Georgia Tech VIP:
Unlocking and Analyzing Historical Texts

Topic Modelling Team

Spring 2025 Summary Report

Team Members:

Alina Lee

Myungjin Shin

Parag S Ambildhuke

Shubahm Dhar

Taehwan Park

How to run

- Repository: https://github.com/VIP-TopicModeling/Topic-Modeling-VIP-EEBO-TCP-Collections-Navigations
- Wiki: https://github.com/VIP-TopicModeling/Topic-Modeling-VIP-EEBO-TCP-Collections-Navigations/wiki
- Algorithms Document (continually updated LaTeX document; email mshin90@gatech.edu for access): https://github.com/VIP-TopicModeling/Topic-Modeling-VIP-EEBO-TCP-Collections-Navigations/blob/master/docs/algorithms/Algorithms.pdf
- Presentation, progress slides from last semester: Slides

How to run

- I recommend perusing all the wiki sections in order, but in short:
  1. Clone the repo, and install required packages as listed here.
  2. The historic documents are stored in XML formats in our repo. In order to build the topic models they must first be preprocessed, separating the text content and metadata. Run commands listed here.
  3. Currently, the topic models we use (primarily pLSI) makes use of the word document matrix as well as TF-IDF. Generate them using commands listed here.

How to run

4. Train the pLSI topic model. To easily track the training process, it is recommended to use Weights & Biases. While you can run `wandb login` command (guide), you can also set the `WANDB_API_KEY` environment variable for automatic log-in, by exporting it or setting it in the .env file in the root of the project.

Then, run the following command to train the pLSI model with 20 topics for 120 iterations:

```
python plsi.py --topics 20 --input_dir out/vectors --output_dir
vectors_in_csv/plsi_vectors --max_iter 120 --tol 1e-5 --pct_docs 100 --
matrix_type tfidf_reweighted_count_vectors –use_wandb
```

Read the wiki entry for details on the training arguments.

Given the computational resource to train the model (may take several hours with above setting), you should use Georgia Tech's PACE-ICE compute; see here.

You may also use these training results from last year: 20250409_vocab100k.

How to run

5. Analyse the results. There are three main programs you can use for result analysis/visualization.

    1) visualization.py (wiki entry)

This script generates a detailed tables and graphs providing an overview of the results consisting of the following and more (from wiki):

- **Table 1: Document Count Per Topic**
  For each topic, the table displays the number of documents that have the highest probability for that topic (from P_dz), along with the corresponding percentage of the total documents.
- **Table 2: Top Words Per Topic (With Average Value)**
  For each topic, the table lists the top N words (default is 20) from P_zw sorted in descending order by value. It also shows the average row value for that topic and, for each word, the rank, word, the value, and the percentage that word contributes to the total for that topic.
- **Table 3: Top 5 Documents Per Topic & Their Topic's Best Words**
  For every topic, this table displays the top 5 documents (based on the highest P_dz values for that topic) with details including document ID, filename, document length, and the probability value. Following the document list, the table also shows the top 5 words for that topic (from P_zw) along with their corresponding values.

You can also use Interactive Terminal Mode (see wiki entry for details).

Command:

```
python visualization.py --matrix_dir vectors_in_csv/plsi_vectors --data_dir
out/vectors --dz_filename PLSI_P_dz_20topics_120iter_reweighted_count_vectors.csv
--zw_filename PLSI_P_zw_20topics_120iter_reweighted_count_vectors.csv --output_dir
vectors_in_csv/plsi_txt_reports
```

Modify the above command depending on your training/preprocessing configuration.

How to run

5. 2) Visualization Dashboard for Document-Topic Similarity Graphs (wiki entry)

As per instructions in the wiki entry, go to `${PROJECT_ROOT}/document_network`, and launch `index.html` to view a network visualization of the documents according to their topic distributions.

3) Cluster analysis

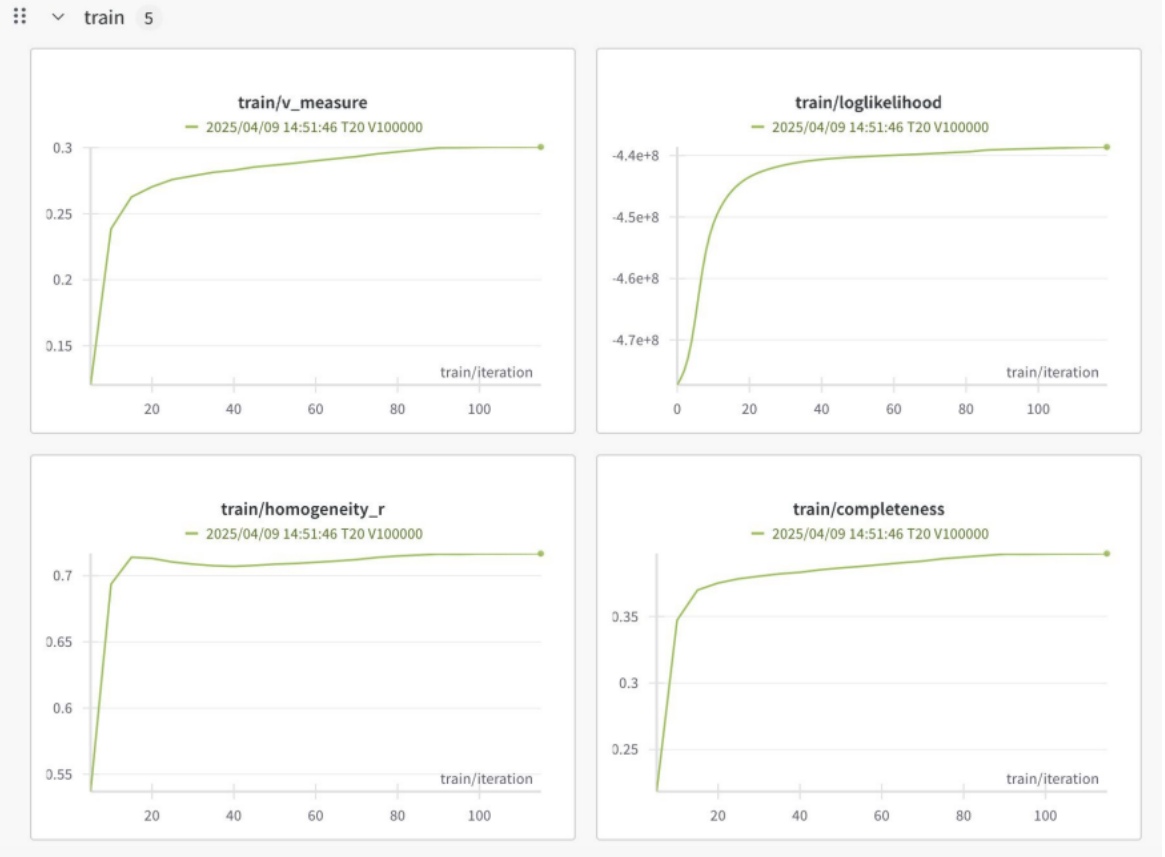
Launch `cluster_analysis.ipynb` in Jupyter Notebook.

Algorithm - pLSI

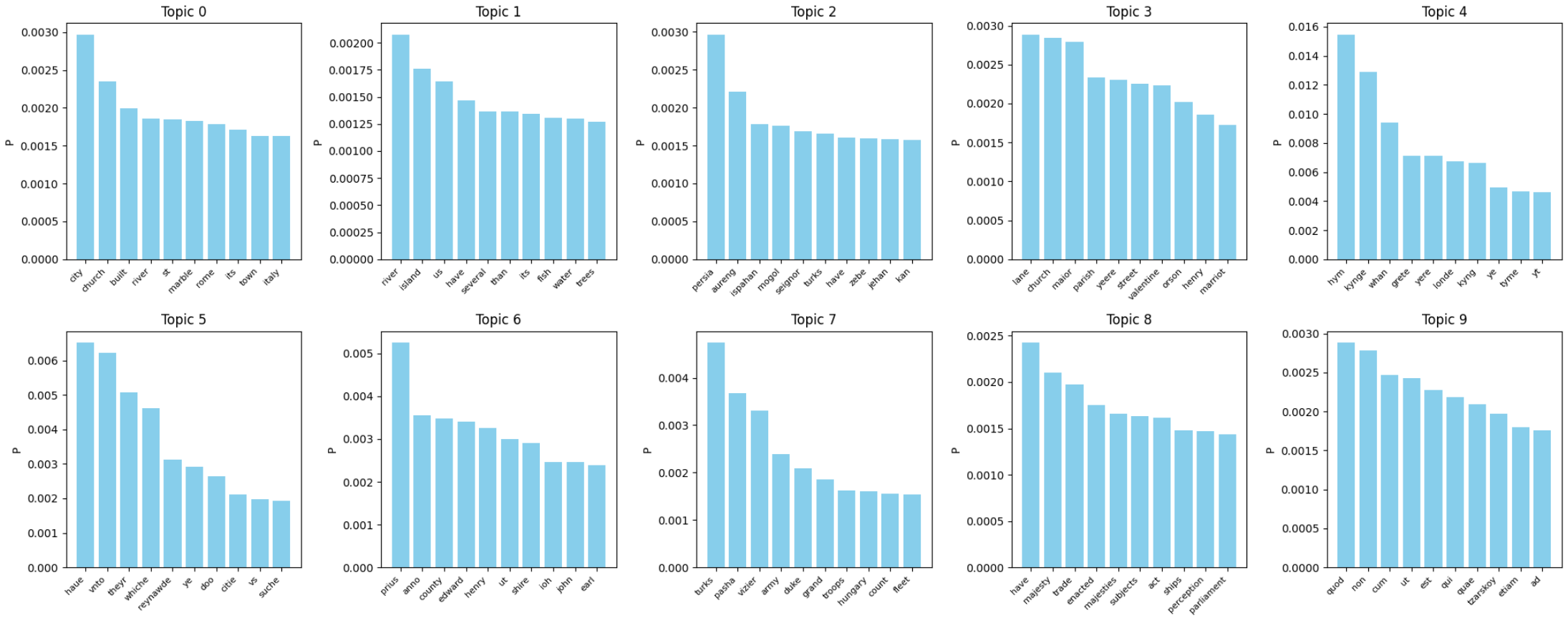
- **Probabilistic Latent Semantic Indexing (pLSI)** is a topic modeling algorithm that models the probability of a word appearing in a document as a mixture of topics. Each document is represented as a distribution over latent topics, and each topic is a distribution over words.
- **Latent topics** are hidden variables that explain the co-occurrence of words and documents.
- The model assumes that:
  - Each word in a document is generated by first selecting a topic based on the document's topic distribution.
  - Then, a word is selected based on the topic's word distribution.
- For more details on the algorithm and terminology, see our Algorithms Document.

Training pLSI

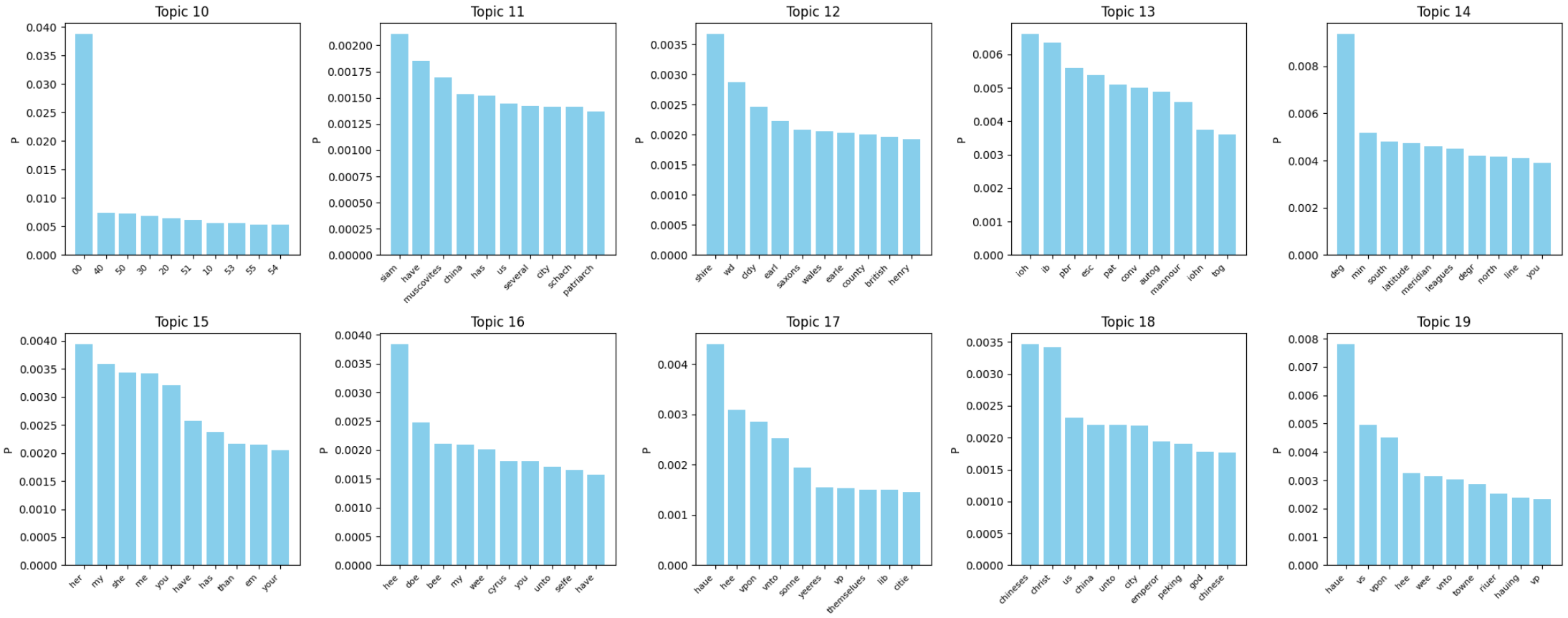
- As you train the pLSI model, you can observe the log likelihoods and clustering metrics (V-measure, homogeneity and completeness) over time, and you should choose the number of training iterations such that they converge by the end.



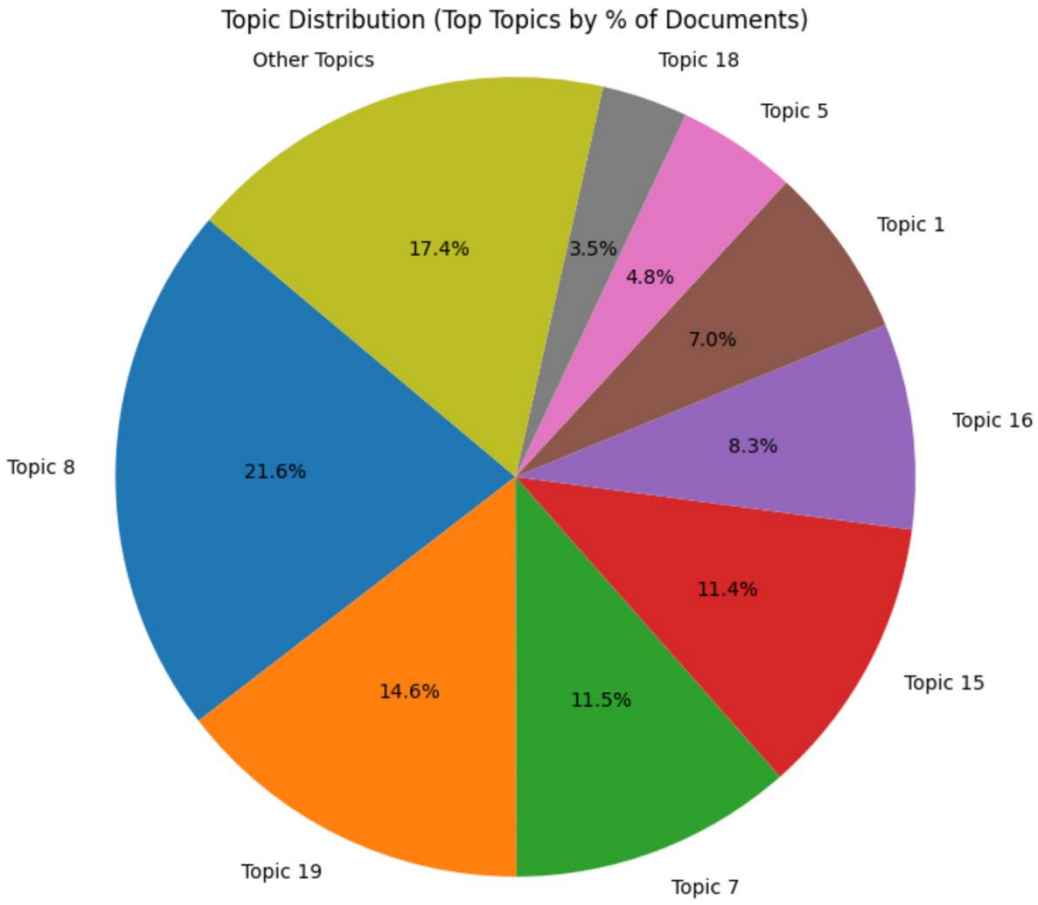
Result analyses – topic word distributions



Result analyses – topic word distributions



Result analyses – topic distribution



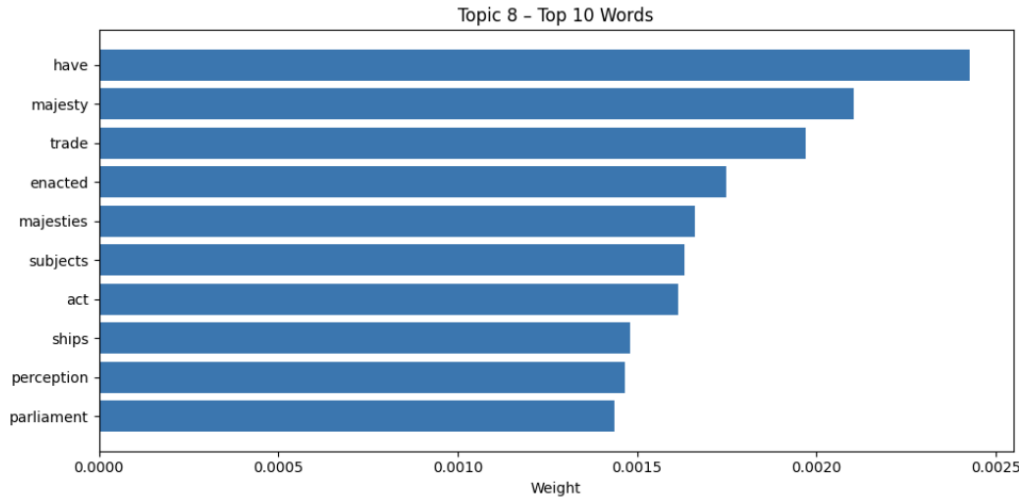
Result analyses – topic analyses

- The topic distribution shows a clear concentration around a few dominant themes. Topic 8 appears most frequently, representing about 21.6% of all documents. Topics 19, 7, and 15 follow closely behind, each making up between 11% and 15% of the corpus. By contrast, topics like 13 (0.14%), 6 (1.02%), and 2 (1.09%) are far less common (all ~1% or less). We can conclude that these topics are more niche or specific content area.

- To better understand what these dominant topics represent, we'll now break down the most common words associated with each one and assign thematic labels based on their vocabulary patterns.

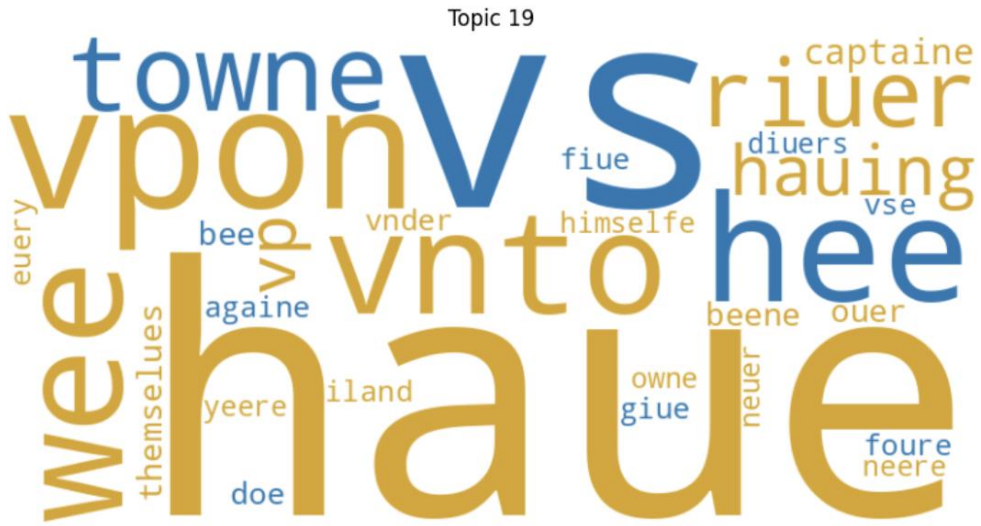
Result analyses – topic 8

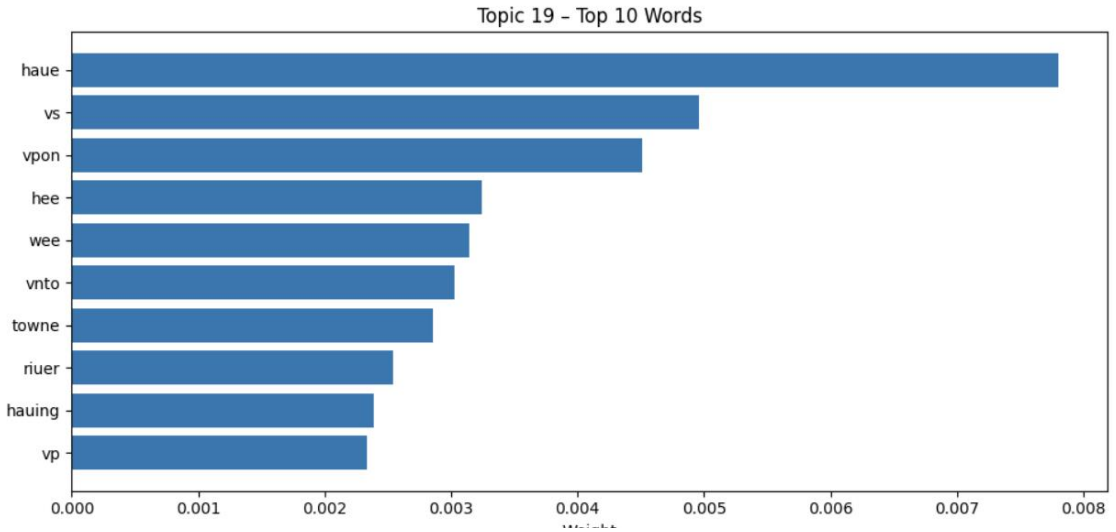
- **<u>Topic 8 centers on themes of trade and governance.</u>** Words that emerge include “majesty,” “parliament,” “act,” and “enacted” which all point to institutional and legal activity. Meanwhile terms such as “trade,” “goods,” and “ships” suggest a focus on commerce and travel. Also something to note is the prominence of proper nouns and formal legal verbiage. “England,” “India,” and “company” suggests a connection to imperial activity. The word cloud highlights the dominance of this political and economic language, and the bar chart reinforces this with the statistical weight of those terms. As the most common topic in the corpus, Topic 8 reflects a strong focus on power structures, policy, and cross-empire exchange.



Result analyses – topic 19

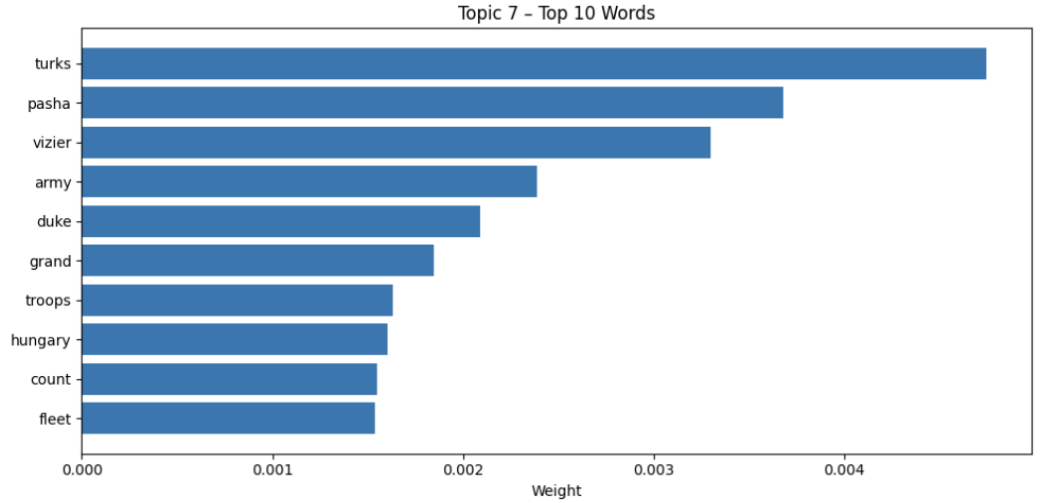
- The early modern language can make this topic difficult to interpret initially, but on upon further research, we can conclude that Topic 19 centers on **themes of navigation and travel.** The most frequent terms include “locational” words like “captaine,” “towne,” “riuer,” and prepositions like “vpon” (variant of the word “upon”) which all point to travel/exploratory missions, possibly tied to naval travel. There is also a use of pronouns such as “we,” “vs,” and “themselves” which suggests these documents were written in the first person, most likely in the form of journals and letters. The word cloud highlights these “location” words. With over 14% of documents linked to this topic, Topic 19 emerges as a key narrative in the corpus, likely tied to first hand accounts of imperial travel.





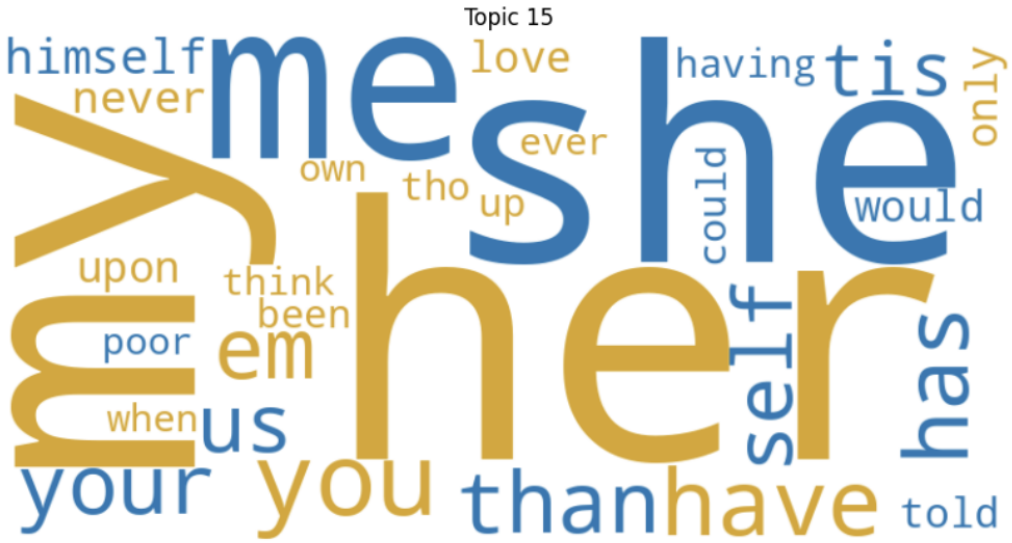
Result analyses – topic 7

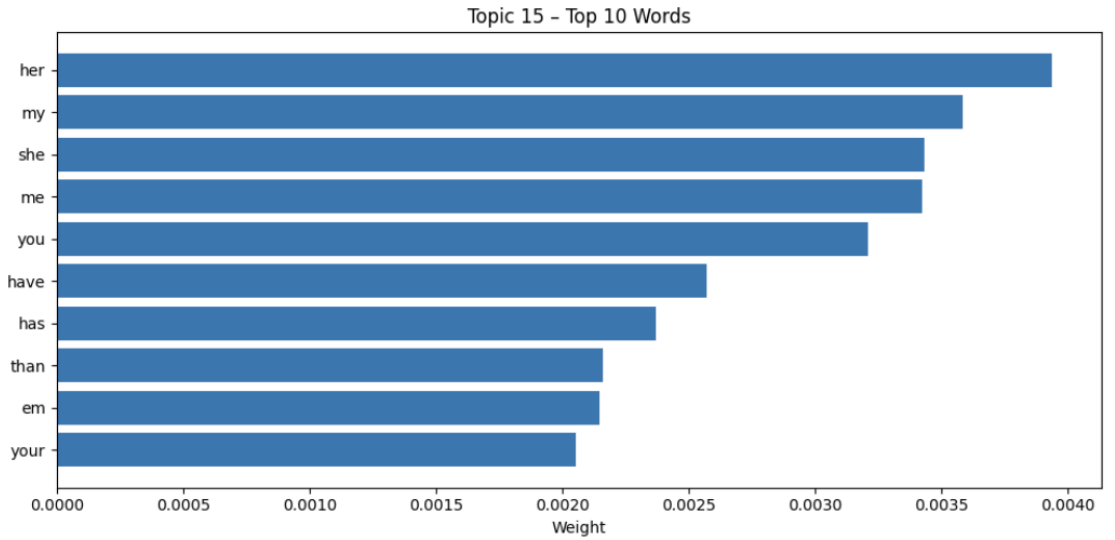
- In analyzing Topic 7, what stands out is the **prominence of political and military language.** Terms like “turks,” “vizier,” “pasha,” “sultan,” and “vienna” point to European-Ottoman tensions, likely involving diplomatic and military interactions between the two powers. Other terms such as “siege,” “fleet,” “troops,” and “enemy” reflect the language of warfare and strategy. Overall, this topic likely captures historical accounts of military campaigns and global dynamics between alliances. The word cloud mainly highlights certain titles or ranks (and possibly names of key figures), while the bar chart reinforces this. Over 11% of documents are tied to this theme, and overall, topic 7 reflects the recurring narrative we see throughout the documents of international warfare/conflict.



Result analyses – topic 15

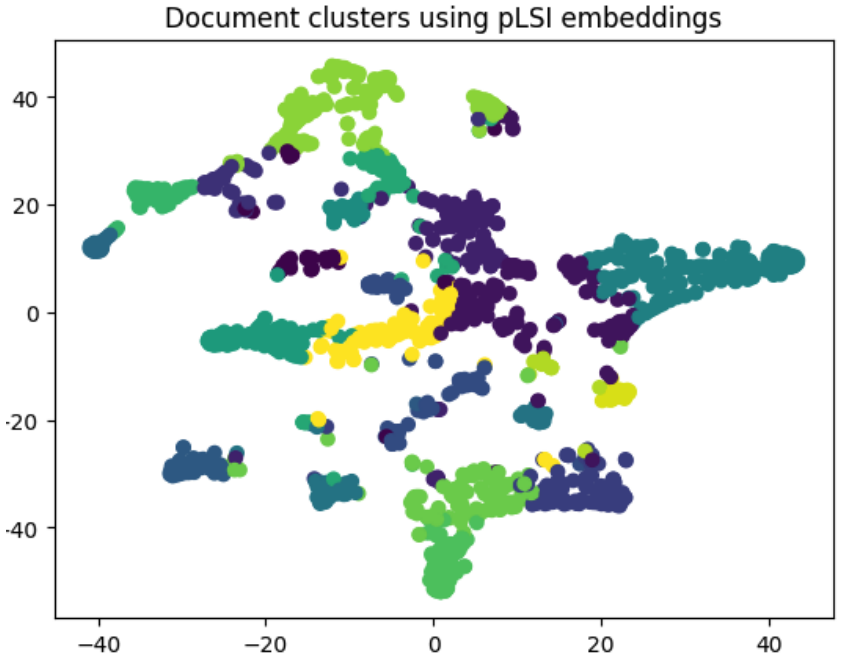
- Topic 15 diverges from the rest of the topics with its strong focus on **<u>emotion and personal relationships.</u>** This is indicated by the strong presence of relational terms such as words like “her,” “she,” “love,” and “told”, as well as the prominence of first-person pronouns such as “me” and “us”. These words suggest a strong presence of first-person narratives, possibly drawn from letters or diaries. The language is intimate and reflective, marking a clear shift from the formal or legal tone of the other topics. The word cloud emphasizes emotional and interpersonal vocabulary, while the bar chart confirms their frequency and weight.





Result analyses – cluster analyses

- Using topic distributions per document derived from pLSI, we can cluster the documents using unsupervised algorithms. We can cluster all 1466 documents into 20 clusters (same number as the number of topics) using K-Means, as can be seen in the figure on the left (projected to 2d via TSNE), which can be compared against other “natural” clusters that can be generated using document metadata.
- Another method is to assign for each document its most dominant topic, which would cause information loss for less likely topics but would be more interpretable for we know the word distribution for each topic.

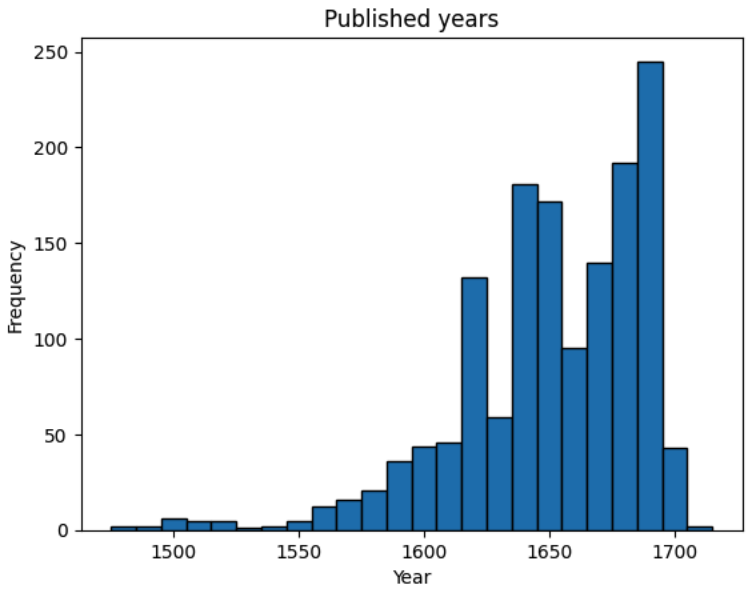


Result analyses – cluster analyses (vs. authorship)

- We would like to compare our derived document clusters against document clusters that can be generated using the author information as well as published year information (in the future we could also incorporate metadata such as published location and more that can be found in EEBO metadata).
- However, after processing the author information, after removing abbreviations such as “d.”, “ca.”, “cent.” and preprocessing to filter out duplicates, we ended up with a rather disproportionate 568 unique authors (out of 1466 documents), and the top authors are as follows (first value in tuple indicating the number of documents attributed to the author; nan indicating unknown authorship):
  - `[(581, 'nan'), (41, 'england and wales parliament'), (23, 'england and wales'), (21, 'england and wales sovereign charles ii'), (20, 'england and wales sovereign charles i'), (15, 'scotland privy council'), (12, 'taylor john'), (11, 'england and wales sovereign james i'), (8, 'ward edward'), (8, 'ogilby john'), (8, 'butter nathaniel'), (7, 'penn william'), (7, 'mather cotton'), (7, 'england and wales sovereign elizabeth i'), (7, 'company of scotland trading to africa and the indies'), (6, 'united provinces of the netherlands staten generaal'), (6, 'r b'), (6, 'mandeville john sir'), (5, 'smith john'), (5, 'howell james'), (5, 'camden william')]`
- Unfortunately, these are not very helpful for clustering, since most authorship here does not convey a lot of information. Therefore, for the time being, we will not be clustering our documents by authorship.

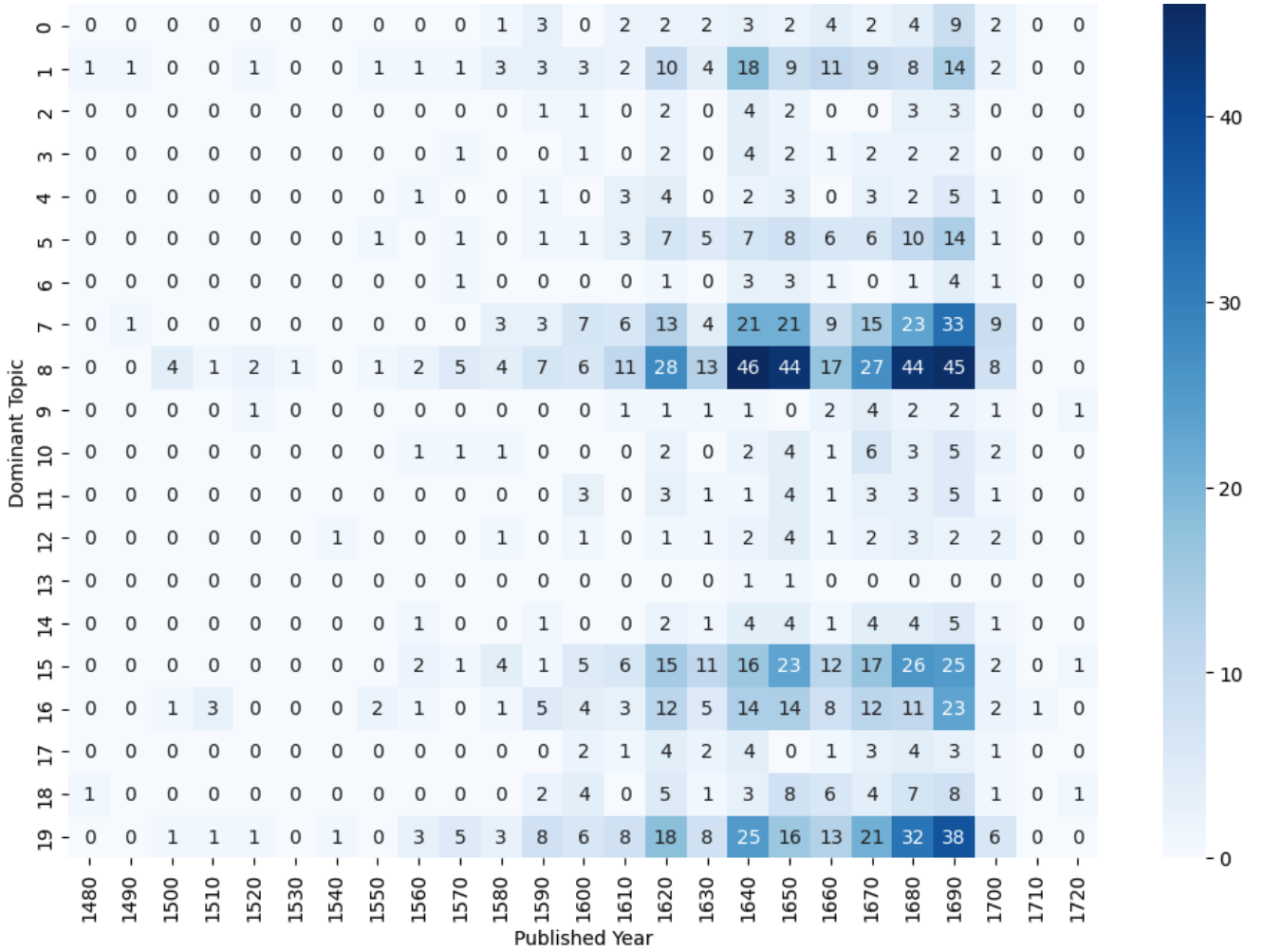
Result analyses – cluster analyses (vs. published year)

- Since all documents come with information about in which year it was published, we can generate a histogram of documents' published years.
- Judging by the distribution, it should be appropriate to cluster each document into its own decade bin, and compare its clustering to its most dominant topic clustering, which can be seen on the next page.



Result analyses – cluster analyses (vs. published year)

- We can observe topics such as 8 (trade & governance) remaining dominant all throughout (although there are periods during which their dominance dip, such trend happen across other topics concurrently).
- We can also observe topic 19 (navigation & travel) gaining popularity towards the end of the 17th century



Final words

- For weekly progress and more detailed analyses, please see Slides

- This report was made by Myungjin Shin, and the detailed per-topic analyses (pp. 11-16) were done by Alina Lee.