Hierarchical Topic Modeling over Financial Documents

Ulises Hernandez [1] Gilberto Garcia [1] Abel Perez [1] Nicolo Ricca [1] Xinyu Wang [1] Yunchen Yao [1]
Simerjot Kaur [2] Keshav Ramani [2] Akshat Gupta [2]

[1]Columbia University Data Science Institute [2]JP Morgan Chase & Co.

COLUMBIA UNIVERSITY
DATA SCIENCE INSTITUTE

JPMORGAN CHASE & CO.

Goal

The objective of this project is to leverage unsupervised learning models to unveil the structure of the public Enron email dataset. We focus on two tasks: hierarchical topic modeling and topic evolution in email threads. To this end, term frequency models like LDA, word embedding approaches like BERTopic, and an ensemble of the two are employed.

Model Frameworks

- **LDA.** Latent Dirichlet Allocation is a mixed membership model that assumes independence between documents in a corpus and models each document as a mixture of topics. A topic is defined as a distribution over terms in the corpus vocabulary. The model specifies a generative process in which the topics and topic proportions are sampled from Dirichlet distributions, and the latent topic assignments and observed words are sampled from a categorical variables.
- **BERTopic.** This hierarchical model generates topic representations in three steps. First, each document is transformed to an embedding representation using a pre-trained language model. Then, the dimensionality of the resulting embeddings is reduced and subsequently clustered. Lastly, topic representations are extracted from the clusters using a custom class-based variation of TF-IDF.
- **Combination of LDA and BERTopic.** From the text of each email, the probabilistic topic assignment vector is extracted via LDA, and the sentence embedding vector through BERT. Afterwards, these two vectors are concatenated resulting in a high-dimensional vector where information is sparse and correlated. These vectors are treated as the new representation of the emails and are used as input for the BERTopic model to obtain the topics.

Significant data cleaning was performed to transform the original data into suitable input formats, which included removal of stopwords and very frequent words. It was found that the interpretability of the topics strongly depends on the data cleaning process: a bag of words approach for LDA, and embeddings approach for BERTopic.

Hierarchical Topic Modeling

Cross validation was used for hyperparameter tuning in all models. As a result, seven topics were selected for LDA, and uni-grams with bi-grams were used for BERTopic and LDA-BERTopic models, together with HDBSCAN clustering for topic construction. The following figures show the result of LDA as an example. Figure 1 shows the top words in each topics and Figure 2 displays the hierarchical structure of topics.

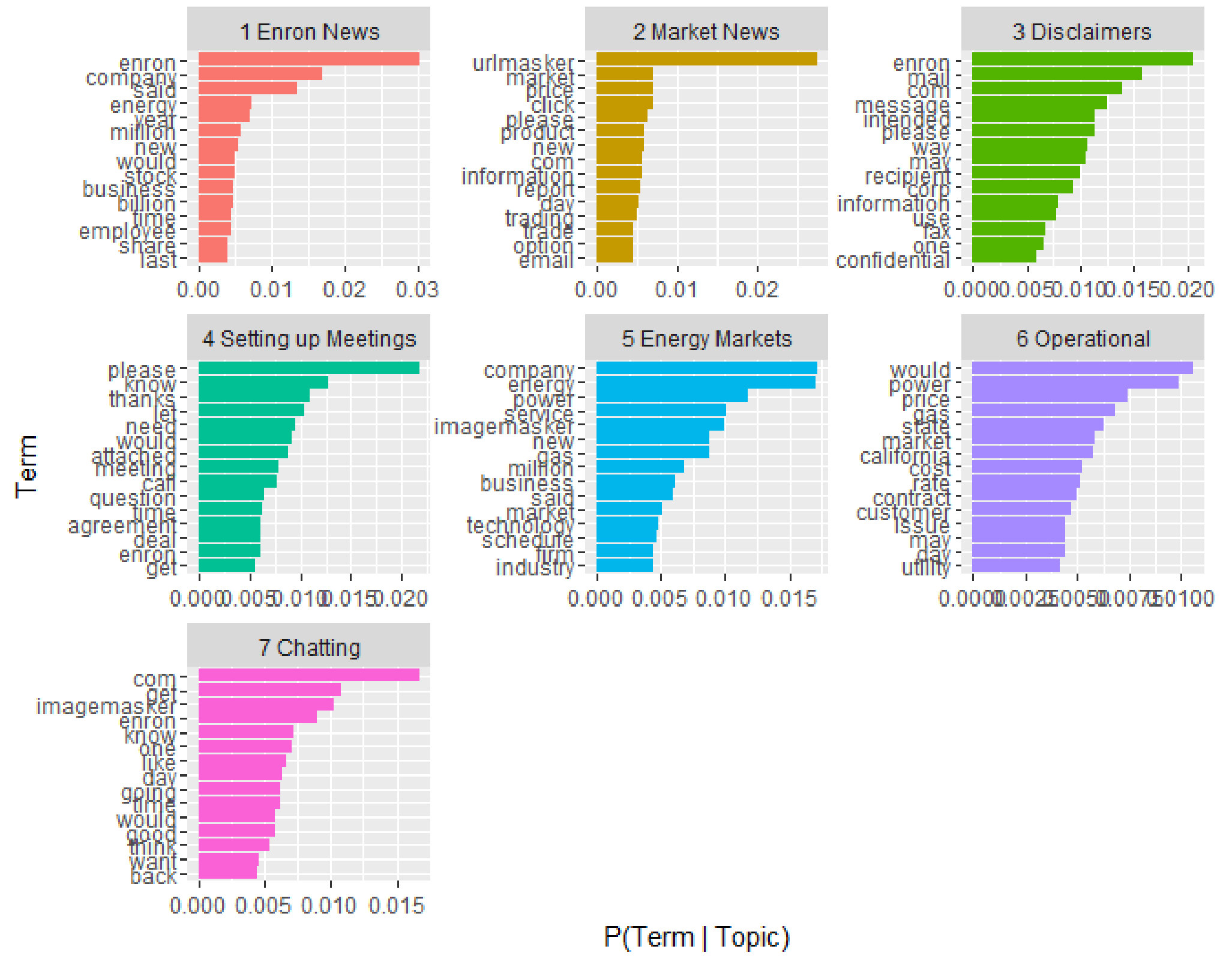


Figure 1. Top words of Topics in LDA

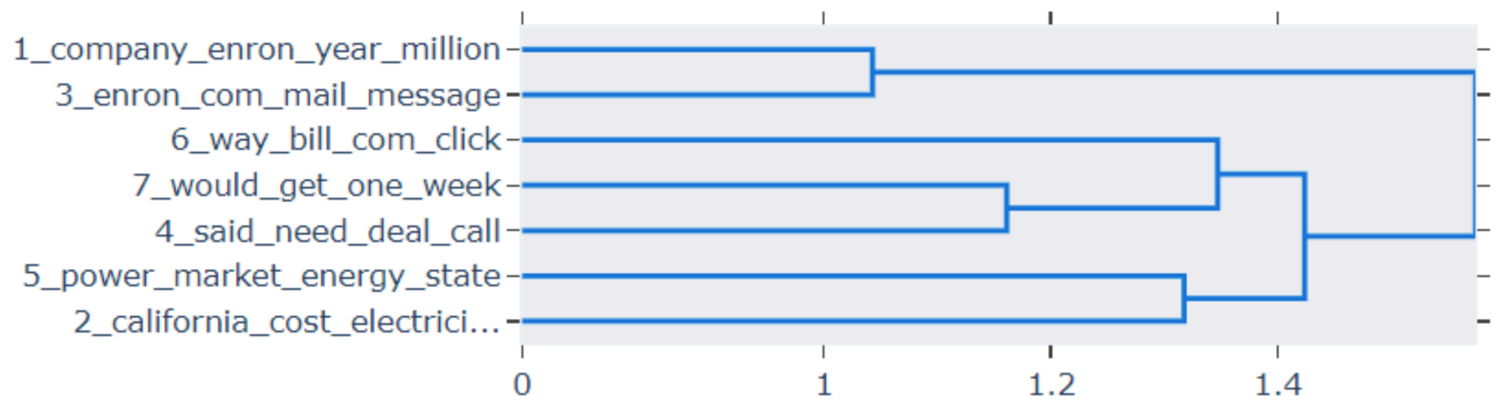


Figure 2. Hierarchical Structure of topics in LDA

Metrics	LDA	BERTopic	LDA+BERTopic
Coherence	**0.5623**	0.5242	0.4164
Diversity	0.8000	**0.9313**	0.8498

Table 1. Evaluation of three topic models

Two metrics, coherence and diversity, were used to evaluate the models. Coherence measures how similar the topmost words in a topic are to each other, while diversity displays how different the resulting topics are. The Table 1 shows the comparison of three models.

Topic Evolution Models

About 64% of emails in the data set are members of a thread. This serves as motivation for analyzing the overall topic **evolution** for an average email thread, as well as how topics **change** within a particular thread. LDA was selected for this task because it provides better interpretability.

- **Overall topic evolution:** LDA topic proportions were averaged among all emails in a common thread position. These averages allow us to see a pattern in email content as thread position increases.
- **Topic change within threads:** For each thread, smoothed topic proportions are computed as a weighted average of current and previous proportions. Topic with highest smoothed proportion is assigned to each email allowing to count the transitions between topics in threads.

Topic Evolution results

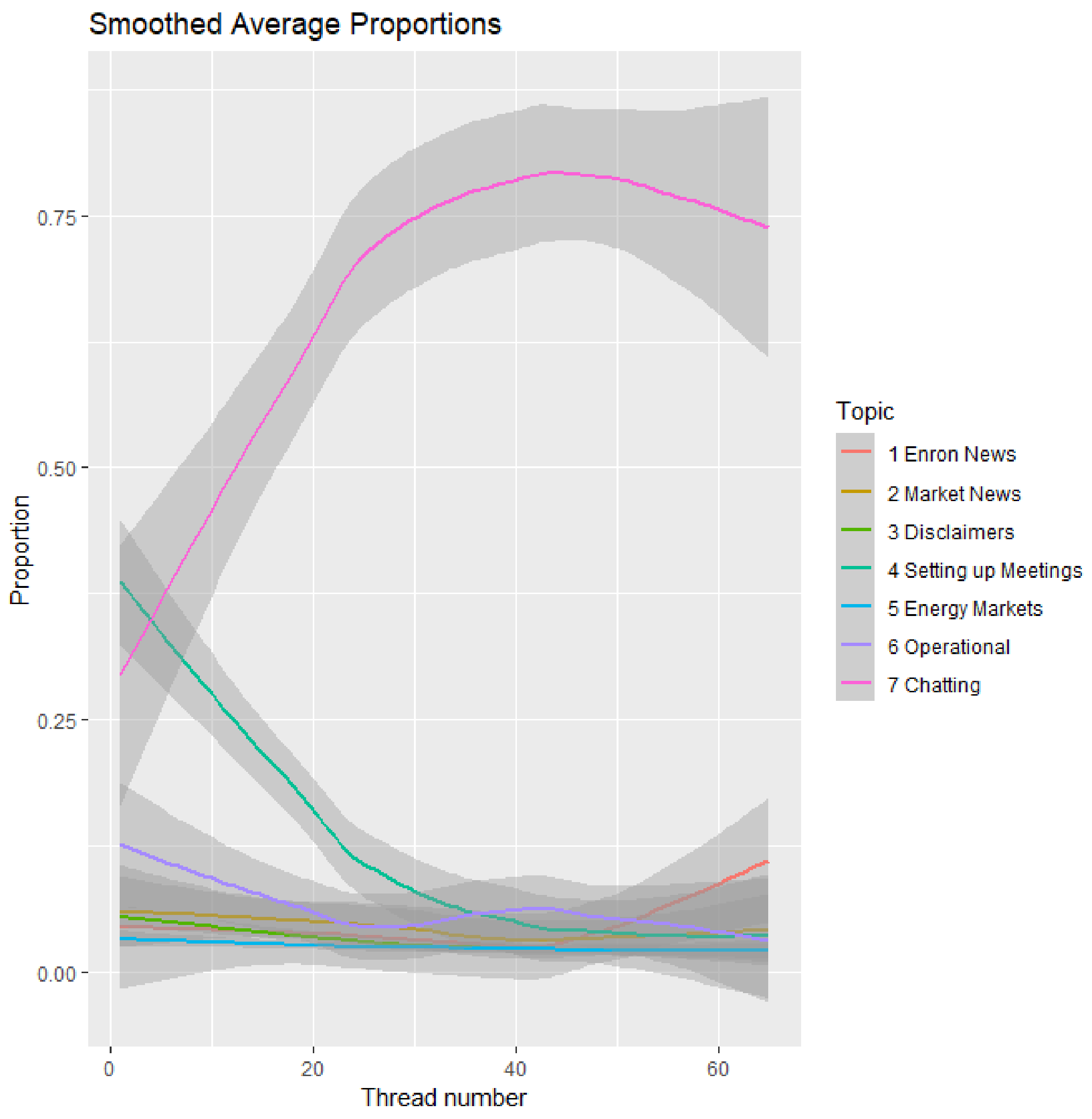
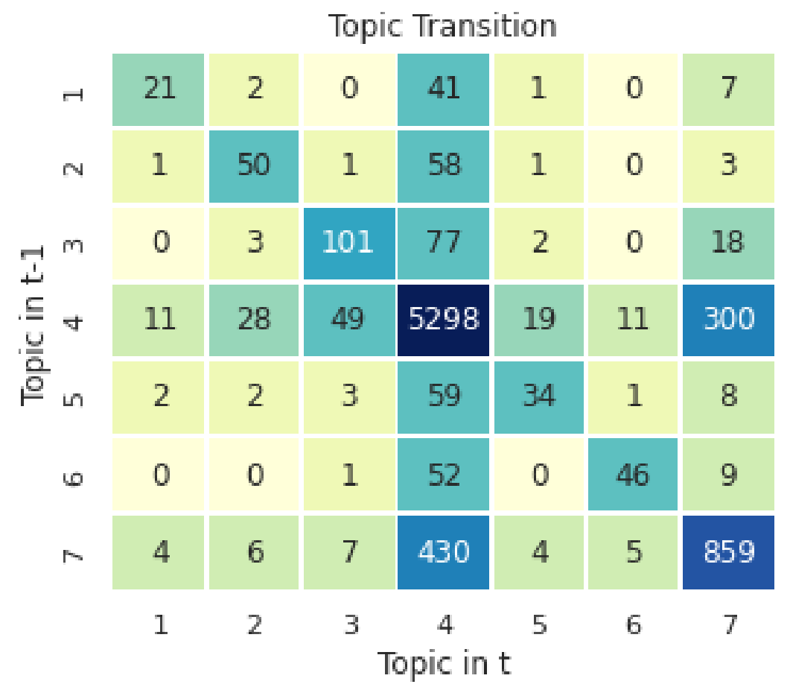
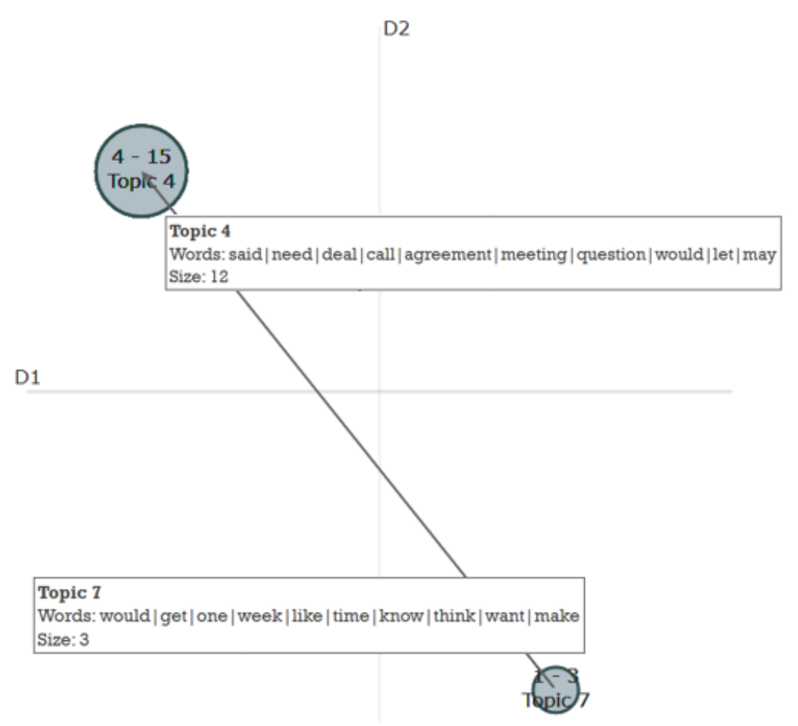


Figure 3. Overall Topic Evolution



	1	2	3	4	5	6	7
1	21	2	0	41	1	0	7
2	1	50	1	58	1	0	3
3	0	3	101	77	2	0	18
4	11	28	49	5298	19	11	300
5	2	2	3	59	34	1	8
6	0	0	1	52	0	46	9
7	4	6	7	430	4	5	859

Figure 4. Topic Transition



Topic 4
Words: said | need | deal | call | agreement | meeting | question | would | let | may
Size: 12

Topic 7
Words: would | get | one | week | like | time | know | think | want | make
Size: 3

Figure 5. Topic change within a thread

Figure 3 shows topic proportions evolution as thread length increases, revealing a movement from formal to informal topics. Figure 4 shows the frequency of the topics at time t given the topics at time $t-1$. Figure 5 shows how the topic evolves within a thread. Every circle represents a topic, the size of the circle represents the frequency of that topic and the distance between circles shows the topic similarity. The arrow can tell how the topic evolves, and the text in the circle indicates the topic and threads that belong to this topic.

Further Extensions

- **Hierarchical LDA (HLDA).** Using the same principles as LDA, this model tries to construct topic trees on multiple levels using the concept of the *nested Chinese restaurant process*. Given the high complexity of the model, it is expected that considerable running time and hardware resources are required.
- **Alternative combination procedure.** Given the success of LDA for modeling topic evolution, implementing a continuous mixed-membership model using word embeddings, derived from a BERTopic model, is an alternative way to combine the LDA and BERTopic frameworks.

https://github.com/UlisesHdzG/Hierarchical-topic-modeling-on-financial-docs

Columbia University - DSI Capstone Project 2022, New York