Topic Modeling with Financial Text Corpora

Zachary Olea

zolea1@jh.edu

Theory of Statistics II

Applied & Computational Mathematics

JHU Whiting School of Engineering EP

May 22, 2025

1 Executive Summary

This paper investigates the application of Latent Dirichlet Allocation (LDA) to analyze textual disclosures from Form 8-K filings of S&P 100 companies. By treating each 8-K filing as a document and the collection of filings as a corpus, LDA was employed to uncover latent topics within these financial texts. The analysis involved preprocessing the data to remove boilerplate language and high-frequency terms, followed by constructing a document-term matrix for topic modeling.

The study revealed significant patterns in the data, such as the prevalence of certain topics across different sectors and the impact of regulatory changes on disclosure practices. Visualizations, including heat maps and hierarchical clustering, were used to illustrate the distribution of topics across sectors and over time. Additionally, dimensionality reduction techniques like t-SNE and UMAP were applied to project the topic distributions into two dimensions, facilitating intuitive exploration of document relationships based on their topic composition.

Overall, the analysis demonstrates the effectiveness of LDA in extracting meaningful insights from complex financial disclosures, highlighting its potential for exploratory data analysis in the financial domain.

2 Project Description

2.1 Project Introduction

This project applies Latent Dirichlet Allocation (LDA) topic modeling to analyze textual disclosures from **Form 8-K** filings submitted by S&P 100 companies. Form 8-K is a report filed with the U.S. Securities and Exchange Commission (SEC) to announce major corporate events that shareholders should know about, such as earnings releases, acquisitions, or leadership changes [1]. LDA assumes that documents consist of various latent topics, and each topic assigns probabilities to words. In this project, each 8-K filing is treated as a document, and the collection of all such filings from the S&P 100 between 2015 and 2025 constitutes the corpus.

LDA is a generative probabilistic model for collections of discrete data, particularly text corpora. LDA is considered generative because topics are not pre-defined; rather, they are discovered automatically during the modeling process. It was introduced by Blei, Ng, and Jordan in 2003 and has since become a foundational method for topic modeling in natural language processing (NLP). The core assumption of LDA is that each document in a corpus is a mixture of topics, where each topic is represented by a distribution over words. The generative process models documents by sampling a topic distribution from a Dirichlet prior, then generating words by repeatedly sampling a topic and then a word conditioned on that topic [2].

2.2 SEC 8-K Parsing and Processing

I began by selecting the domain for my project, ultimately choosing financial data due to the unique challenges it presents for LDA modeling. The complexity of financial disclosures provided an opportunity to explore methods for mitigating known limitations of LDA in noisy, domain-specific corpora. Although the original scope included all S&P 500 companies, computational constraints necessitated narrowing the dataset to the S&P 100. This index comprises the 100 largest and most established firms within the S&P 500 and offers broad sectoral representation across U.S. equities. To further reduce complexity, I restricted the analysis to companies that were current constituents of the S&P 100 during a fixed date range, rather than attempting to model historical index composition—which is difficult to reconstruct due to frequent constituent changes and the variable timing of 8-K filings. The current S&P 100 membership is accessible in WRDS under `comp_na_daily_all.wrds_idx_cst_current`.

To capture intra-industry topic variation, I also collected sectoral index classifications corresponding to each S&P 100 constituent. These sector assignments are particularly useful because each company is mapped to exactly one sector, ensuring a balanced classification scheme. Sector labels are available in the same WRDS dataset and were merged by filtering on the S&P 100 index and joining on `gvkey`, a unique company identifier. For additional clarity and consistency, I parsed the sector names to retain only the core industry label—e.g., “Health Care”, “Consumer Staples”,

or "Financials". Finally, to link these records to SEC filings, I joined Central Index Key (CIK) codes using the `comp.company` dataset on WRDS, which provides a lookup between `gvkey` and `CIK`, the permanent company identifier used by the SEC.

Subsequently, I collected Form 8-K filings—semi-structured XML documents filed with the SEC. These filings typically include boilerplate legal language, attached exhibits, and sectioned items labeled with standard identifiers (e.g., `Item 2.02 Results of Operations and Financial Condition`). The raw XML files are accessible through WRDS's file system, with file paths indexed in `comp_na_daily_all.wrds_idx_cst_current`, which stores SEC form locations by CIK identifier and form type. I used `rdate`, the reporting/publication date, for each 8-K filing to sort and organize documents chronologically. Additionally, I converted each date to the month end date to simplify aggregation across time.

To prepare the data for modeling, I wrote a script to extract and compile the full text of each 8-K filing into a structured `DataFrame`, covering all S&P 100 companies from 2015 to 2025. Within the context of LDA, each **document** corresponds to a single 8-K filing, each **word** is a token parsed from that filing, and each **topic** is a latent distribution over those tokens, inferred via probabilistic topic modeling. The data size is roughly 1 GB per 100 companies per 10 years.

2.3 Preprocessing and Vectorization for LDA

To prepare the data for topic modeling, I developed a custom parser to extract the relevant text from each 8-K filing, focusing specifically on the main body content, `Item` sections, and associated `Exhibit` entries, while discarding irrelevant tags and signature blocks. All residual XML and HTML markup was removed, whitespace was normalized, and special characters were converted to plain text. The cleaned text was then lowercased, stripped of punctuation and numerical digits, and tokenized. I applied natural language preprocessing using NLTK's English stopword list and WordNet lemmatizer—for instance, converting "running" to "run" and "companies" to "company". This pipeline yielded a clean bag-of-words representation appropriate for count-based topic modeling, where word order is not considered.

To further refine the textual corpus and eliminate boilerplate language and overly common terms—such as "section" or "report"—I applied a document frequency filter to the vocabulary. If a word w appears in $d_f(w)$ documents out of a total D, then it is removed if $\frac{d_f(w)}{D} > 0.75$. Effectively, I excluded any term that appeared in more than 75% of the documents.This approach prevents high-frequency, low-information words from dominating the topic modeling process and improves the interpretability of learned topics.

Once the high-frequency terms were pruned, I constructed a document-term matrix $\mathbf{X} \in \mathbb{N}^{D \times V}$, where each entry x_{ij} represents the count of term j in document i, and V is the size of the filtered vocabulary. I then applied the LDA model which assumes each document is a mixture over K latent topics, and each topic is a distribution over words. For each document d, LDA assumes,

$$\theta^{(d)} \sim \text{Dirichlet}(\alpha), \quad z_n^{(d)} \sim \text{Multinomial}(\theta^{(d)}), \quad w_n^{(d)} \sim \text{Multinomial}(\beta_{z_n^{(d)}}) \tag{1}$$

where $\theta^{(d)}$ is the topic distribution for document d, $z_n^{(d)}$ is the topic assignment for the n-th word in the document, and β_k is the word distribution for topic k, drawn from a Dirichlet prior $\beta_k \sim \text{Dirichlet}(\eta)$. In this project, I set $K = 10$ for ease of demonstration, which resulted in ten latent topics.

3 Results

The finalized cleaned data was fed into the model where each row represents one 8-K filing:

monthdate	coname	sector	cleaned_content
2024-01-31	INTEL	IT	Item 2.02 Results of Operations and Financial...
2021-10-31	AMEX	Financials	Item 7.01 Regulation FD Disclosure American Ex...
2019-01-31	GOLDMAN SACHS	Financials	Item 2.02 Results of Operations and Financial ...
2018-04-30	WELLS FARGO	Financials	Item 9.01. Financial Statements and Exhibits E...
2015-06-30	NVIDIA	IT	Item 1.01. Entry into a Material Definitive Ag...

Table 1: Sample of Cleaned 8-K Filings

Note that one 8-K may have multiple items and in a future study should be separated into different rows with any corresponding exhibits included to that item. Next we see the top eight reoccurring words per the ten latent topics. This is to help identify the overall topic and relevance for the words in the topic.

Topic	**Top Words**
Topic 0	note, shall, date, indenture, trustee, payment, prospectus, principal
Topic 1	agreement, shall, section, party, parent, subsidiary, date, transaction
Topic 2	net, income, total, loan, million, loss, expense, asset
Topic 3	net, income, operating, revenue, cash, nongaap, quarter, tax
Topic 4	note, offer, llc, subsidiary, time, date, restricted, indebtedness
Topic 5	plan, share, award, stock, shall, agreement, date, participant
Topic 6	index, day, calculation, time, agent, date, trading, indenture
Topic 7	million, cost, net, customer, energy, asset, december, tax
Topic 8	lender, shall, agent, borrower, section, administrative, loan, agreement
Topic 9	shall, corporation, director, meeting, stock, share, board, section

Table 2: Top Words for Each LDA Topic

We can see that each topic still contains a fair amount of boilerplate financial jargon; however,

let us assume that these words are significant contributors to the topical relevance for each 8-K. Next, I plotted a heat map of the topic distributions over each sector to see if any particular sector has higher percentages of any topic and conversely if any topic is prevalently only from one sector.

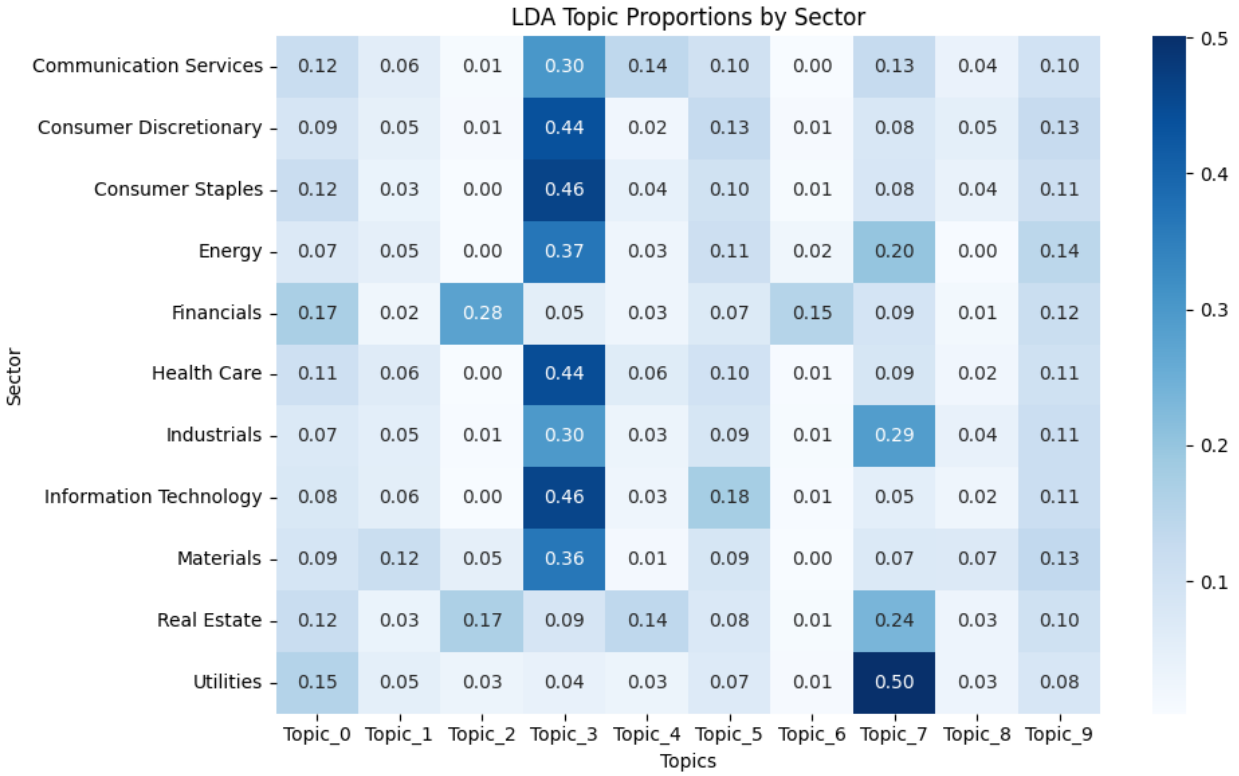

Figure 1: Topic Distribution Heat Map

Some topics, like `Topic 9` are very evenly distributed across each sector, indicating that this is most likely regulatory reporting that is recurring and mandatory, the word distribution for this topic indicates that it relates to ownership and is most likely the `Item 5` filings which are about executive stock compensation. On the other hand, `Topic 7`, which refers to energy costs, is predominantly related to the Utilities sector with Industrials and Real Estate having some significance as well.

We can see that `Topic 3` is the most prevalent in all sectors except for Financials, Real Estate and Utilities.This seems to be a bit odd as this topic relates to financial statement jargon and particularly "Non-GAAP" reporting, which is reporting that is not audited (GAAP) in the same standard as normal quarterly and annual financial statements are. Non-GAAP reporting is often

used as a metric for investors to understand the true health of a firm, and this heatmap is showing that this type of reporting is not common for these three areas.

We observe that some sectors and some topics have matching patterns in this distribution we can quantify through hierarchical clustering.

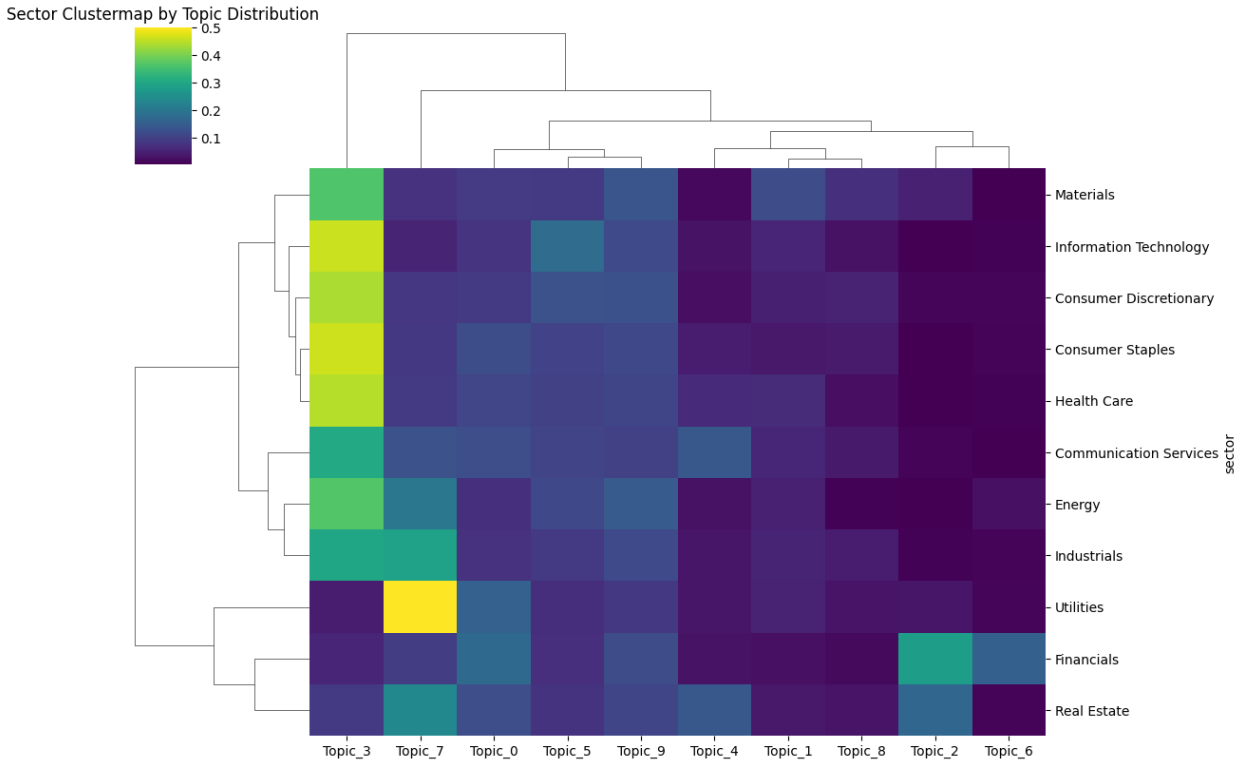

Figure 2: Hierarchy

Here we can see the relations between each topic and each sector more clearly. For instance, `Topic 3` stands apart, but `Topic 1 and Topic 8` are closely related. Additionally, Real Estate, Financials, and Utilities share topical similarity distinctly different than other sectors.

However, this heat map does not show the distributions over time, which is a feature of this dataset. Since all topics distributions will add to 1, I plot the topic distributions over time stacked.

Figure 3: Topics Over Time by Sector

From this, we can observe that there is some amount of seasonality in the topics, most likely due to filings that are mandatory on a fixed basis. An interesting observation that stands out is the disappearance of `Topic 6` around the end of 2020.

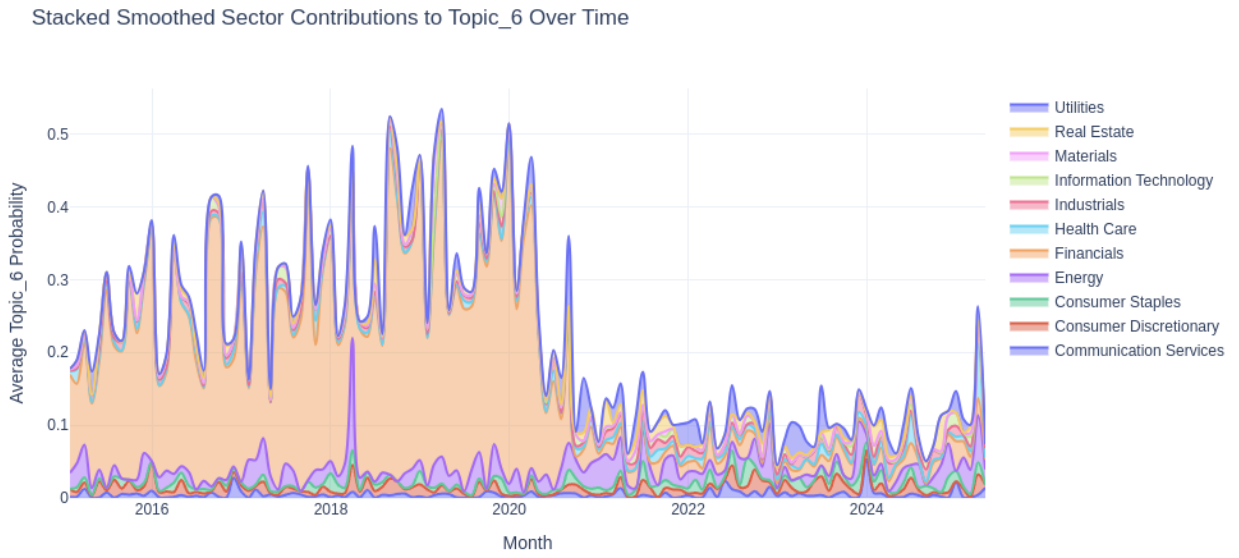

Figure 4: Sector Distribution for Topic 6

Isolating `Topic 6`, we can see that there is a structural break in the time series for just the Financials sector.

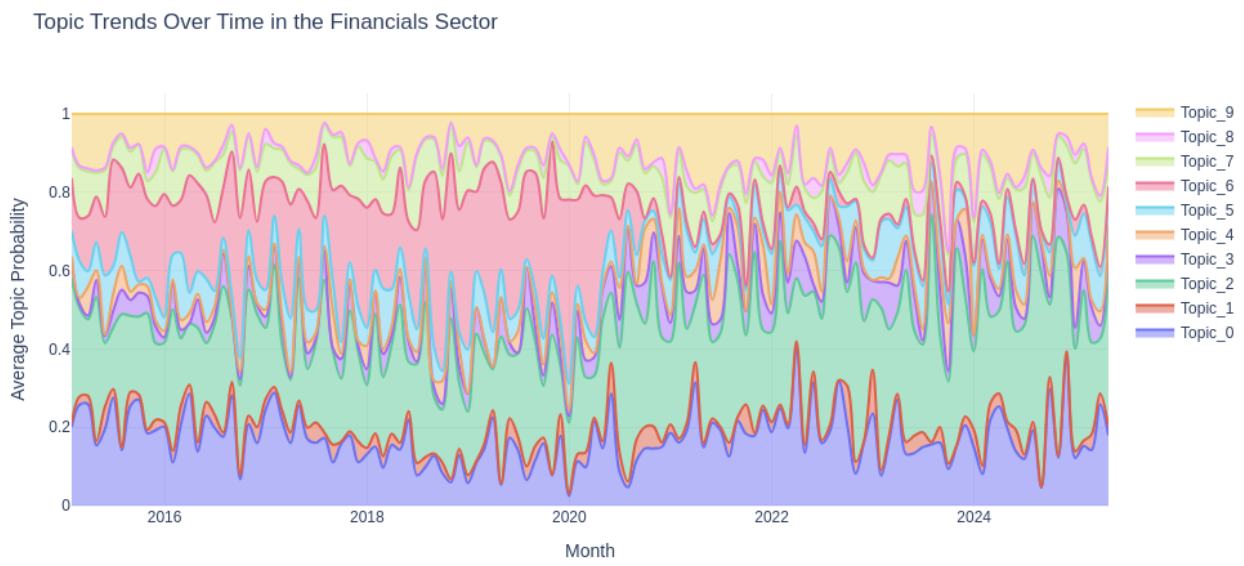

Figure 5: Topic Distribution for Financial Sector

Additionally, when examining all topics for the Financial sector, we observe that only Topic 6 exhibits a notable change during this period. Accounting researchers are often interested in how regulatory changes impact disclosure practices, as these shifts can serve as natural experiments for measuring the effects of new rules. In this case, a review of SEC regulatory activity reveals that, effective November 9, 2020, significant amendments to Regulation S-K were implemented. Most notably, Item 101 was revised to eliminate the requirement for companies to disclose general business developments over specific time frames (previously the past three or five years). Instead, companies are now only required to report material developments, allowing for more concise and tailored disclosures.

In addition to this change, amendments to Items 105, 103, and 303 further streamlined reporting requirements. Item 105 now requires disclosure only of material risk factors, organized for clarity; Item 103 raises the threshold for reporting certain legal proceedings and allows for cross-referencing to avoid duplication; and Item 303 (Management's Discussion and Analysis) was modernized to focus on material trends and uncertainties, eliminating some prescriptive requirements such as selected financial data and detailed discussions of inflation unless material. Collectively, these changes reduced the volume and specificity of required disclosures, which likely contributed

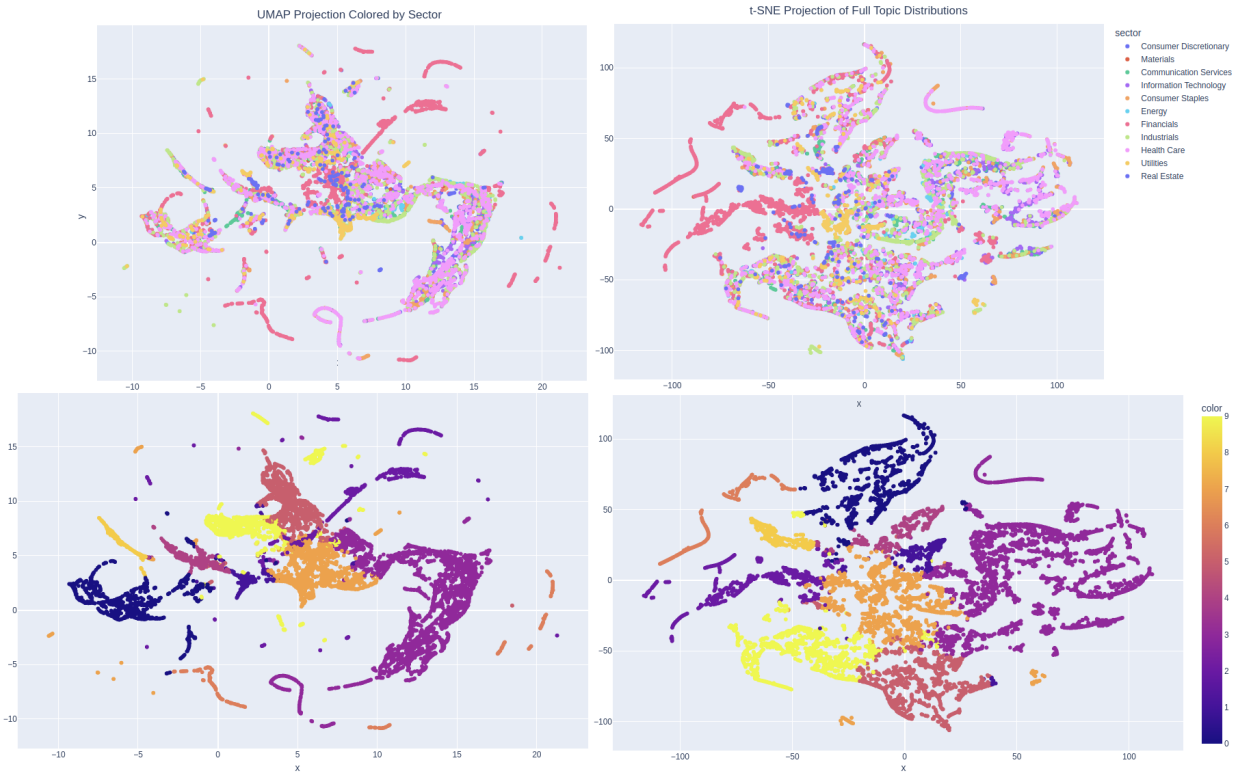

Figure 6: UMAP & t-SNE

to the observed decline in Topic 6 for financial sector companies during this period.

Topic distributions by each sector and sector distributions for each topic are available in the appendix:7, 8.

To visualize the high-dimensional topic distributions generated by LDA, I applied dimensionality reduction techniques, t-SNE and UMAP, to project the document-topic vectors into two dimensions for visualization. This allows for intuitive exploration of the relationships between documents based on their topic composition. Since each document is a mixture of topic, ideally we would use a blended colors to represent topic strength (fuzzy coloring); however, for simplicity I plotted the predominant topic per document. Here we can see clear formation of tails, clusters, islands, bridges, and other clustering shapes. We can see that some of the isolation in groups can be explained by the predominant topic distribution as well as with the sector. We can also observe that by taking predominant topic alone, we can get a fair amount of separation for clustering with just the 10 topics; however, testing with higher topic selection may result in better fit.

4 Methods

4.1 Latent Dirichlet Allocation: Model Architecture

Diving deeper into LDA's architecture, for each document d, a topic distribution $\theta_d \sim \text{Dir}(\alpha)$ is sampled, representing the topic proportions for that document. For each word position n in the document,

- A topic assignment $z_{dn} \sim \text{Multinomial}(\theta_d)$ is chosen.
- A word $w_{dn} \sim \text{Multinomial}(\phi_{z_{dn}})$ is sampled, where each topic's word distribution $\phi_k \sim \text{Dir}(\beta)$.

Let $\mathbf{w}_d = (w_{d1}, \ldots, w_{dN_d})$ denote the sequence of words in the d-th 8-K filing, where N_d is the number of words in document d. The latent variables in LDA are,

- θ_d: topic distribution for document d.
- z_{dn}: topic assignment for word n in document d.
- ϕ_k: word distribution for topic k.

Given corpus-level hyperparameters α and β, the joint distribution over topic proportions, topic assignments, and observed words in a single document is,

$$p(\theta_d, \mathbf{z}_d, \mathbf{w}_d \mid \alpha, \beta) = p(\theta_d \mid \alpha) \prod_{n=1}^{N_d} p(z_{dn} \mid \theta_d)\, p(w_{dn} \mid z_{dn}, \beta). \tag{2}$$

[2]

The marginal likelihood of the observed words in a document is obtained by integrating out the latent variables,

$$p(\mathbf{w}_d \mid \alpha, \beta) = \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d)\, p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d. \tag{3}$$

[2]

In this setting, each **document** corresponds to an individual 8-K filing, each **word** w_{dn} is a token from the cleaned text, and the **topics** are multinomial distributions over words, shared across all filings. Since exact inference of the posterior is intractable, approximate methods are used. The `scikit-learn` implementation employs online variational Bayes to estimate the topic-word distributions ϕ_k and document-topic proportions θ_d from the data. By fitting the LDA model to the parsed and lemmatized 8-K corpus, we uncover a set of interpretable topics representing latent themes in regulatory disclosures, which can be analyzed further across sectors and over time.

4.2 Dimension Reduction

Let $\Theta \in \mathbb{R}^{N \times K}$ be the document-topic matrix, where N is the number of documents and K is the number of topics. Each row θ_d is a probability vector for document d. For each document $d \in D$, LDA produces a topic distribution vector,

$$\theta d = [\theta d, 1, \theta_{d,2}, \ldots, \theta_{d,K}]$$

where $\theta_{d,k}$ is the probability of topic k in document d, and $\sum_{k=1}^{K} \theta_{d,k} = 1$. Combining these vectors into a matrix we have,

$$\Theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{|D|} \end{bmatrix} \in \mathbb{R}^{|D| \times K}$$

Next, I employed dimension reduction techniques to view the topic modelings ability to cluster. UMAP finds a low-dimensional embedding $\mathbf{Y} \in \mathbb{R}^{|D| \times 2}$ that preserves the local structure of the data by minimizing the cross-entropy between high-dimensional and low-dimensional fuzzy simplicial sets. t-SNE computes pairwise similarities between documents in the high-dimensional topic space and seeks a 2D embedding that preserves these similarities by minimizing the Kullback-Leibler

divergence between the high- and low-dimensional distributions.

4.2.1 UMAP

For each document i, we compute the local connectivity using Euclidian distance in the topic space. We define a fuzzy simplicial set (weighted graph) where the edge weight between i and j is,

$$w_{ij} = \exp\left(-\frac{\max(0, \|\theta_i - \theta_j\| - \rho_i)}{\sigma_i}\right)$$

where ρ_i is the distance to the nearest neighbor and σ_i is a normalization factor. In the low-dimensional space, we define a similar fuzzy set with edge weights,

$$w'_{ij} = \left(1 + a\|\mathbf{y}_i - \mathbf{y}_j\|^{2b}\right)^{-1}$$

where a, b are hyperparameters controlling the embedding. UMAP optimizes the low-dimensional embeddings $\{\mathbf{y}_i\}$ by minimizing the cross-entropy between the high- and low-dimensional fuzzy sets:

$$C = \sum_{i<j}\left[w_{ij}\log\frac{w_{ij}}{w'ij} + (1 - wij)\log\frac{1 - w_{ij}}{1 - w'_{ij}}\right]$$

[3]

4.2.2 t-SNE

For each pair of documents i, j, we define the conditional probability that document j is a neighbor of i as,

$$p_{j|i} = \frac{\exp\left(-\|\theta i - \theta_j\|^2/2\sigma_i^2\right)}{\sum k \neq i \exp\left(-\|\theta_i - \theta_k\|^2/2\sigma_i^2\right)}$$

where σ_i is set so that the perplexity of the distribution equals a user-specified value. The joint probability is,

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

Let $\mathbf{y}_i, \mathbf{y}_j \in \mathbb{R}^2$ be the 2D embeddings. We define,

$$q_{ij} = \frac{(1 + \|\mathbf{y}i - \mathbf{y}_j\|^2)^{-1}}{\sum k \neq l (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

t-SNE finds the embedding $\{\mathbf{y}_i\}$ that minimizes the Kullback-Leibler divergence between the high- and low-dimensional distributions,

$$\mathrm{KL}(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

[4]

5 Conclusions

In this study, I applied LDA to analyze textual disclosures from Form 8-K filings of S&P 100 companies. The results demonstrate that LDA is a powerful tool for exploratory data analysis and uncovering hidden patterns in large text corpora. Despite the legal and financial boilerplate jargon prevalent in 8-K filings, LDA successfully extracted meaningful topics, such as regulatory changes, which are crucial for understanding corporate disclosures.

However, several challenges were encountered during the analysis. The primary difficulty was the inherent complexity and noise within financial disclosures. While LDA provided a robust framework for topic modeling, it did not inherently process sectoral and temporal dimensions of the data. Rather, these were only used in analysis, post-LDA. Incorporating these aspects into the training process would require more sophisticated and modified versions of LDA, which could potentially yield more nuanced insights.

Future analyses could benefit from using GENSIM, a library that offers additional capabilities for topic modeling. GENSIM's implementation of LDA is highly efficient and allows for greater flexibility in model tuning and optimization. Additionally, exploring alternative methods such as word embeddings and large language models (LLMs) could address some limitations of pure sta-

tistical LDA. For instance, LDA does not account for semantic similarity, terms like ”merger” and ”acquisition” may be treated as unrelated, despite their contextual similarity in financial texts. Similarly, LDA does not capture sentence or document-level context, which can lead to misinterpretations. For example, ”self-driving” might not be recognized as part of ”artificial intelligence” unless explicitly mentioned.

Using advanced techniques like BERT-based topic modeling (BERTopic) can incorporate semantic similarity and contextual information, providing a more comprehensive understanding of the data. These methods leverage pre-trained language models to capture deeper linguistic patterns, making them well-suited for complex financial texts.

While LDA has proven effective for initial EDA and topic discovery, integrating more advanced models and techniques will enhance the analysis of financial disclosures, offering more nuanced insihgts.

6 Appendices

6.1 Additional Plots

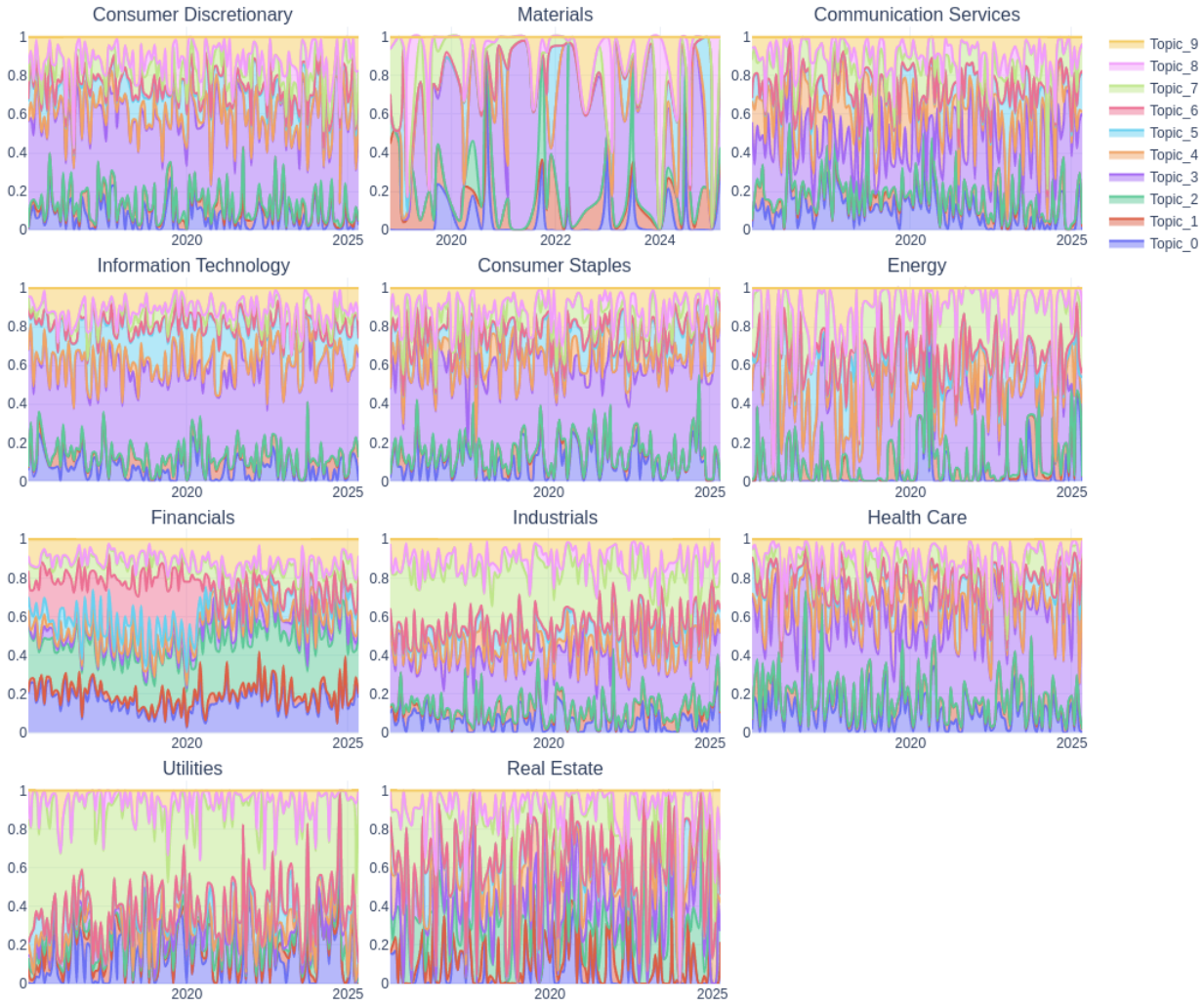

Figure 7: Sectors Over Time by Topic

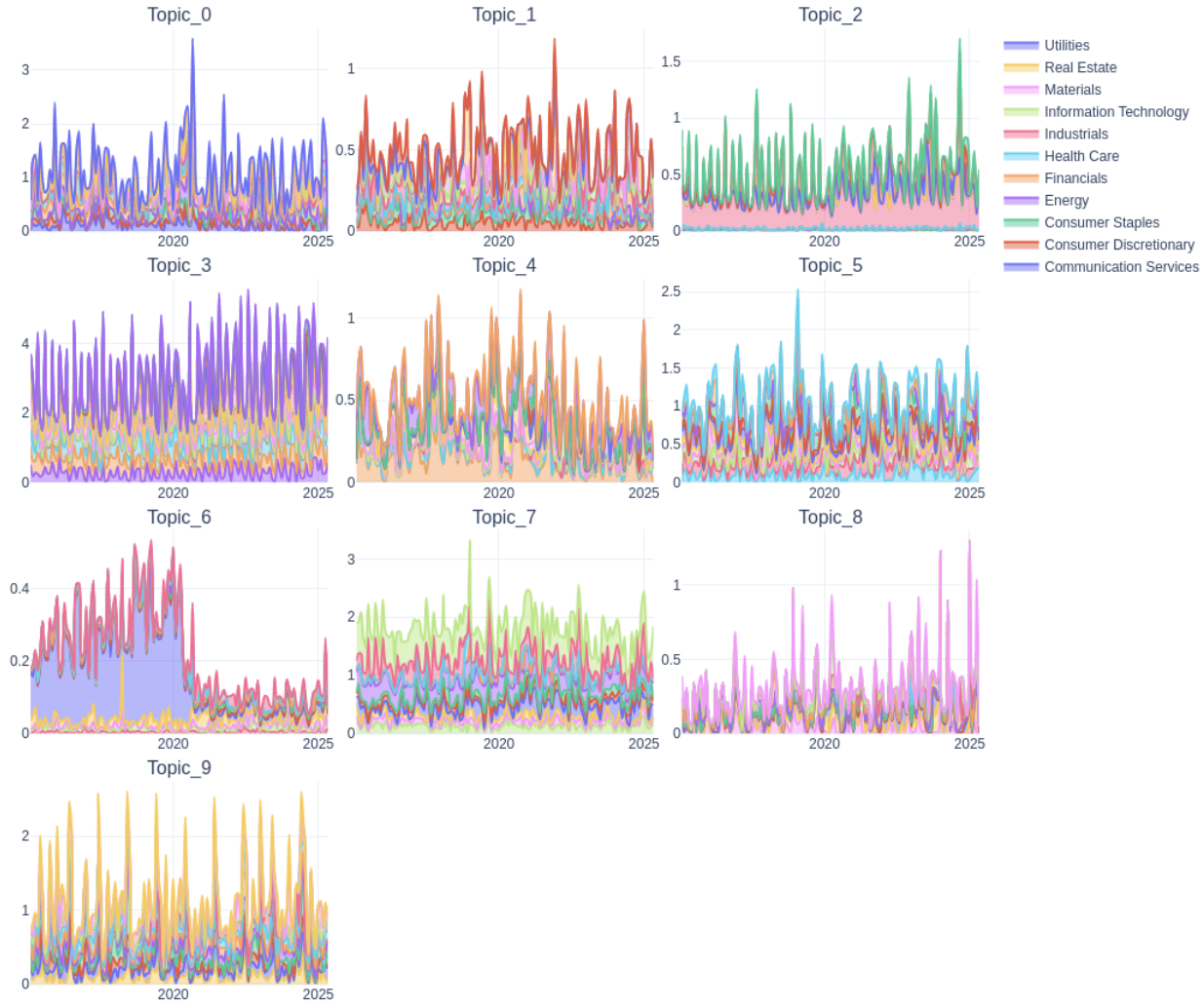

Figure 8: Topics Over Time by Sector

6.2 8-K Example (reduced)

```
<SEC-DOCUMENT>0001193125-24-154299.txt : 20240604
<SEC-HEADER>0001193125-24-154299.hdr.sgml : 20240604
<ACCEPTANCE-DATETIME>20240604162955
ACCESSION NUMBER:  0001193125-24-154299
CONFORMED SUBMISSION TYPE: 8-K
PUBLIC DOCUMENT COUNT:  26
CONFORMED PERIOD OF REPORT: 20240604
ITEM INFORMATION:  Other Events
ITEM INFORMATION:  Financial Statements and Exhibits
FILED AS OF DATE:  20240604
DATE AS OF CHANGE:  20240604

FILER:
```

```
 COMPANY DATA:
  COMPANY CONFORMED NAME:   LINDE PLC
  CENTRAL INDEX KEY:   0001707925
  STANDARD INDUSTRIAL CLASSIFICATION: INDUSTRIAL INORGANIC CHEMICALS [2810]
  ORGANIZATION NAME:            08 Industrial Applications and Services
  IRS NUMBER:    000000000
  STATE OF INCORPORATION:   L2
  FISCAL YEAR END:   1231

</SEC-HEADER>
<DOCUMENT>
<TYPE>8-K
<SEQUENCE>1
<FILENAME>d815371d8k.htm
<DESCRIPTION>8-K
<TEXT>
xml version='1.0' encoding='ASCII'?8-K
```

6.3 Bag-of-Words (reduced)

item result operation financial condition february tesla motor inc released financial result quarter fiscal year ended december posting fourth quarter full year shareholder letter website full text shareholder letter attached hereto exhibit incorporated herein reference information intended furnished item form result operation financial condition shall deemed filed purpose section security exchange act amended exchange act incorporated reference filing security act amended exchange act except shall expressly set forth specific reference filing item financial statement exhibit exhibit exhibit description tesla motor inc fourth quarter full year shareholder letter dated february exhibit exhibit tesla motor fourth quarter full year shareholder letter introduced allwheel drive dual motor model autopilot record quarterly production vehicle delivered vehicle pd production delay pushed expanded supercharger network cover u coast coast europe expecting growth vehicle delivery model begin shipping six month february dear fellow shareholder tesla continues drive global transition sustainable transport pure electric car revenue company combined tesla nongaap revenue grew almost gaap revenue grew almost gross margin simultaneously increased unusually high level automotive standard moreover implied graph vehicle production demand expected accelerate also increased number store service center expanded supercharger network started construction gigafactory introduced numerous advance model result progress entered order model almost reservation model built vehicle thus achieving production target model vehicle required herculean effort held back release performance allwheel drive dual motor car pd ensure would truly great experience owner able recover lost production end quarter delivering car physically impossible due combination customer vacation severe winter weather shipping problem actual ship result vehicle slipped december delivered financials reflect delivery shortfall onetime manufacturing inefficiency related introduction pd autopilot functionality impact strong dollar even though dollar continued strengthen versus euro believe able sustain nongaap automotive gross margin stabilizing production improve manufacturing efficiency without sharp increase dollar strength gross margin would tesla serf international market global supply chain model built north america strong dollar slightly negative net effect profitability vehicle enhancement market expansion launched pd november enthusiastic reception automotive reporter car buyer motor trend called model pd quickest sedan world youtube video thrilled passenger went viral customer responded placing record number model order quarter level order likely reflects

6.4 Data & Code

Full data and code will be made available on my github at: https://github.com/takeru240/

6.4.1 Code Excerpts

```
def parse_8k_content(content):
    cleaned_parts = []
    type_blocks = re.split(r'(?=<TYPE>)', content, flags=re.IGNORECASE)
    for block in type_blocks:
        type_match = re.search(r'<TYPE>(.*?)\s', block, flags=re.IGNORECASE)
        if not type_match:
            continue
        type_name = type_match.group(1).strip()
        text_match = re.search(r'<TEXT>(.*?)</TEXT>', block, flags=re.IGNORECASE
        ↪  | re.DOTALL)
        if not text_match:
            continue
        text = text_match.group(1).strip()
        if type_name.lower() == '8-k':
            item_sections = re.split(r'\b(Item\s+\d[\.\d]*)', text,
            ↪  flags=re.IGNORECASE)
            if len(item_sections) > 1:
                for i in range(1, len(item_sections), 2):
                    item_header = item_sections[i].strip()
                    item_text = item_sections[i+1]
                    item_text = re.split(r'Signature', item_text,
                    ↪  flags=re.IGNORECASE)[0]
                    cleaned_parts.append(item_header + " " + item_text)
            else:
                text = re.split(r'Signature', text, flags=re.IGNORECASE)[0]
                cleaned_parts.append(text)
        elif type_name.upper().startswith('EX-'):
            exhibit_part = type_name[3:].strip()
            if re.match(r'^\d+(\.\d+)?$', exhibit_part):
                cleaned_parts.append(text)
            else:
                continue
    return '\n\n'.join(cleaned_parts)

def basic_text_cleaning(text):
    text = re.sub(r'<[^>]+>', ' ', text)
    text = re.sub(r' |&amp;|&lt;|&gt;', ' ', text)
    text = re.sub(r'\s+', ' ', text)
    text = text.strip()

    return text
```

```python
def clean_text(text):
    if not isinstance(text, str):
        return ''
    text = text.lower()
    text = text.translate(str.maketrans('', '', string.punctuation))
    text = re.sub(r'\d+', '', text)
    words = re.findall(r'\b[a-z]{2,}\b', text)
    words = [lemmatizer.lemmatize(word) for word in words if word not in stop_words]
    cleaned_text = ' '.join(words)
    return cleaned_text
df_8k['lda_cleaned_content'] = df_8k['cleaned_content'].apply(clean_text)

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
vectorizer = CountVectorizer(
    stop_words='english',
    lowercase=True,
    max_df=0.75,
    )
doc_term_matrix = vectorizer.fit_transform(df_8k['lda_cleaned_content'])
lda = LatentDirichletAllocation(n_components=10, random_state=0)
lda.fit(doc_term_matrix)

feature_names = vectorizer.get_feature_names_out()
def display_topics(model, feature_names, no_top_words):
    for topic_idx, topic in enumerate(model.components_):
        print(f"Topic {topic_idx}: ", end="")
        print(", ".join([feature_names[i] for i in topic.argsort()[:-no_top_words - 1:-1]]))
display_topics(lda, feature_names, no_top_words=8)

doc_topic_dist = lda.transform(doc_term_matrix)
topic_columns = [f"Topic_{i}" for i in range(lda.n_components)]
df_8k_topics = pd.DataFrame(doc_topic_dist, columns=topic_columns)
df_8k_topics = pd.concat([df_8k[['sector','monthdate', 'cik', 'coname']].reset_index(drop=True), df_8k_topics], axis=1)
df_8k_topics.sample(5)
```

References

[1] U.S. Securities and Exchange Commission, *Form 8-K: Current Report*, https://www.sec.gov/fast-answers/answersform8khtm.html, Accessed: 2025-05-06.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018. [Online]. Available: https://arxiv.org/abs/1802.03426.

[4] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.