

# Text Analysis on Corporate 8-K Filings: Sustainability as a Trading Signal\*

**Zhengqing (Cody) Wan**

CODY.WAN@NYU.EDU

*MS. in Mathematics in Finance*

*Cournat Institute, New York University*

*New York, NY 10012, USA*

## Abstract

This paper tests the hypothesis that companies may file an 8K current report that discusses issues and actions related to sustainability and ESG, for perceived publicity and marketing values, which may negatively affect stock performance in the near future. In doing so, I introduce text analysis methods, from text preprocessing to topic modeling with Latent Dirichlet Allocation (LDA) and apply them to identify 8K filing documents on sustainability. The backtested result shows that the number of such 8K filings increases over the years, and the average returns following the release of 8K are statistically significantly positive in the early days and statistically significantly negative in more recent years. This paper then suggests that sustainability can be viewed as a signal for shorting stocks and also discusses other implications following the result.

**Keywords:** ESG, Sustainability, SEC Filings, Text Analysis, Latent Dirichlet Allocation, Event-Driven Strategy

## 1. Introduction

Text analysis, also commonly referred to as text mining or natural language processing (NLP) in some loose sense, has been around since the 1980s, when a set of algorithms had been proposed to find a subset of a word’s characters that can best capture its intrinsic meaning. One of the most prominent algorithms is the Porter stemmer by Porter et al. (1980), which is still commonly used today. As we enter the 21st century, more algorithms that seek to understand the meaning of a text document, represented as a list of stemmed words, have come into place. They range from probabilistic models such as the Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003) to transformer-based neural-network architecture like Google BERT from Devlin et al. (2018), making once hopeless NLP tasks amenable.

With the democratization of computing powers and the prevalence of open-source NLP algorithms, text analysis is making its way into finance. It does not necessarily require heavy machineries to achieve interesting results. For instance, Cohen et al. (2020) compared the year-over-year dissimilarities of 10K<sup>1</sup> reports using a bag-of-words approach and discovered that companies are more likely to alter a large chunk of previous year’s report only in anticipation of underperformance this year. Therefore, a portfolio that shorts “changers” and buys “non-changers” could earn up to 22% per year in alpha in the future. Lopez-Lira (2020) used LDA to extract risk factors that firms themselves identify in their 10K report

---

\*. I would like to thank Prof. Robert Reider for his guidance and supervision on this project. Responsibility for any errors in this work remains my own.

1. for more on 10K, see <https://www.sec.gov/fast-answers/answersreada10khtm.html>

and found that a model that uses only firm identified risk factors performs at least as well as traditional factor models, despite not using any information from past prices or returns.

Much of recent research has been on the 10K report, and little has been on the 8K<sup>2</sup> report, which is to notify investors of a current event. Any changes to a company’s material information should be reported within four business days to provide an update to previously filed quarterly or annual report. Circumstances that apply may include but are not limited to:

- Acquisition and merger
- Incurrence of material debt
- Senior officer appointments and departures
- Amendments or waivers of code of ethics
- Results of shareholder votes

The release of an 8K report, by nature of being current, often correlates with stock price jumps as well as the longer-term price movement. The report content is likely to be of predictive power for subsequent stock performance. Several attempts have been made by Kim et al. (2018) and others that try to fit a classifier based on the filing content to predict the direction of stock price movement. Anecdotally, a hedge fund named Derwent Capital Markets<sup>3</sup> have tried this general idea with twitter feeds and other texts data in the early 2010s and failed not too long after.

This paper takes a different approach from estimating the relationship between text structures and stock price movements, and instead applies text analysis to identify signals to aid stock selection based on some qualitative reasoning. More specifically, I focus on testing the effect of company releasing any material information related to sustainability in 8K report on the stock price in the near future. This choice of topic is based on the observation that while some aspect of sustainability is beneficial, there is certain perceived value for a company to discuss or act on trendy issues such as sustainability and ESG (environmental, social, and governance) without regard to its bottom-line, which may or may not translate into worse stock performance.

This study finds that there has been a large increase in 8K reports on sustainability since 2000, from the early days of sustainable investing when the Global Compact just produced the landmark report “Who Cares Wins” (Compact, 2004) to the end of 2019 when several sustainability ETFs have been offered and ESG has become a household name in the financial community. It also finds that the average 90 and 120 trading days excess return following the release of such 8K reports has gone from statistically significantly positive to statistically significantly negative over a 20 year period, suggesting the narrative that late-comers may have now drowned out the positive effect on stock performance associated with the early adopters and 8K reports on sustainability could now be seen as a negative trading signal to short certain stocks for an expected positive return. This paper also entertains other implications from this result. The following sections will introduce the methodologies and data for reaching aforementioned results as well as additional analysis.

---

2. for more on 8K, see <https://www.sec.gov/fast-answers/answersform8khtml.html>

3. [https://en.wikipedia.org/wiki/Derwent\\_Capital\\_Markets](https://en.wikipedia.org/wiki/Derwent_Capital_Markets)

## 2. Text Analysis

This section introduces methods for preprocessing a text document and analyzing its content.

### 2.1 Preprocessing

A text in its raw form is stored as a long string. Preprocessing is often needed before applying any text analysis algorithms. The NLTK package in Python provides most of the functionality one would need for such tasks. First, a long string is often chopped into several pieces, which is called tokenization and each piece is called a token. The general goal of tokenization is to itemize each word and punctuation from a text string, and can be done by calling the `word_tokenize` method, where we can also specify which language the text is written in for added accuracy. Next, we may choose to remove punctuations and convert all words into lowercase by call Python’s `isalpha` and `lower` methods on a string object.

In addition, a word may come in different forms, such as *perform*, *performs*, and *performing* for grammatical reasons, and in many cases, we would like them to be recognized as the same word. To do this, we can apply stemming or lemmatization to reduce words to a common base form. An important distinction of these two methods is that a stemmer will return a subset of the word by chopping off its ends with some heuristic and perhaps crude rule, which is not necessarily equal to its morphological root, whereas a lemmatizer takes speech context into consideration and will return a dictionary form of the word, which itself must also be a legitimate word. For instance, for the word “sustainability”, a stemmer and a lemmatizer may give “sustaina” and “sustainsbility”(no change) respectively. Which one to use depends on the context.

Another common procedure in text preprocessing is to remove stop-words, which refer to words such as *the*, *that*, and *and* that contribute little to the meaning of a sentence. NLTK has its own collection of stop-words by language and one can use it to filter out stop-words in a document. Additionally, there often exist multi-word expressions in a document and after tokenization, they come as separate words. There are several way to re-construct multi-word expressions. One is to manually specify which words make up a multi-word expression and iterate over the entire list of words to combine such words into one item, which can be done by calling `mweTokenizer`. Another way is to apply phrase-modeling to a document. For any two tokens  $A$  and  $B$ , we can specify the condition for when they will be recognized as a phase by

$$\frac{\text{count}(A, B) - \text{count}_{\min}}{\text{count}(A) * \text{count}(B)} * N \geq \text{threshold}$$

and combine words that meet such condition.

### 2.2 Dictionary Method

Now that the text document is a list of words, the dictionary method refers to computing the frequency for a pre-defined set of words and use this information to gauge the content of a document. This straightforward method is often used as a benchmark for more sophisticated models. In the context of sustainability, we can compute the frequency of certain words

such as *carbon neutral*, *diverse workforce*, and *community impact*, and identify documents with a high count of such words as sustainability documents.

## 2.3 Topic Model: Latent Dirichlet Allocation

In natural language processing, a topic model is a statistical model for discovering latent structures of a document, and Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003) is a generative statistical model that assumes each document has a mixture distribution over an underlying set of topics, and each topic is modeled as a mixture over an underlying set of word probabilities. More specifically, LDA assumes the following generative process for a corpus  $D$  consisting of  $M$  documents each of length  $N_i$  and there are  $K$  topics to be identified:

- 1 Choose  $\theta_i \sim \text{Dir}(\alpha)$ , where  $i \in \{1, \dots, M\}$  and  $\text{Dir}(\alpha)$  is a Dirichlet distribution with parameter  $\alpha$  which is often set small ( $\alpha < 1$ ), suggesting each document would contain a mixture of very few topics
- 2 Choose  $\varphi_k \sim \text{Dir}(\beta)$ , where  $k \in \{1, \dots, K\}$  and  $\beta$  is usually set small as well
- 3 For each of the word positions  $i, j$ , where  $i \in \{1, \dots, M\}$ , and  $j \in \{1, \dots, N_i\}$ 
  - a Choose a topic  $z_{i,j} \sim \text{Multinomial}(\theta_i)$ .
  - b Choose a word  $w_{i,j} \sim \text{Multinomial}(\varphi_{z_{i,j}})$ .

Note that multinomial distribution here refers to the multinomial with only one trial, which is known as the categorical distribution and the lengths  $N_i$  are treated as independent of all the other data generating variables ( $w$  and  $z$ ). Inference and parameter estimations for LDA make use of variational methods and the EM algorithm, and details can be found in the original paper by Blei et al. (2003). To aid with understanding LDA, Lopez-Lira (2020) presents the model in terms of the matrix factorization, as shown in figure 1.

## 3. Data

There are three main sources of data used in this paper: CRSP and other databases available in WRDS to fetch price and fundamental data, the SEC Edgar database to scrape for company filings, and corporate websites to scrape for content on sustainability for training a topic model.

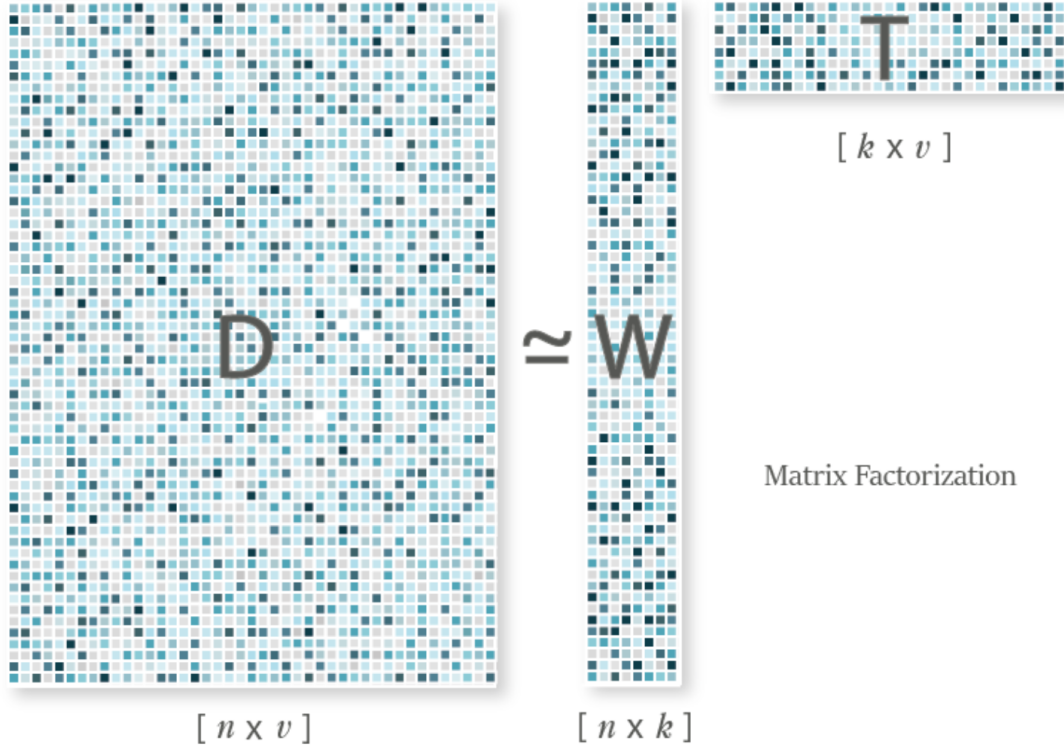
### 3.1 CRSP and other databases in WRDS

CRSP is used to construct the stock universe for this study. First, only US common stocks are considered, which have a share code of 10 or 11. Since the SEC starts to require companies to submit filings electronically as of May 1996<sup>4</sup> and hence all historical filings are available on its Edgar database thereafter, only stocks that are listed from May 1996 to Dec. 2019 are considered. They need not be listed throughout the entire period. Stocks that have less than 24 months of return history or fall under 500 million market capitalization at any point in this period are excluded. There is a total of 728 stocks in the stock universe.

---

4. <https://help.edgar-online.com/edgar/history.asp?site=pro>

Figure 1: LDA as Matrix Factorization. Source: Lopez 2020



The matrix  $D$  on the left is a document-term matrix where each cell represents the term-frequency or term-frequency-inverse-document-frequency (tf-idf) for word  $v_j$  in document  $n_i$ . With LDA, we obtain, for document  $n_i$ , the probabilities of being assigned to each one of  $k$  topics, which corresponds to row  $i$  in matrix  $W$ , and for topic  $k_j$ , the probabilities of being given each one of  $v$  words, which corresponds to row  $j$  in matrix  $T$ .  $W \times T$  can be thought of as approximating matrix  $D$  via  $k$  rank 1 matrices and each rank 1 matrix is the product of  $i$ th column vector in  $W$  and  $i$ th row vector in  $T$  for  $i \in \{1, \dots, k\}$ . With the choice of  $k$  reasonably small (less than the rank of  $D$ ), LDA can also be viewed as a dimension reduction technique applied to matrix  $D$ .

When computing the returns after the release of a 8K report, CRSP is used to fetch price history. Due to my subscription limit, price history is only available through the end of 2019, so for events that happened during the second half of 2019 and require price history in 2020, Yahoo Finance data are used. Other variables are also used to evaluate the price movement after an event which include trading volumes from CRSP, earnings announcement dates from I/B/E/S, Fama-French three factor portfolios and 49 industry portfolios from their website.

### 3.2 8K Filings

The SEC Edgar database offers a web interface to query historical filings. By entering the CIK number (the unique identifier for companies registered with the SEC), type of filings and some date range, it returns a page of search results with hyperlinks for every filing, ordered from the most recent to the least. Each filing is often consisted of multiple documents that include the main report, extension reports, and other miscellaneous files such as PowerPoint slides and images. Figure 2 and figure 3 are examples of search and filing document pages respectively.

Figure 2: Edgar Search Result Page

The screenshot shows the SEC Edgar Search Results page for International Business Machines Corp. (CIK: 0000051143). The page displays a list of filings with the following columns: Filings, Format, Description, Filing Date, and File/Film Number. The filings are ordered from most recent to least recent.

Filings	Format	Description	Filing Date	File/Film Number
8-K	Documents	Current report, Item 8.01 Acc-no: 0001104659-19-066657 (34 Act) Size: 447 KB	2019-11-22	001-02360 191241907
8-K	Documents	Current report, Items 5.02 and 9.01 Acc-no: 0001104659-19-062171 (34 Act) Size: 457 KB	2019-11-12	001-02360 191206436
8-K	Documents	Current report, Items 7.01 and 9.01 Acc-no: 0001558370-19-008942 (34 Act) Size: 1 MB	2019-10-17	001-02360 191155387
8-K	Documents	Current report, Items 2.02, 7.01, and 9.01 Acc-no: 0001558370-19-008923 (34 Act) Size: 4 MB	2019-10-16	001-02360 191152911
8-K	Documents	Current report, Items 5.02 and 9.01 Acc-no: 0001144204-19-045696 (34 Act) Size: 441 KB	2019-09-24	001-02360 191109362
8-K/A	Documents	[Amend] Current report, Item 9.01 Acc-no: 0001558370-19-008675 (34 Act) Size: 1 MB	2019-09-20	001-02360 191105162
8-K	Documents	Current report, Items 7.01 and 9.01 Acc-no: 0001558370-19-006914 (34 Act) Size: 1 MB	2019-08-02	001-02360 19094981
8-K	Documents	Current report, Items 7.01 and 9.01 Acc-no: 0001558370-19-006802 (34 Act) Size: 6 MB	2019-08-02	001-02360 19094940
8-K	Documents	Current report, Items 1.01, 2.03, and 9.01 Acc-no: 0000950157-19-000769 (34 Act) Size: 942 KB	2019-07-19	001-02360 19064079
8-K	Documents	Current report, Items 7.01 and 9.01 Acc-no: 0001558370-19-006151 (34 Act) Size: 1 MB	2019-07-18	001-02360 19061636
		Current report, Items 2.02, 7.01, and 9.01		001-02360

I develop a script for scraping 8K filings by mimicking how a person would go about searching and downloading desired filing documents. It does the following:

#### Algorithm 1: Scraping SEC Filing Reports


```

for each (CIK, Filing.type, Date.range) do
    request search result that contains meta-data for all applicable filings;
    save meta-data and each filing's URL;
    for each URL do
        request filing page and save filing text content;
    end
end

```

More specifically, to request a page of search result, a uniform resource locator (URL), also commonly known as the web address, is generated by taking the end-point `r"https://www.sec.gov/cgi-bin/browse-edgar"` and a parameter dictionary that specifies the request to search filings with filing-specific strings such as `CIK`, `type`, and `dateb`, and this URL will return a user-oriented web page like that in figure 2. For scraping

Figure 3: Edgar Filing Document Page



[Home](#) | [Latest Filings](#) | [Previous Page](#)

U.S. Securities and Exchange Commission

Filing Detail

[Search the Next-Generation EDGAR System](#)

[SEC Home](#) » [Search the Next-Generation EDGAR System](#) » [Company Search](#) » [Current Page](#)

Form 8-K - Current report

Filing Date

2020-12-18

Accepted

2020-12-18 16:13:17

Documents

15

Period of Report

2020-12-15

Items

Item 5.02: Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers; Compensatory Arrangements of Certain Officers  
Item 5.03: Amendments to Articles of Incorporation or Bylaws; Change in Fiscal Year  
Item 8.01: Financial Statements and Exhibits

SEC Accession No.

0001104659-20-137423

Document Format Files

Seq	Description	Document	Type	Size
1	FORM 8-K	tm2038625-2_8k.htm	8-K	89998
2	EXHIBIT 3.2	tm2038625d2_ex3-2.htm	EX-3.2	165533
3	EXHIBIT 99.1	tm2038625d2_ex99-1.htm	EX-99.1	15284
	Complete submission text file	0001104659-20-137423.txt		672143

Data Files

Seq	Description	Document	Type	Size
4	XBRL TAXONOMY EXTENSION SCHEMA	ibm-20201215_xsd	EX-101.SCH	8771
5	XBRL TAXONOMY EXTENSION DEFINITION LINKBASE	ibm-20201215_def.xml	EX-101.DEF	36758
6	XBRL TAXONOMY EXTENSION LABEL LINKBASE	ibm-20201215_lab.xml	EX-101.LAB	59833
7	XBRL TAXONOMY EXTENSION PRESENTATION LINKBASE	ibm-20201215_pre.xml	EX-101.PRE	34750
8	EXTRACTED XBRL INSTANCE DOCUMENT	tm2038625-2_8k.htm.xml	XML	28721

INTERNATIONAL BUSINESS MACHINES CORP (Filer) CIK: 0000051143 (see all company filings)

IRS No.: 130871985 | State of Incorp.: NY | Fiscal Year End: 1231

Type: 8-K | Act: 34 | File No.: 001-02360 | Film No.: 201400758

SIC: 3570 Computer & office Equipment

Office of Technology

Business Address

1 NEW ORCHARD ROAD  
ARMONK NY 10504  
9144991900

Mailing Address

1 NEW ORCHARD RD  
ARMONK NY 10504

purposes, I specify an additional string output as `atom`, which will return the page in Extensible Markup Language (XML) format, which clearly indicates a page's sub-content via meta-data labels. Figure 4 shows the same page as that in figure 2 in XML.

Figure 4: Edgar Search Result Page in XML

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<feed xmlns="http://www.w3.org/2005/Atom">
  <author>
    <email>webmaster@sec.gov</email>
    <name>Webmaster</name>
  </author>
  <company-info>
    <addresses>
      <address type="mailing">
        <city>ARMONK</city>
        <state>NY</state>
        <street1>1 NEW ORCHARD RD</street1>
      </address>
      <address type="business">
        <city>ARMONK</city>
        <phone>9144991900</phone>
        <state>NY</state>
        <street1>1 NEW ORCHARD ROAD</street1>
      </address>
    </addresses>
    <assigned-sic>3570</assigned-sic>
    <assigned-sic-desc>COMPUTER & OFFICE EQUIPMENT</assigned-sic-desc>
    <assigned-sic-href>https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&SIC=3570&owner=exclude&count=100</assigned-sic-href>
    <cik>0000051143</cik>
    <cik-href>https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=0000051143&owner=exclude&count=100</cik-href>
    <conformed-name>INTERNATIONAL BUSINESS MACHINES CORP</conformed-name>
    <fiscal-year-end>1231</fiscal-year-end>
    <office>Office of Technology</office>
    <state-location>NY</state-location>
    <state-location-href>https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&State=NY&owner=exclude&count=100</state-location-href>
    <state-of-incorporation>NY</state-of-incorporation>
  </company-info>
  <entry>
    <category label="form type" scheme="https://www.sec.gov/" term="8-K" />
    <content type="text/xml">
      <accession-number>0001104659-19-066657</accession-number>
      <act>34</act>
      <file-number>001-02360</file-number>
      <file-number-href>https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&filenum=001-02360&owner=exclude&count=100</file-number-href>
      <filing-date>2019-11-22</filing-date>
      <filing-href>https://www.sec.gov/Archives/edgar/data/51143/000110465919066657/0001104659-19-066657-index.htm</filing-href>
      <filing-type>8-K</filing-type>
      <file-number>191241907</file-number>
      <form-name>Current report</form-name>
      <items-desc>Item 8.01</items-desc>
      <size>447 KB</size>
      <xbrl_href>https://www.sec.gov/cgi-bin/viewer?action=view&cik=51143&accession_number=0001104659-19-066657&xbrl_type=v</xbrl_href>
    </content>
  </entry>
</feed>
```

The search result page provides meta-data for each filing including its URL but not its content. Additional step is needed to visit the filing page and scrape for its content. The tricky part is that not every filing page comes in XML format<sup>5</sup>, so its complete submission text file is parsed, which include filing documents, HTML tags and CSS tags. By identifying filing related tags such as `8K` and `EX-*` which stands for extension document and is part of the official filing documents, I extract content strings and use `unicodedata.normalize` method to normalize any unicode to human-readable strings.

Observing that pages in a filing document are split by setting `hr` to have 100% width, each document in a filing is parsed as a list of strings with each item corresponding to a page. This granularity will help with selecting sustainability filings as this topic may only come up in a small portion of the filing, which may be drowned out if all pages are viewed as a single string.

For every tuple of CIK/company, filing type (8K), and date range (set as before Dec. 31 2019 for this study), a Python dictionary is populated with the structure shown in figure 5 and is saved locally as a JSON file for readability. There are a handful of companies where certain meta-data labels do not exist, and an empty string is placed as exception handling. A detailed log of the scraping process is available on my Github page<sup>6</sup> and the entire process took around 48 hours to complete on my personal laptop with 16 CPUs and a 64G RAM. Multi-threading or multi-processing cannot be used to speed things up since SEC Edgar imposes a 10 requests per second limit<sup>7</sup> and an exceedance may result in the IP address sending the request getting blocked for 10 minutes. Only 1 out of 728 companies I try to scrape for run into issues that could not be fixed in a reasonable amount of time (when parsing an extension document, the immediately following image file was also parsed as text, which broke `BeautifulSoup`), and therefore the success rate is 99.86%. It is worth noting that, though not extensively tested, the design of the script allows scraping any type of SEC filings amenable with little to no adjustment needed.

### 3.3 Sustainability Corpora

Another source of text data needed for this study is a collection of corpora pre-labeled as sustainability, so that we could apply the Latent Dirichlet Allocation (LDA) and extract sustainability topics and its distribution of words, to identify sustainability texts in a filing document with. Ideally, I would like to obtain three general topics, which correspond to three aspects of ESG, that is, environmental, social and governance. There are several collections of text data freely available such as Reuters-21578 and Yelp reviews, and some companies like Parable.ai have manually curated sustainability documents for internal use, but as of writing, I am not aware of any text collection that is pre-labeled as sustainability and also freely available to the public. Considering the circumstances and inspired by the work of Sozzi (2017), I decide to create my own collection sustainability corpora by scraping for sustainability content from corporate websites.

I build a scraper based on the `scrapy` framework, and the general idea is to extract content from a web page whose URL contains sustainability keywords. To increase effi-

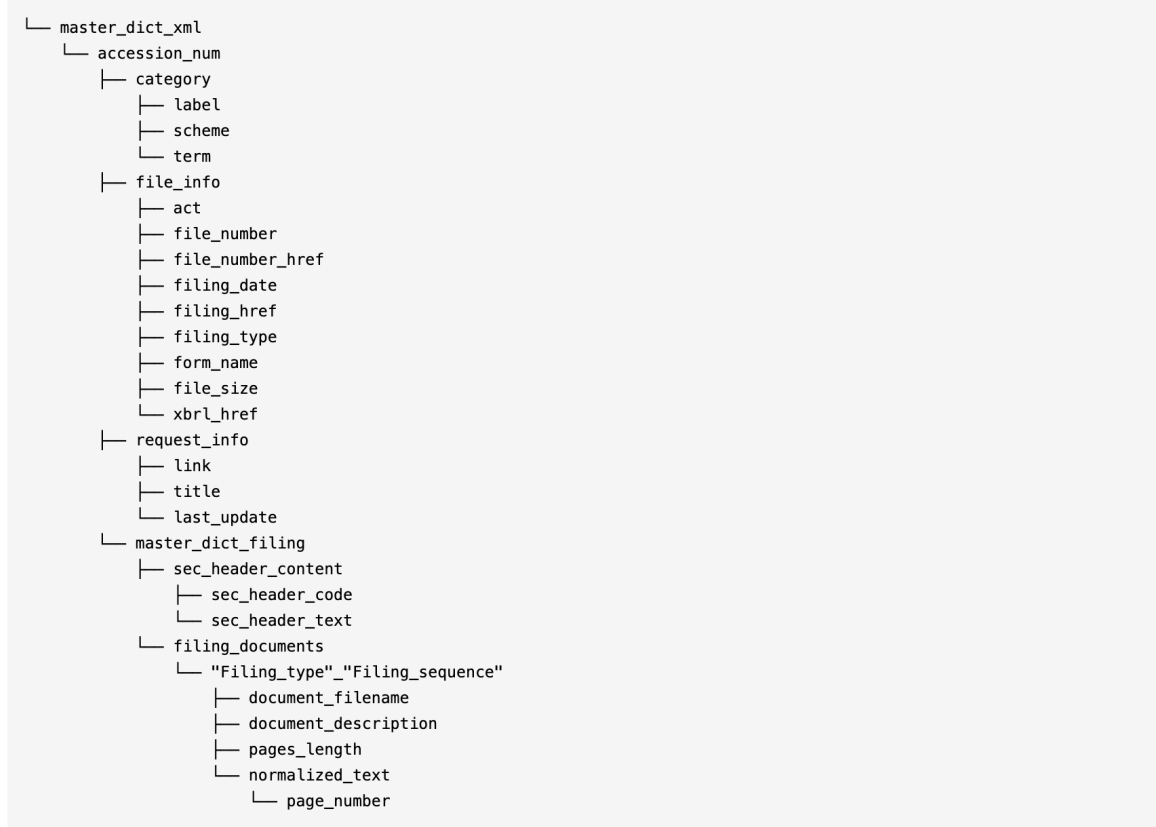
5. SEC is phasing in the requirement to submit filings in XBRL format, which is similar to XML, but historical filings will likely stay in their original `txt` format

6. [https://github.com/cccccody/MathFin\\_Project\\_20Fall/logs/download\\_filings/](https://github.com/cccccody/MathFin_Project_20Fall/logs/download_filings/)

7. <https://www.sec.gov/privacy.htm#security>



Figure 5: Structure of Dictionary to Store Scraped Content



Each filing is uniquely identified by its accession number. The search result page provides information for **category**, **file\_info**, and **request\_info**, and every filing's complete submission text file provides for entries in **master\_dict\_filing**.

ciency, I also create a list of filtering keywords to identify pages most likely not related to sustainability such as *forums*, *support*, and *developer* so that the web crawler can skip it if the URL does not contain any of the sustainability keywords to begin with. A complete list of sustainability keywords used by the scraper is given by table 1, and filtering keywords by table 2.

Table 1: Sustainability Keywords for URL

sustainability	inclusive-environment	social-justice	carbon-footprint
clean-energy	carbon-neutral	corporate-governance	diverse-workforce
community-impact	clean-water	climate-change	social-impact
social-responsibility	corporate-responsibility	global-warming	consumer-privacy

Table 2: Filtering Keywords for URL

document	blog	product	profit
accessories	shop	support	developer
revenue	archive	search	login
dmg	forums	investor	news

In short, my **scrapy** scraper does the following:

---

**Algorithm 2:** Scraping Websites for Sustainability Text

---

```

for each  $URL_{seed}$  do
    web-request  $URL_{seed}$ ;
    if success then
        domain_origin = urlparse( $URL_{seed}$ ).netloc;
        URL_links = extract_links(response_ $URL_{seed}$ );
        remove URL_links that do not contain sustainability keywords but contain
        filtering keywords;
        for each  $URL_{link}$  do
            domain_this = urlparse( $URL_{link}$ ).netloc;
            if  $domain\_origin == domain\_this$  then
                web-request  $URL_{link}$ ;
                extract text content;
                save to local database;
            end
        end
    else
        continue ;
    end
end

```

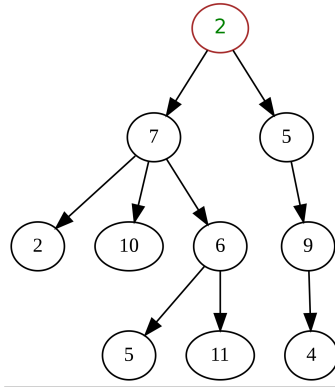
---

It starts with a list of seed URLs and I use the corporate website addresses provided by Compustat from all stocks in the stock universe. Websites react to a scraping bot with different levels of sophistication and most would deny access if the bot's identify and intend are not properly specified. At a minimum, I define the **USER\_AGENT** variable with information specific to my computing environment, which grants me access to most of the web pages.

The domain information of each URL is extracted and checked if it is the same as the seed URL before making a request, in order to make sure the bot does not crawl outside into the vast world of the internet. Strings in each out-going link/URL are checked for sustainability and filtering keywords to determine the set of URLs to visit next for increased efficiency.

Most websites assume a tree structure shown in figure 6. The default crawling mode of **scrapy** is Depth-First-Search (DFS). Given that various links to sustainability content are likely on the same level falling under the same node and that DFS may cause the bot to get stuck in one branch of the tree, I specify the bot to traverse in Breath-First-Search (BFS). In comparison, BFS generates 10 times more sustainability content than DFS on average.

Figure 6: Tree Data Structure, Source: wikipedia



In addition to HTML and CSS syntax, which are relatively easy to exclude, a web page often includes non-essential text content such as navigation panels, comment sections and advertisements. To extract a web page’s essential content, I use **DragNet** developed by Utiu and Ionescu (2018), which is shown to have the highest precision and lower recall than using **Readability**, **BeautifulSoup.get\_text()**, and **<p>**. **DragNet**’s high precision suits this study particularly well since I would like to obtain the essential content of a sustainability page with a high degree of confidence to reduce the mislabeling rate, and the fact that it may miss other content is of little problem, which can be alleviated by running the scraper longer or on more websites.

Some of the websites visited are shown in languages other than English. Though **DEFAULT\_REQUEST\_HEADERS** is set to explicitly ask for content in English, it is up to the web developer on the other side whether or not their web server shall respond to this request and present the content in English. I use **langid** developed by Lui and Baldwin (2012) to detect and skip web pages not in English.

The extracted content as well as meta-data such as company name, keywords identified are written to a local **MongoDB** database, which is a document-oriented NoSQL database and suitable for storing unstructured data such as text.

With over 700 companies, many of which have a well-developed website with hundreds of thousands of links, it is not clear how long a complete traversal will take. To increase the speed, I implement a concurrency-enabled **scrapy** scraper that can process 16 URLs at the same time. To avoid getting stuck with a particular website, I run the scraper three times overnight and rearrange the order of web addresses before each run, with companies having a larger market capitalization placed in the front. In the end, I obtain around 50,000 entries and 14,000 unique entries. Top keywords for identifying sustainability web pages are *sustainability*, *climate-change*, and *corporate-responsibility*.

**Visualization of topics** With content extracted, we can apply LDA and check for the distribution of topics and words to see if they resemble what would be expected from sustainability texts. To do this, I use **GenSim** by Rehřek and Sojka (2011) to compute LDA topics and distributions. I then use **LDavis** by Sievert and Shirley (2014) to visualize LDA topics. Figure 7 shows the visualization that assumes  $K = 20$  topics.

Figure 7: LDA Topics from Sustainability Texts



The size of the circle is proportionate to its marginal topic distribution, and the center positions are computed by projecting inter-topic distances to 2-dimensions via multidimensional scaling. There are three groups of topics farthest apart from each other, in the upper left, lower left and far right region. To view the word distribution of a topic, let us first define the *relevance* of word  $w$  to topic  $k$  given a weight parameter  $\lambda$  (where  $0 \leq \lambda \leq 1$ ) as:

$$r(w, k | \lambda) = \lambda \log(\phi_{kw}) + (1 - \lambda) \log\left(\frac{\phi_{kw}}{p_w}\right)$$

where  $\lambda$  determines the weight between the estimated frequency of word  $w$  within topic  $k$  and that frequency normalized by the overall frequency of word  $w$ , both measured in log. Sievert and Shirley (2014) estimated  $\lambda = 0.6$  to be the optimal value for topic interpretation,

so I rank each word in a topic by relevance with  $\lambda = 0.6$  and the top words for topic 1, 2, and 8 are shown in table 3.

Table 3: Top Words in Topic by Relevance

Topic 1	Topic 2	Topic 8
energy	compliance	society
emission	conduct	environment
facility	ethic	action
renewable	division	culture
company	code	materializes
water	corruption	solidarity
maintenance	business	dividend
environmental	company	climate

Despite shared words such as *company*, and *environment*, these three topics represent fairly distinct themes, which roughly correspond to three aspects of ESG. Topic 1 has words *emission*, *water*, and *renewable* which refer to environmental issues, topic 2 has *compliance*, *ethic*, and *corruption* which refer to corporate governance, and topic 8 has *society*, *culture*, and *solidarity* which refer to social impact. The themes of these topics can be extrapolated to other topics nearby, which largely represent the three clustered regions in the plot. In other words, the scraped sustainability texts are indeed about sustainability and can capture a good amount of diversity and granularity in sustainability topics. That said, the distribution of topics depend on the parameter  $K$ . One might set  $K = 3$  and expect to see three topics corresponding perfectly to three aspects of ESG, but this is generally not the case due to the imbalance of texts on different ESG topics in the corpora.

#### 4. Trading Sustainability

To test the idea that companies may issue 8K reports with some portion dedicated to sustainability/ESG only for marketing and other non-performance related reasons, which could result in a decrease in stock price, I select historical filings on sustainability, and compute stock returns after the event of releasing such 8K filings. This is essentially to backtest an event-driven trading strategy.

#### 4.1 Benchmark: Keywords Matching

There is more than one way to determine the content of a filing with varying degrees of sophistication. A natural benchmark method is to select ones that contain any of pre-specified sustainability keywords:

---

**Algorithm 3:** Benchmark: Keywords Matching

---

```
for each filing do
  for each document do
    for each page do
      preprocess text;
      if contains keywords then
        save filing company, and date;
        continue to next filing;
      end
    end
  end
end
```

---

The text is preprocessed using methods introduced in the previous section. Tokenization, case-lowering, punctuation-removal, lemmatization, stop-words-removal, and multi-word-reconstruction are applied in this order. The keywords used are given in table 4, and lemmatization is also applied to keywords before being used for matching.

Table 4: Keywords for Keywords Matching

carbon neutral	carbon footprint	clean water
pollution mitigation	climate change	global warming
waste management	community impact	consumer privacy
diverse workforce	labor standard	corporate responsibility
executive compensation	sustainability	social responsibility

A single pass of running keywords matching on all filings takes around 24 hours. I implement multi-processing on my personal laptop with 16 CPUs and reduce the time to about 4 hours.

#### 4.2 Identifying Filings with LDA

Another way is to compare the empirical distribution of words in each filing to that of sustainability topics. I take the topic distributions produced by the LDA model trained using web-scraped sustainability corpora, and compute the probability of assigning each topic to a given filing text. The filing texts used in this case are the ones that have already been selected by keywords matching. There is very little chance for filings to contain sustainability content when they do not have any sustainability keywords to begin with.

The filing that is assigned to some sustainability topic with a high probability above some threshold (set as 0.75) is selected.

---

**Algorithm 4:** Identifying Filings with LDA

---

```

for each filing do
  for each document do
    for each page do
      preprocess text;
      Infer LDA-produced sustainability topics' distribution on text;
      if largest probability > threshold then
        save filing company, and date;
        continue to next filing;
      end
    end
  end
end

```

---

The underlying thinking is that when a given text is not related to sustainability, any LDA-produced sustainability topic will be just as well or just as poorly at explaining the filing text, so the total probability 1 would be split among all topics and each computed probability remains small. When there is a high degree of confidence in assigning a sustainability topic to a filing, it suggests the content of the filing is likely about sustainability.

### 4.3 Results

Most of the identified filings are concentrated in a few NAICS<sup>8</sup> sectors and the percentage of filings selected with LDA is roughly the same as shown in figure 8.

Chronologically, we see more sustainability filings over the years as shown in figure 9. The labels on the x axis Year<sub>1</sub> – Year<sub>2</sub> denote the time period from the beginning of year 1 to the end of year 2. It is worth noting that the percentage of filings selected from keywords matching by LDA increases from 2000 to 2012, and has been on the decrease since then. We are past the peak in terms of both the absolute number of filings and the percentage of those indicated by LDA, perhaps suggesting that sustainability is no longer a novel concept and everything-sustainability has entered a stable period and likely to stay for the long run.

Historical returns are shown in table 5 and table 6 for selected filings from keywords matching. The market factor in Fama-French three factor portfolios is used as the market benchmark, and Fama-French 49 industry portfolio returns are used as the industry benchmark. Values in parentheses are p-values of t-test on all event returns in a given period, with the null hypothesis being the returns are no different from 0. 8K report release that coincides with known events that cause significant drift in stock prices both short-term and long-term are excluded which include earnings and merger/acquisition announcements. Trading volume on the event day is used to compare with historical average and if abnormal trading volumes take place on the release day, such 8K filing is also excluded. The returns for LDA selected filings are computed the same way, and are given in table 7 and table 8.

There is a clear pattern in both keywords matching and LDA excess returns in that the early days show a statistically significant positive return while more recently, the return

---

8. [https://en.wikipedia.org/wiki/North\\_American\\_Industry\\_Classification\\_System](https://en.wikipedia.org/wiki/North_American_Industry_Classification_System)

Figure 8: Sustainability Filings Count by Sector

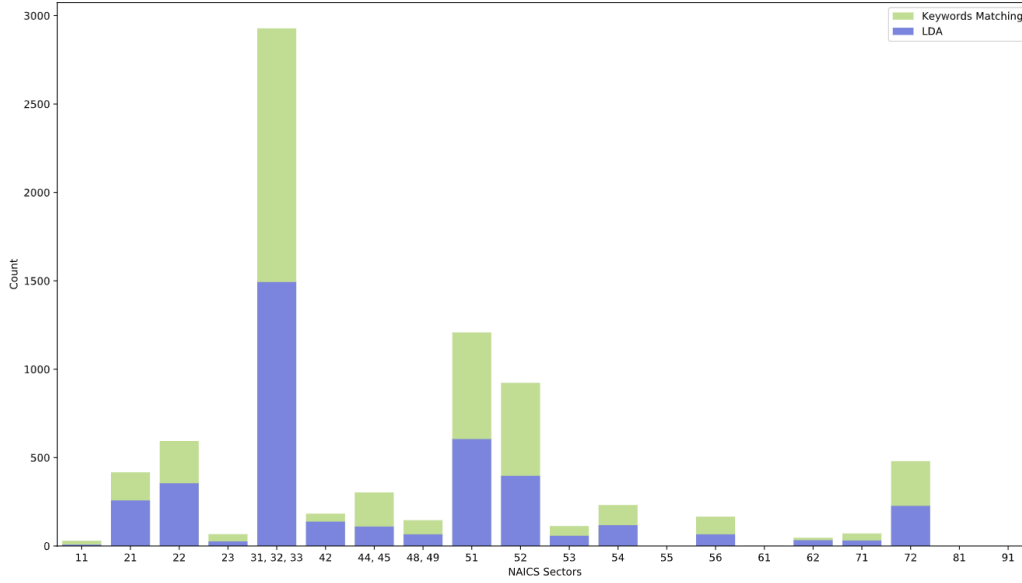
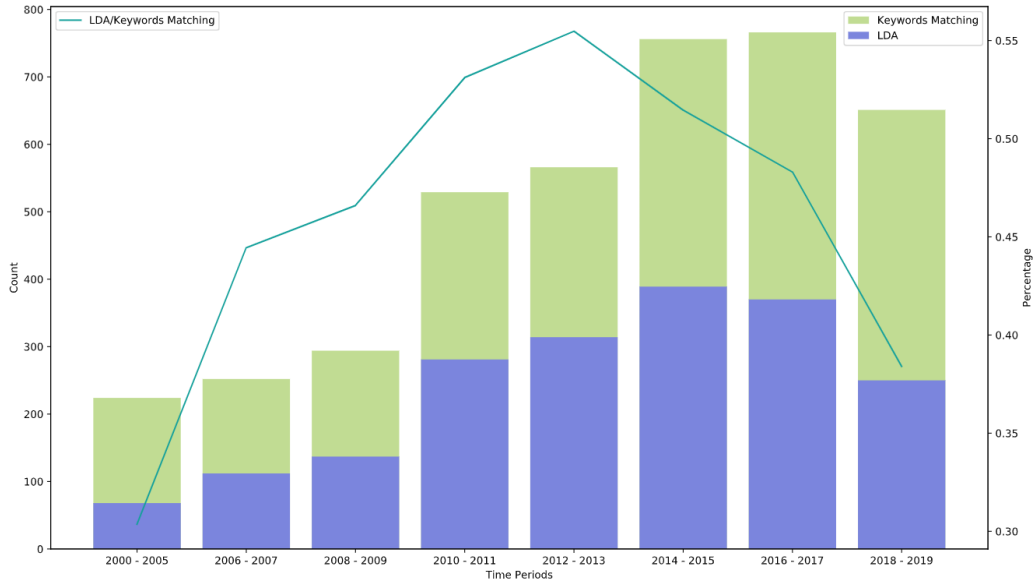


Figure 9: Sustainability Filings Count by Time Periods



has gone to a statistically significant negative, at the end of 90 and 120 trading days after event respectively. Such pattern also exists in keywords matching excess returns for 30 and 60 trading days but not so much for LDA excess returns.

Suppose the LDA model is able to correctly select filings with sustainability content, then the return pattern shows how at the beginning, a focus on sustainability correlates with strong corporate operations and stock performance. One can conjecture that when the



concept of sustainability was first introduced and seriously considered by the early adopters, much emphasis was on how it can boost a company’s bottom line and be integrated into a company’s fundamentals. As sustainability enters mainstream media and the activist investing movement takes shape with issues such as corporate social responsibility, carbon foot-print and diversity on top of its priority list, there is an increasing publicity value in disclosing a company’s position and action related to sustainability. For large-cap and client-facing companies, this is especially true. Companies such as Apple, Microsoft or Goldman Sachs have a fully developed web space on everything-sustainability that goes on in the company and issues sustainability report<sup>9</sup> and outlooks on a regular basis.

Other companies may follow suit, which could explain the increase in 8K reports on sustainability over the years. The average negative returns in recent years suggest that the release of sustainability related information indicates a poor stock performance in near future. A closer look at the early sustainability 8K reports show that most of the content is on corporate governance, such as the appointment of senior directors, vote result of a new compensation plan and the establishment of new divisions or committees. A recent study by Pedersen et al. (2020) shows that the E (carbon emission) and S (non-sin) in ESG have a strong negative relationship with gross profit, while G (low accruals) has a strong positive relationship. Therefore, it is possible that the increased number of filings on E and S cause the overall negative returns. That said, sustainability or ESG as a whole, is a fairly strong indicator of poor stock performance in today’s market, and this could be used either as a selection criterion for shorting stocks or a filtering criterion when constructing a stock portfolio.

## 5. Conclusion and Future Steps

In this paper, I have presented methods for text analysis, from preprocessing to topic modeling and apply them to SEC filing documents. I have also introduced procedures for scraping filings document from SEC Edgar and for scraping targeted content from corporate websites. Given the observation that some companies sometimes report on trendy issues such as sustainability for publicity values without regards to its corporate performance, I have tested how a release of 8K filing mentioning sustainability affects stock performance in the near future. The number of such 8K filings increases over the years, and the return results suggest a strong positive return associated with sustainability filings in the early days and a strong negative return in more recent years. There is more than one interpretation that could apply to this result. For instance, when late-comers start to report on sustainability in their 8K filings, they might have drowned out the positive effect with sustainability reporting associated with the early adopters. On the other hand, as the early sustainability filings seem to be predominantly about G or corporate governance, which is well documented to have a positive effect on stock performance, and the more recent ones to include environmental and social issues, such shift may also explain the change in returns.

There are certainly areas in this study that could benefit from additional work and analysis. The selection of sustainability filings with LDA based on the probability distribution of topics is not entirely robust. While it is plausible that in general, when a filing is not related to sustainability, every sustainability topic produced by LDA will be just as poorly

---

9. Example: <https://www.goldmansachs.com/s/sustainability-report/index.html>

at explaining the content, which results in low probability values for all topics. It is possible that a topic is assigned to a filing with a high degree of confidence spuriously, simply because the empirical distribution of words in the filing happens to be much farther from the other topics which inflate the actual resemblance. A potential fix is to use a LDA model trained on both sustainability and generic filing documents. In addition, a classifier can be trained using LDA topic distributions as input to differentiate between sustainability and generic filing content.

It would also be interesting to test the effects of E, S, and G in ESG separately. As noted previously, much of the early sustainability filings are about corporate governance issues. How each aspect of ESG affects stock performance would offer additional insight in evaluating sustainability as a potential trading signal. Last but not least, more statistical analyses are needed to assess the robustness of backtested results. Confounding variables such as industry classification, market capitalization, B2B/B2C, should be included.

The next steps will start by implementing aforementioned ideas. To stay up-to-date on this project, please feel free to check my Github<sup>10</sup> page. Thoughts and comments are welcome.

---

10. [https://github.com/cody-wan/Text\\_8K\\_ESG](https://github.com/cody-wan/Text_8K_ESG)

Table 5: Keywords Matching - Excess Return relative to Market

t0	t1	sample size	1d - MKT	10d - MKT	30d - MKT	60d - MKT	90d - MKT	120d - MKT
1/1/00	12/31/05	224	0.36% (9.89%)	0.49% (22.03%)	2.13% (0.6%)	3.84% (0.09%)	6.3% (0.01%)	7.86% (0.0%)
1/1/06	12/31/07	252	0.47% (5.13%)	0.28% (50.8%)	0.71% (24.47%)	2.67% (0.33%)	3.64% (0.22%)	5.16% (0.02%)
1/1/08	12/31/09	294	0.14% (67.14%)	-0.41% (55.3%)	0.92% (31.37%)	2.45% (6.47%)	3.42% (3.18%)	3.29% (7.82%)
1/1/10	12/31/11	529	0.24% (6.27%)	0.07% (78.39%)	0.08% (84.35%)	0.38% (55.49%)	-0.32% (63.95%)	-0.43% (61.16%)
1/1/12	12/31/13	566	-0.04% (71.12%)	0.61% (1.65%)	0.73% (10.94%)	1.65% (1.54%)	2.33% (0.75%)	3.62% (0.01%)
1/1/14	12/31/15	756	0.0% (97.08%)	-0.27% (25.7%)	-1.25% (0.2%)	-2.45% (0.0%)	-3.52% (0.0%)	-4.93% (0.0%)
1/1/16	12/31/17	766	0.02% (89.62%)	-0.32% (18.77%)	-0.6% (8.84%)	-0.78% (11.17%)	-0.97% (10.17%)	-0.74% (30.65%)
1/1/18	12/31/19	651	-0.02% (90.07%)	-0.23% (34.53%)	-0.94% (2.36%)	-3.11% (0.0%)	-5.5% (0.0%)	-6.76% (0.0%)

Table 6: Keywords Matching - Excess Return relative to Industry

t0	t1	sample size	1d - IDT	10d - IDT	30d - IDT	60d - IDT	90d - IDT	120d - IDT
1/1/00	12/31/05	224	0.35% (10.79%)	0.41% (29.09%)	1.8% (1.18%)	3.03% (0.49%)	4.95% (0.07%)	5.79% (0.02%)
1/1/06	12/31/07	252	0.3% (20.22%)	-0.02% (95.79%)	-0.0% (99.41%)	0.89% (28.66%)	0.76% (49.02%)	1.3% (29.23%)
1/1/08	12/31/09	294	0.18% (53.14%)	-0.08% (89.67%)	1.3% (9.47%)	1.72% (12.46%)	1.94% (15.42%)	1.78% (24.65%)
1/1/10	12/31/11	529	0.22% (7.37%)	-0.05% (83.62%)	-0.31% (38.85%)	-0.14% (82.03%)	-0.79% (21.55%)	-0.87% (26.23%)
1/1/12	12/31/13	566	-0.01% (91.52%)	0.66% (0.58%)	0.8% (5.9%)	1.69% (0.72%)	2.18% (0.68%)	2.99% (0.07%)
1/1/14	12/31/15	756	-0.02% (86.27%)	-0.23% (33.02%)	-1.11% (0.39%)	-1.83% (0.06%)	-2.69% (0.0%)	-3.89% (0.0%)
1/1/16	12/31/17	766	0.02% (86.68%)	-0.16% (47.01%)	-0.51% (12.23%)	-0.55% (23.88%)	-0.91% (11.33%)	-0.99% (16.16%)
1/1/18	12/31/19	651	-0.1% (36.91%)	-0.27% (26.38%)	-0.68% (8.07%)	-2.5% (0.0%)	-4.2% (0.0%)	-5.2% (0.0%)

Table 7: LDA - Excess Return relative to Market

t0	t1	sample size	1d - MKT	10d - MKT	30d - MKT	60d - MKT	90d - MKT	120d - MKT
1/1/00	12/31/05	68	0.005 (27.32%)	0.0078 (19.3%)	0.0187 (11.52%)	0.04 (2.25%)	0.089 (0.05%)	0.113 (0.01%)
1/1/06	12/31/07	112	0.0054 (18.39%)	0.0063 (40.95%)	0.016 (12.26%)	0.0305 (2.69%)	0.0413 (2.88%)	0.0634 (0.54%)
1/1/08	12/31/09	137	0.0003 (95.54%)	0.0033 (70.54%)	0.0187 (13.02%)	0.0164 (40.57%)	0.0201 (41.14%)	0.0187 (53.52%)
1/1/10	12/31/11	281	0.0018 (25.26%)	0.0013 (65.5%)	-0.0005 (91.16%)	-0.0037 (68.0%)	-0.0151 (8.46%)	-0.0139 (21.28%)
1/1/12	12/31/13	314	-0.0003 (86.91%)	0.0052 (14.29%)	0.0013 (81.59%)	0.0079 (36.04%)	0.009 (38.8%)	0.0238 (5.26%)
1/1/14	12/31/15	389	0.0009 (68.69%)	-0.0024 (52.12%)	-0.0176 (0.26%)	-0.0352 (0.0%)	-0.0454 (0.0%)	-0.0661 (0.0%)
1/1/16	12/31/17	370	-0.0002 (91.33%)	-0.0028 (44.48%)	-0.0036 (51.72%)	-0.0078 (31.05%)	-0.0055 (55.43%)	0.0013 (90.93%)
1/1/18	12/31/19	250	-0.0033 (17.89%)	-0.0052 (25.51%)	-0.0102 (18.33%)	-0.031 (0.58%)	-0.0582 (0.0%)	-0.0723 (0.0%)

Table 8: LDA - Excess Return relative to Industry

t0	t1	sample size	1d - IDT	10d - IDT	30d - IDT	60d - IDT	90d - IDT	120d - IDT
1/1/00	12/31/05	68	0.0027 (55.54%)	0.0053 (41.86%)	0.0155 (21.51%)	0.025 (13.89%)	0.0687 (0.43%)	0.0836 (0.31%)
1/1/06	12/31/07	112	0.0028 (47.27%)	0.0031 (66.32%)	0.0049 (60.96%)	0.0072 (55.97%)	0.0096 (57.69%)	0.0183 (35.63%)
1/1/08	12/31/09	137	0.0017 (72.79%)	0.0042 (57.88%)	0.0201 (6.78%)	0.0108 (49.62%)	0.0091 (64.62%)	0.0114 (62.76%)
1/1/10	12/31/11	281	0.0021 (16.4%)	0.0005 (86.63%)	-0.004 (36.38%)	-0.0076 (36.29%)	-0.0181 (2.08%)	-0.0168 (9.58%)
1/1/12	12/31/13	314	0.0006 (70.65%)	0.0062 (6.12%)	0.004 (46.22%)	0.0124 (13.29%)	0.0129 (18.21%)	0.0256 (2.76%)
1/1/14	12/31/15	389	0.0005 (81.88%)	-0.0032 (38.22%)	-0.0176 (0.13%)	-0.0317 (0.0%)	-0.0401 (0.0%)	-0.0544 (0.0%)
1/1/16	12/31/17	370	-0.0 (98.62%)	-0.0001 (97.4%)	-0.0023 (64.99%)	-0.004 (58.41%)	-0.0029 (74.47%)	0.0036 (74.5%)
1/1/18	12/31/19	250	-0.003 (20.9%)	-0.0034 (44.09%)	-0.0061 (38.42%)	-0.0213 (3.95%)	-0.0377 (0.23%)	-0.0471 (0.12%)

## References

- David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- Lauren Cohen, Christopher Malloy, and Quoc Nguyen. Lazy prices. *The Journal of Finance*, 75(3):1371–1415, 2020.
- Global Compact. Who cares wins: Connecting financial markets to a changing world. *New York*, 2004.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Misuk Kim, Eunjeong Lucy Park, and Sungzoon Cho. Stock price prediction through sentiment analysis of corporate disclosures using distributed representation. *Intelligent Data Analysis*, 22(6):1395–1413, 2018.
- Alejandro Lopez-Lira. Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*, 2020.
- Marco Lui and Timothy Baldwin. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30, 2012.
- Lasse Heje Pedersen, Shaun Fitzgibbons, and Lukasz Pomorski. Responsible investing: The esg-efficient frontier. *Journal of Financial Economics*, 2020.
- Martin F Porter et al. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.
- Alessandra Sozzi. Measuring sustainability reporting using web scraping and natural language processing, Aug 2017.
- Nichita Utu and Vlad-Sebastian Ionescu. Learning web content extraction with dom features. In *2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 5–11. IEEE, 2018.
- Radim Řehřek and Petr Sojka. Gensim—statistical semantics in python. *Retrieved from genism. org*, 2011.