural components and adherence frameworks present in each of these regimes was initially hypothesized as providing ample motivation for firms to produce varied usage and discussion of legal terms, regulatory speech, and comment on the implications or “blame” words oriented at pertinent regulators. What seems instead to result, is a more nuanced set of regulatory response difference, where firm group trends, while differencing slightly, do not seem to reveal marked differences in usage and blame.**

** A full table of data used in the three SIC group comparison is provided at rear in Appendix 4. 1300 Group: Petroleum and Natural Gas, 5000 Group: Wholesale and Retail, 6000 Group: Financial and Investment Institutions
While each sector shows a marked difference in the number of filers, the relative behavior of each of the other groups shows a more nuanced, and to some degree, unexpected differences in the number of regulatory words flagged, and similarities in the length of regulatory strings reviewed and in the number of blame words used is, again, surprising, given the differences in regulatory frameworks applicable. While increased reporting guidelines passed in 2008 may be cause for the explosion in the length of Financial and
Investment Institution filings following 2008, the constant deficit in general regulatory speech, indicated in regulatory words flagged and in regulatory paragraphs overall, is relatively surprising. In short, this simple comparison demonstrates the need for a nuanced approach to using legal or regulation oriented speech as a meaningful left-hand side indicator in the context of 10-K filings. While words used pertaining to regulatory concepts and legal issues may lead to within industry interest, for example, as a motivating component for greater research into the difference in form length between industries following the financial regulations instituted as effective in 2008, using regulatory speech as an overall indicator in tying firm outcomes may not provide meaningful reflection on actual applicable legal regimes. Hence, this project seeks to provide this brief industry comparison to highlight the difficulty in comparing these legal and regulatory mentions where this language and these general regulator words are not, by the structure of the information requested, guided, or compelled for inclusion in these reports.

There is a theory which states that if ever anyone discovers exactly what the Universe is for and why it is here, it will instantly disappear and be replaced by something even more bizarre and inexplicable. There is another theory which states that this has already happened.

*Douglas Adams, Hitchhiker's Guide to the Galaxy*
SECTION V

Concluding Remarks

While other projects have tied left-hand side outcomes to language contained in 10-K reports, using the legal and regulatory oriented speech in these reports in the same way may provide difficult or meaningless without implementing a nuanced, sector-by-sector, industry-by-industry, or some more specific approach. Without a specific call or uniform guidelines for the inclusion of such regulatory information or discussion for means of comparison, like other sections included in these filings, using regulatory speech for such purposes in supposing ties to left-hand side outcomes may be unwarranted, difficult, or may fail to provide statistical significance. While this outcome has been disheartening to the author, some minor further discussion below as to why this particular issue may be the case below may help to further other research paths oriented at dissecting this type of report information.

In discussion with several lawyers who file Form 10-K on behalf of their clients, several possible reasons for these nuance issues arose, if, unlike tying such legal discussion to left-hand side outcomes, the study of these particular topics is undertaken in the context of these filings, perhaps those topical focuses, could yield a more reasonable, but less NLP focused output data of interest. For many of these filings, these legal references surfaced in boilerplate language detailing possible risks, especially those in the case of foreign or other applicable legal regimes. For many filing attorneys and firms, the use of stock boilerplate or copy boilerplate language is common, if not the norm, and should a small firm lawyer be filing the same, finding appropriate language by using boiler plate already included in a prior filing by a third party is not uncommon, it is quite the case that copying is flattery, and best practice to some regard. The same seems to hold true for the operating inclusion of legal disclaimers. Recycling disclaimer and boilerplate language both within firm and filer to filer is common, lawyers, not unlike the author, feel there is hardly any need to reinvent the wheel, and as the old adage goes, it is the non-conforming nail that receives the hammer after all. Finally, industry standard for disclosure, risk boilerplate, and the discussion of risks, regimes, and legal doctrines may also be at fault for these nuances. Smaller public companies may follow their well-
received larger industry bedfellows in conforming to reporting regularity when it comes to boilerplate language, disclaimers, and risk summaries, for the simple reason that, where all companies report identically in such matters, an operating industry standard seemingly results. In short, such behavioral keys in the law firms and accounting departments issuing these reports, while fruit for specific projects on these concepts, muddies the water in using the legal and regulatory comments included in these reports as meaningful identifiers for left-hand side outcomes.

*I love deadlines. I love the whoooshing noise they make as they go by.*

*Douglas Adams*
BIBLIOGRAPHY

Bibliographical information included below has been grouped by topic, research field, and provider.

Useful Securities and Exchange Commission Provided Links, Materials, and Guides:

ECFR for 1934 Act: https://www.ecfr.gov/cgi-bin/text-idx?amp;node=17:4.0.1.1.1&rgn=div5

Landing Page for Sec Corporate Finance Division: https://www.sec.gov/page/corpfin-section-landing


Quick Information listing 10-K by SEC: https://www.sec.gov/fast-answers/answers-form10khtm.html

SEC Form 10-K Sample PDF with Sample form and Instructions for Filers: https://www.sec.gov/about/forms/form10-k.pdf


SEC Forms List: https://www.sec.gov/forms

Resources and Citations by Tom Loughran and Bill McDonald:

Link to Notre Dame repository maintained by Bill McDonald: https://sraf.nd.edu


**Helpful Natural Language Processing and Application Summaries and Surveys**


**Finance, Accounting, and Organizational Management:**


Legal:


A common mistake that people make when trying to design something completely foolproof is to underestimate the ingenuity of complete fools.

Douglas Adams
APPENDICES

Due to the length and form of appended supplements, each has been attached with cover page and description following the final page of this document.

Appendix 1: NLP Code

PDF Printout of the Natural Language Processing Python Code used to pull the words as data for this project. Includes: initial package calls and downloads, paragraph by paragraph counting mechanism, regulatory word list counting and vector input code, “Blame” word counting mechanism, and Out File CSV closure program.

Appendix 2: Sample Out File

Columnized CSV file output, shows column by column sample data produced from a single year reporting pull from the NLP Code.

Appendix 3: SEC SIC Firm Groups List

While a full list of SEC SIC Codes can be found on the “Division of Corporation Finance: Standard Industrial Classification (SIC) Code List” at: www.sec.gov/info/edgar/siccodes.htm, this appendix includes the subset of SIC codes used in the 3-industry comparison employed in this project.

Appendix 4: 3 Group SIC Firm Group Comparison Full Data Table

Full table of the data compiled in this project for the 3-industry comparison employed in this project. Also includes a comparison of the “highest overall” from the entire sample, with firm name and relevant SIC, at right for reference.

He gazed keenly into the distance and looked as if he would quite like the wind to blow his hair back dramatically at that point, but the wind was busy fooling around with some leaves a little way off.

Douglas Adams
I may not have gone where I intended to go, but I think I have ended up where I needed to be.

*Douglas Adams*
Isn’t it enough to see that a garden is beautiful without having to believe that there are fairies at the bottom of it too?

Douglas Adam
# -*- coding: utf-8 -*-

Spyder Editor

This is a temporary script file.

#In Core Packages

#Overall Progs
import numpy
import os.path
import urllib.request
from bs4 import BeautifulSoup, SoupStrainer
import re
import zipfile
import sys
import webbrowser
import urllib.request
import codecs
import csv
import sys
import collections
import glob
import string
from string import punctuation

#Word List Progs
import os
import shutil
import nltk
from string import punctuation
from nltk.stem.snowball import SnowballStemmer
from nltk.corpus import stopwords
from nltk.tokenize import RegexpTokenizer
from collections import OrderedDict
from sklearn.feature_extraction.text import TfidfVectorizer
from collections import defaultdict
import re

#Set two word functions for help later with NLTK
tokenizer= RegexpTokenizer(r'\w+') #Splits up strings into little strings
NOPunct = set(string.punctuation) #set of punctuation symbols for later removal

#Step 1: Load Word Vectors from word Lists

#Open Empty Vectors for Word Lists (note: NOT STEMMING since reg titles full)
RegWords= [] #Storage for General Regulation words list
DueToWords = [] #Storage for Blame words list

#Path to Word List.csv files
Wordpath = "/Users/KMills/Desktop/Hart Fellowship FALL/"
# Word File Names
RegCSV = "LawWords_Hart2018.csv"
DueToCSV = "DueToWords_Hart2018.csv"

# Setting General Reg Words Set into RegWords vector
with open(Wordpath+RegCSV, "r") as reg:
    line=reg.readline().rstrip('\n')
    RegWords.append(line)  # Adding each word to our word list for positive words
    while(line):
        if line !=RegWords[-1]:  # Checking for repeats before we place things into the list
            RegWords.append(line)  # adds the new words only when it is not a repeat
            line=reg.readline().rstrip('\n')

    # Testing the the List of regulatory words
    print(line)
    print(len(RegWords))  # Notes the number of pos. words (unstemmed)
    reg.close()

# Setting Blame (DueTo) Words Set into DueToWords vector
with open(Wordpath+DueToCSV, "r") as dt:
    line=dt.readline().rstrip('\n')
    DueToWords.append(line)  # Adding each word to our word list for positive words
    while(line):
        if line !=DueToWords[-1]:  # Checking for repeats before we place things into the list
            DueToWords.append(line)  # adds the new words only when it is not a repeat
            line=dt.readline().rstrip('\n')

    # Testing the the List of regulatory words
    print(line)
    print(len(DueToWords))  # Notes the number of pos. words (unstemmed)
    reg.close()

print("Word lists converted into Vector Form")

# Step 2: Send the program to go return data
# Setting path to 10-K Files
PracticePath = "/Users/KMills/Desktop/Sample_Processed"

# Setting matrix to house each data line (document by document)
StatLines = []

# Sending the program into the files in loop
for file in os.listdir(PracticePath):        # Selects the file from the directory file
if file.endswith(".txt"):
    # Opens each one of the txt files in the directory
    print("CURRENT FILE: "+os.path.join(PracticePath,file))  # Prints the path to the current FILE
    doc = open(os.path.join(PracticePath,file))  # Opens the individual 10K file itself
    print("File Opened: "+os.path.join(PracticePath,file))  # Initial reg File Opened Sanity Check

    ###################################################################
    #### Setting the initial empty vector for word string for file ####
    ###################################################################

    DocCorpus = []  # setting clean repository for reg-word-wholetext
    print("-- [] -- Empty Corpus Opened")
    MiniCorpus = []  # opening initial mini-corpus
    # will have writing procedures over mini-corpus later

    ###################################################################
    ### Words_Flagd: Reg or Act words counted                     ###
    ### Blame_Words: Blame words counted in reg/act paragraphs     ###
    ### Para_Count: Total Paragraphs Counter                      ###
    ### RegPara_Count: Total Paragraphs with Reg/Act Word         ###
    ###################################################################

    # Setting Word Counters to Zero
    Words_Flagd = 0
    Blame_Words = 0
    Para_Count = 0
    RegPara_Count = 0

    # Setting word list (General Reg or Act Words) to be used
    List_Name = RegWords  # use same code for list of gen/specific

    ###################################################################
    #### Sending the program in to read lines in the cleaned form  ####
    ###################################################################

    lines = doc.readlines()  # sets lines as reading line-by-line
    print("Quiet! I'm Trying to Read these lines!!")  # Read lines sanity check

    for line in lines:

        ###################################################################
        ##### Grabbing Vitals before beginning the language processing ####
        ###################################################################

        # CIK Number coded to: CIK_Number
        # Report Date coded to: Rep_Date
```python
if "ACCESSION NUMBER:" in line and re.findall(r"\D\d{10}\D", line):
    CIK_Number = str(re.findall(r"\D\d{10}\D", line))
    print("CIK Located: "+CIK_Number)

if "FILED AS OF DATE:" in line and re.findall(r"\D\d{8}\D", line):
    Rep_Date = str(re.findall(r"\D\d{8}\D", line))
    print("Report Date Located: "+Rep_Date)

if "COMPANY CONFORMED NAME:" in line:
    try:
        Name_Line = line.split("\t\t")
        # print(str(Name_Line)) # Line structure
        # print(str(Name_Line[1])) # Name structure
        Dirty_Name = Name_Line[1] # Un-Chomped Name
        Co_Name = str(Dirty_Name).rstrip() # Chomps new line char
        print("Company Name Found: "+Co_Name) # Update Check
    except:
        Co_Name = "Null_Split_Co_Name"
        print(Co_Name)

if "STANDARD INDUSTRIAL CLASSIFICATION:" in line and re.findall(r"\D\d{4}\D", line):
    SIC_Number = str(re.findall(r"\D\d{4}\D", line))
    print("SIC Found: "+SIC_Number)
```
### If we hit a new paragraph, we want to: ###
### 1. Test the Mini-Corpus for Reg Words ###
### 2. If we find a reg word we append to Doc Corpus ###
### 3. If we find NO Reg Word, we move on to new para ###
### -- So each Mini-Corpus is going to be appended ###
### to the main corpus ONLY IF it contains a reg word ###
### -- Counts the number of paragraphs total (ie total ###
### tested paragraphs) and then counts the number of the ###
### paragraphs we have tested which also contain reg/act ###
### words - this gives us a good look at volume of reg ###
### -- speech relative to document length for each txt file ###

```
print("Blank Space Found -- TESTING PARAGRAPH")
Para_Count = Para_Count + 1
Dirty_Words = str(MiniCorpus)
words = ''.join(c for c in Dirty_Words if c not in NOPunct)
for word in List_Name:
    WC = words.count(word) #Counts the offurance of the word
    #print(str(word)+": "+str(WC)) #tracking as we pull words
    Words_Flagd = Words_Flagd + WC #Updates overall counter
    if WC > 0:
        DocCorpus.append(MiniCorpus)
        #DocCorpusTracker = DocCorpusTracker + 1
        #print("Document Corpus Updated: "+str(DocCorpus))
        RegPara_Count = RegPara_Count + 1
MiniCorpus = []
print("MC Reset for New")
```

### Now we want to text the overall corpus of regulatory ###
### paragraphs for the bame words. So now we want to take our # ### DocCorpus and do a word count inside that for the words # ### that we have labeled as blame words :-) #

```
Dirty_Corpus = str(DocCorpus)
corpus = ''.join(c for c in Dirty_Corpus if c not in NOPunct )
```
Str_Leng = len(corpus)

for blameword in DueToWords:
    BC = corpus.count(blameword)
    #print(str(blameword)+" : "+str(BC)) #tracking as we pull words
    Blame_Words = Blame_Words + BC #Updates overall counter

#Cleaning a few of my strings before we append to csv file
Report_Date = str(Rep_Date)
Report_Date = ''.join(c for c in Report_Date if c not in NOPunct )
Company_Name = str(Co_Name)
Company_Name = ''.join(c for c in Company_Name if c not in NOPunct )
CIKNumber = str(CIK_Number)
CIKNumber = ''.join(c for c in CIKNumber if c not in NOPunct )
SICNumber = str(SIC_Number)
SICNumber = ''.join(c for c in SICNumber if c not in NOPunct )

#Creating call line Fact list:
ID_Line = str(Report_Date)+","+str(Company_Name)+","+str(CIKNumber) +","+str(SICNumber)
Data_Line = str(Para_Count)+","+str(RegPara_Count)+","+str(Words_Flagd) +","+str(Str_Leng)+","+str(Blame_Words)
Stats_Line = str(ID_Line+"","+Data_Line)

print(Stats_Line)

#Appending the stat-line to the stat line Matrix for overall
StatLines.append(Stats_Line)

print("Stats Line Appended total files: "+str(len(Stats_Line)))

 tidydoc.close
print("File: "+os.path.join(PracticePath,file)+" CLOSED")
#Creating the output CSV file
LocalOut = "/Users/KMills/Desktop/SAMPLEOUT/" #Local Output file folder
Stats_CSV = "10K_SampleOut_Hart2018.csv"

GoTo = str(str(LocalOut)+str(Stats_CSV)) #Directions for outfile

with open(GoTo, "w") as OutCSV:
    wr = csv.writer(OutCSV, lineterminator='\n')
    for stat in StatLines:
        wr.writerow([stat])

OutCSV.close()
Appendix 2 - Sample Out-File CSV Table

<table>
<thead>
<tr>
<th>Report Date</th>
<th>Firm Name</th>
<th>CIK Number</th>
<th>SIC Number</th>
<th>Total Paragraphs</th>
<th>Regulatory Paragraphs</th>
<th># Reg Words</th>
<th>Reg String Length</th>
<th># Blame Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>20080331</td>
<td>ADAMS RESOURCES ENERGY INC</td>
<td>2178</td>
<td>5172</td>
<td>3237</td>
<td>10</td>
<td>10</td>
<td>810</td>
<td>1417</td>
</tr>
<tr>
<td>20080228</td>
<td>ALEXANDER BALDWIN INC</td>
<td>3453</td>
<td>4400</td>
<td>47004</td>
<td>13</td>
<td>15</td>
<td>1589</td>
<td>44021</td>
</tr>
<tr>
<td>20080225</td>
<td>ALEXANDERS INC</td>
<td>3499</td>
<td>6798</td>
<td>6563</td>
<td>20</td>
<td>21</td>
<td>5523</td>
<td>9039</td>
</tr>
<tr>
<td>20080604</td>
<td>AMERCO NV</td>
<td>4457</td>
<td>7510</td>
<td>4614</td>
<td>3</td>
<td>3</td>
<td>274</td>
<td>0</td>
</tr>
<tr>
<td>20080220</td>
<td>AMR CORP</td>
<td>4515</td>
<td>4512</td>
<td>8718</td>
<td>4</td>
<td>4</td>
<td>243</td>
<td>0</td>
</tr>
<tr>
<td>20080220</td>
<td>AMERICAN AIRLINES INC</td>
<td>4515</td>
<td>4512</td>
<td>7475</td>
<td>4</td>
<td>4</td>
<td>243</td>
<td>0</td>
</tr>
<tr>
<td>20080222</td>
<td>AMERICAN ELECTRIC POWER CO INC</td>
<td>4904</td>
<td>4911</td>
<td>19824</td>
<td>38</td>
<td>39</td>
<td>5126</td>
<td>28361</td>
</tr>
<tr>
<td>20080228</td>
<td>OHIO POWER CO</td>
<td>4904</td>
<td>4911</td>
<td>19732</td>
<td>35</td>
<td>36</td>
<td>5089</td>
<td>28588</td>
</tr>
<tr>
<td>20080228</td>
<td>APPALACHIAN POWER CO</td>
<td>4904</td>
<td>4911</td>
<td>19922</td>
<td>38</td>
<td>39</td>
<td>5130</td>
<td>28509</td>
</tr>
<tr>
<td>20080228</td>
<td>PUBLIC SERVICE CO OF OKLAHOMA</td>
<td>4904</td>
<td>4911</td>
<td>19139</td>
<td>38</td>
<td>39</td>
<td>5220</td>
<td>28555</td>
</tr>
<tr>
<td>20080228</td>
<td>COLUMBUS SOUTHERN POWER CO OI</td>
<td>4904</td>
<td>4911</td>
<td>19437</td>
<td>40</td>
<td>43</td>
<td>7508</td>
<td>343887</td>
</tr>
<tr>
<td>20080228</td>
<td>SOUTHWESTERN ELECTRIC POWER CO</td>
<td>4904</td>
<td>4911</td>
<td>19109</td>
<td>37</td>
<td>38</td>
<td>5283</td>
<td>28553</td>
</tr>
<tr>
<td>20080228</td>
<td>INDIANA MICHIGAN POWER CO</td>
<td>4904</td>
<td>4911</td>
<td>19080</td>
<td>37</td>
<td>38</td>
<td>5221</td>
<td>28227</td>
</tr>
<tr>
<td>20080328</td>
<td>AEP Texas Central Transition Funding I</td>
<td>4904</td>
<td>6189</td>
<td>566</td>
<td>5</td>
<td>5</td>
<td>460</td>
<td>0</td>
</tr>
<tr>
<td>20080714</td>
<td>AMREP CORP</td>
<td>6207</td>
<td>6500</td>
<td>1146</td>
<td>11</td>
<td>11</td>
<td>8709</td>
<td>2549</td>
</tr>
<tr>
<td>20080305</td>
<td>ANALYSTS INTERNATIONAL CORP</td>
<td>6292</td>
<td>7371</td>
<td>3108</td>
<td>5</td>
<td>5</td>
<td>455</td>
<td>835</td>
</tr>
<tr>
<td>20080401</td>
<td>STAGE STORES INC</td>
<td>6885</td>
<td>5651</td>
<td>4041</td>
<td>17</td>
<td>17</td>
<td>2066</td>
<td>0</td>
</tr>
<tr>
<td>20080314</td>
<td>ARABIAN AMERICAN DEVELOPMENT C</td>
<td>7039</td>
<td>2911</td>
<td>3390</td>
<td>12</td>
<td>12</td>
<td>1126</td>
<td>209</td>
</tr>
<tr>
<td>20080415</td>
<td>ARABIAN AMERICAN DEVELOPMENT C</td>
<td>7039</td>
<td>2911</td>
<td>434</td>
<td>5</td>
<td>5</td>
<td>417</td>
<td>603</td>
</tr>
<tr>
<td>20080829</td>
<td>ARCHER DANIELS MIDLAND CO</td>
<td>7084</td>
<td>2070</td>
<td>4464</td>
<td>12</td>
<td>12</td>
<td>663</td>
<td>0</td>
</tr>
<tr>
<td>20080828</td>
<td>SOUTHWESTERN ENERGY CO</td>
<td>7332</td>
<td>1311</td>
<td>17816</td>
<td>12</td>
<td>16</td>
<td>9077</td>
<td>5274</td>
</tr>
<tr>
<td>20081126</td>
<td>ATWOOD OCEANICS INC</td>
<td>8411</td>
<td>1381</td>
<td>7116</td>
<td>29</td>
<td>30</td>
<td>35266</td>
<td>75362</td>
</tr>
<tr>
<td>20080512</td>
<td>AZZ INC</td>
<td>8947</td>
<td>3640</td>
<td>3179</td>
<td>6</td>
<td>6</td>
<td>590</td>
<td>179</td>
</tr>
<tr>
<td>20080404</td>
<td>BAGDAD CHASE INC</td>
<td>9128</td>
<td>1000</td>
<td>240</td>
<td>4</td>
<td>4</td>
<td>244</td>
<td>0</td>
</tr>
<tr>
<td>20081118</td>
<td>BAGDAD CHASE INC</td>
<td>9128</td>
<td>1000</td>
<td>243</td>
<td>4</td>
<td>4</td>
<td>245</td>
<td>0</td>
</tr>
<tr>
<td>20080314</td>
<td>BALDWIN LYONS INC</td>
<td>9346</td>
<td>6331</td>
<td>35697</td>
<td>8</td>
<td>8</td>
<td>381</td>
<td>931</td>
</tr>
</tbody>
</table>

It is a mistake to think you can solve any major problems just with potatoes.
Douglas Adams
Appendix 3 - SEC SIC Codes List Firm Groups Selected for Comparison

SIC Subsets Chosen:
1311 - 1389: Petroleum and Natural Gas
5000 - 5990: Wholesale and Retail
6021 - 6282: Financial and Investment Institutions

1311-1389 Petroleum and Natural Gas
[All Overseen by A/D Office 4]
1311: CRUDE PETROLEUM AND NATURAL GAS
1381: DRILLING OIL AND GAS WELLS
1382: OIL AND GAS FIELD EXPLORATION
1389: OIL AND GAS FIELD SERVICES, NEC

5000-5990: Wholesale and Retail
[All Overseen by A/D Offices: 2, 3, 4, 5, 6, 9, 10 ]
5000: WHOLESALE-DURABLE GOODS
5010: WHOLESALE-MOTOR VEHICLES & MOTOR VEHICLE PARTS & SUPPLIES
5013: WHOLESALE-MOTOR VEHICLE SUPPLIES & NEW PARTS
5020: WHOLESALE-FURNITURE & HOME FURNISHINGS
5030: WHOLESALE-LUMBER & OTHER CONSTRUCTION MATERIALS
5031: WHOLESALE-LUMBER, PLYWOOD, MILLWORK & WOOD PANELS
5040: WHOLESALE-PROFESSIONAL & COMMERCIAL EQUIPMENT & SUPPLIES
5045: WHOLESALE-COMPUTERS & PERIPHERAL EQUIPMENT & SOFTWARE
5047: WHOLESALE-MEDICAL, DENTAL & HOSPITAL EQUIPMENT & SUPPLIES
5050: WHOLESALE-METALS & MINERALS (NO PETROLEUM)
5051: WHOLESALE-METALS SERVICE CENTERS & OFFICES
5063: WHOLESALE-ELECTRICAL APPARATUS & EQUIPMENT, WIRING SUPPLIES
5064: WHOLESALE-ELECTRICAL APPLIANCES, TV & RADIO SETS
5065: WHOLESALE-ELECTRONIC PARTS & EQUIPMENT, NEC
5070: WHOLESALE-HARDWARE & PLUMBING & HEATING EQUIPMENT & SUPPLIES
5072: WHOLESALE-HARDWARE
5080: WHOLESALE-MACHINERY, EQUIPMENT & SUPPLIES
5082: WHOLESALE-CONSTRUCTION & MINING (NO PETRO) MACHINERY & EQUIP
5084: WHOLESALE-INDUSTRIAL MACHINERY & EQUIPMENT
5090: WHOLESALE-MISC DURABLE GOODS
5094: WHOLESALE-JEWELRY, WATCHES, PRECIOUS STONES & METALS
5099: WHOLESALE-DURABLE GOODS, NEC
5110: WHOLESALE-PAPER & PAPER PRODUCTS
5122: WHOLESALE-DRUGS, PROPRIETARIES & DRUGGISTS' SUNDRIES
5130: WHOLESALE-APPAREL, PIECE GOODS & NOTIONS
5140: WHOLESALE-GROCERIES & RELATED PRODUCTS
5141: WHOLESALE-GROCERIES, GENERAL LINE
5150: WHOLESALE-FARM PRODUCT RAW MATERIALS
5160: WHOLESALE-CHEMICALS & ALLIED PRODUCTS
5171: WHOLESALE-PETROLEUM BULK STATIONS & TERMINALS
5172: WHOLESALE-PETROLEUM & PETROLEUM PRODUCTS (NO BULK STATIONS)
5180: WHOLESALE-BEER, WINE & DISTILLED ALCOHOLIC BEVERAGES
5190: WHOLESALE-MISCELLANEOUS NONDURABLE GOODS
5200: RETAIL-BUILDING MATERIALS, HARDWARE, GARDEN SUPPLY
5211: RETAIL-LUMBER & OTHER BUILDING MATERIALS DEALERS
5271: RETAIL-MOBILE HOME DEALERS
5311: RETAIL-DEPARTMENT STORES
5331: RETAIL-VARIETY STORES
5399: RETAIL-MISC GENERAL MERCANTILE STORES
5400: RETAIL-FOOD STORES
5411: RETAIL-GROCERY STORES
5412: RETAIL-CONVENIENCE STORES
5500: RETAIL-AUTO DEALERS & GASOLINE STATIONS
5531: RETAIL-AUTO & HOME SUPPLY STORES
5600: RETAIL-APPAREL & ACCESSORY STORES
5621: RETAIL-WOMEN'S CLOTHING STORES
5651: RETAIL-FAMILY CLOTHING STORES
5661: RETAIL-SHOE STORES
5700: RETAIL-HOME FURNITURE, FURNISHINGS & EQUIPMENT STORES
5712: RETAIL-FURNITURE STORES
5731: RETAIL-RADIO, TV & CONSUMER ELECTRONICS STORES
5734: RETAIL-COMPUTER & COMPUTER SOFTWARE STORES
5735: RETAIL-RECORD & PRERECORDED TAPE STORES
5810: RETAIL-EATING & DRINKING PLACES
5812: RETAIL-EATING PLACES
5900: RETAIL-MISCELLANEOUS RETAIL
5912: RETAIL-DRUG STORES AND PROPRIETARY STORES
5940: RETAIL-MISCELLANEOUS SHOPPING GOODS STORES
5944: RETAIL-JEWELRY STORES
5945: RETAIL-HOBBY, TOY & GAME SHOPS
5960: RETAIL-NONSTORE RETAILERS
5961: RETAIL-CATALOG & MAIL-ORDER HOUSES
5990: RETAIL-RETAIL STORES, NEC

6021 - 6282: Financial and Investment Institutions

[All Overseen by A/D Offices: 7, 8, and OSF (Office of Structured Finance) ]

44
6021: NATIONAL COMMERCIAL BANKS
6022: STATE COMMERCIAL BANKS
6029: COMMERCIAL BANKS, NEC
6035: SAVINGS INSTITUTION, FEDERALLY CHARTERED
6036: SAVINGS INSTITUTIONS, NOT FEDERALLY CHARTERED
6099: FUNCTIONS RELATED TO DEPOSITORY BANKING, NEC
6111: FEDERAL & FEDERALLY-SPONSORED CREDIT AGENCIES
6141: PERSONAL CREDIT INSTITUTIONS
6153: SHORT-TERM BUSINESS CREDIT INSTITUTIONS
6159: MISCELLANEOUS BUSINESS CREDIT INSTITUTION
6162: MORTGAGE BANKERS & LOAN CORRESPONDENTS
6163: LOAN BROKERS
6172: FINANCE LESSORS
6189: ASSET-BACKED SECURITIES
6199: FINANCE SERVICES
6200: SECURITY & COMMODITY BROKERS, DEALERS, EXCHANGES & SERVICES
6211: SECURITY BROKERS, DEALERS & FLOTATION COMPANIES
6221: COMMODITY CONTRACTS BROKERS & DEALERS
6282: INVESTMENT ADVICE

Shee, you guys are so unhip it’s a wonder your bums don’t fall off.
Douglas Adams
Appendix 4 - 3 Group SIC Firm Group Comparison Full Data Table Cover Page

Full Data Table pdf follows.

Those who study the complex interplay of cause and effect in the history of the Universe say that this sort of thing is going on all the time, but that we are powerless to prevent it. 'It's just life,' they say.

Douglas Adams
<table>
<thead>
<tr>
<th>SIC: 1311 - 1389 Petroleum / Natural Gas</th>
<th>SIC: 5000 - 5990 Wholesale and Retail</th>
<th>SIC: 6021 - 6282 Financial and Investment Institutions</th>
<th>All Filings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year</strong></td>
<td><strong>Total Filings</strong></td>
<td><strong>Average Paragraphs</strong></td>
<td><strong>Average Paragraphs</strong></td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>------------------------------------------</td>
<td>------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>2012</td>
<td>578</td>
<td>574.850017</td>
<td>11.00800081</td>
</tr>
</tbody>
</table>