Constructing Utopias in Restoration London

Audrey Liu, Leona Lu, Erika Wang

Project 17: Constructing Utopias in Restoration London

Project Leads: Nicholas Smolenski, Astrid Giugni
Team: Audrey Liu, Leona Lu, Erika Wang

Project Overview

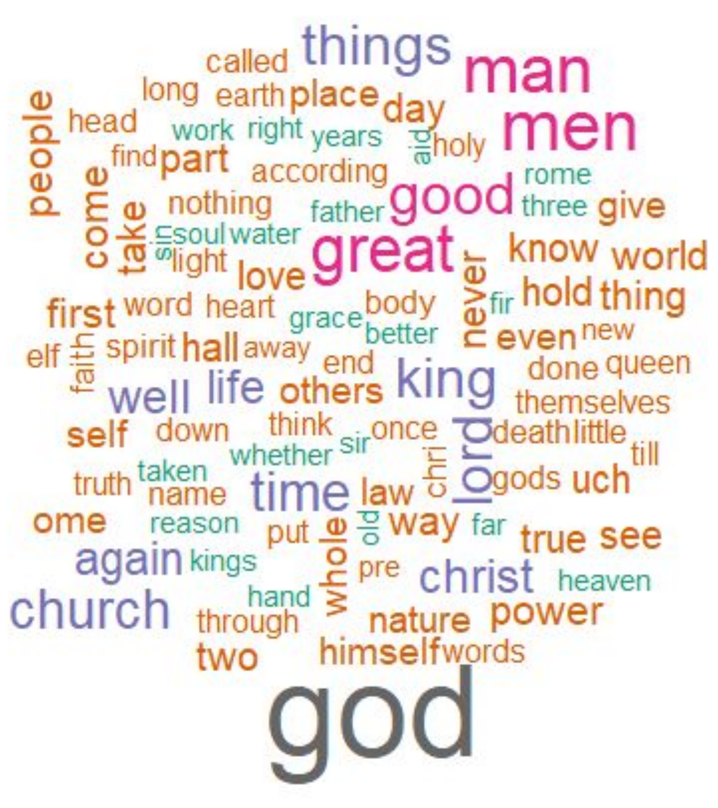


Figure 1: Top 20 Tokens appeared in 17th century texts

After London was destroyed during the Great Fire of 1666, it was reconstructed into a utopia of Europe. Who was this utopia constructed for? Who determined its structure? And what did it look like?

We use Natural Language Processing to analyze semantic trends in digitized texts from EEBO-TCP to answer these questions. Using methods such as word-embedding, sentiment analysis, and hapax richness, we provide a macro-view on themes in the seventeenth century, as well as specific case studies on coal taxes and St Paul's Cathedral.

Methods

Text Cleaning- VARD 2.0
We used Vard 2.0 to normalize non-standard spellings in seventeenth-century printed text. We used 50% threshold to avoid over-normalization.
Word Embedding
We applied Word2Vec and GloVe to create geometric representations of words. We evaluated semantic relationships across word vectors, using distance to calculate cosine similarity.
Sentiment Analysis: BING-Dictionary
We used the BING dictionary to assign sentiment of either positive or negative to words in our text. *Sentiment Score = # positive words / # total*
Hapax Richness:
We evaluated the uniqueness of text by calculating the number of words that appear once divided by total number of words.

Macro-Analysis: Word-Embedding

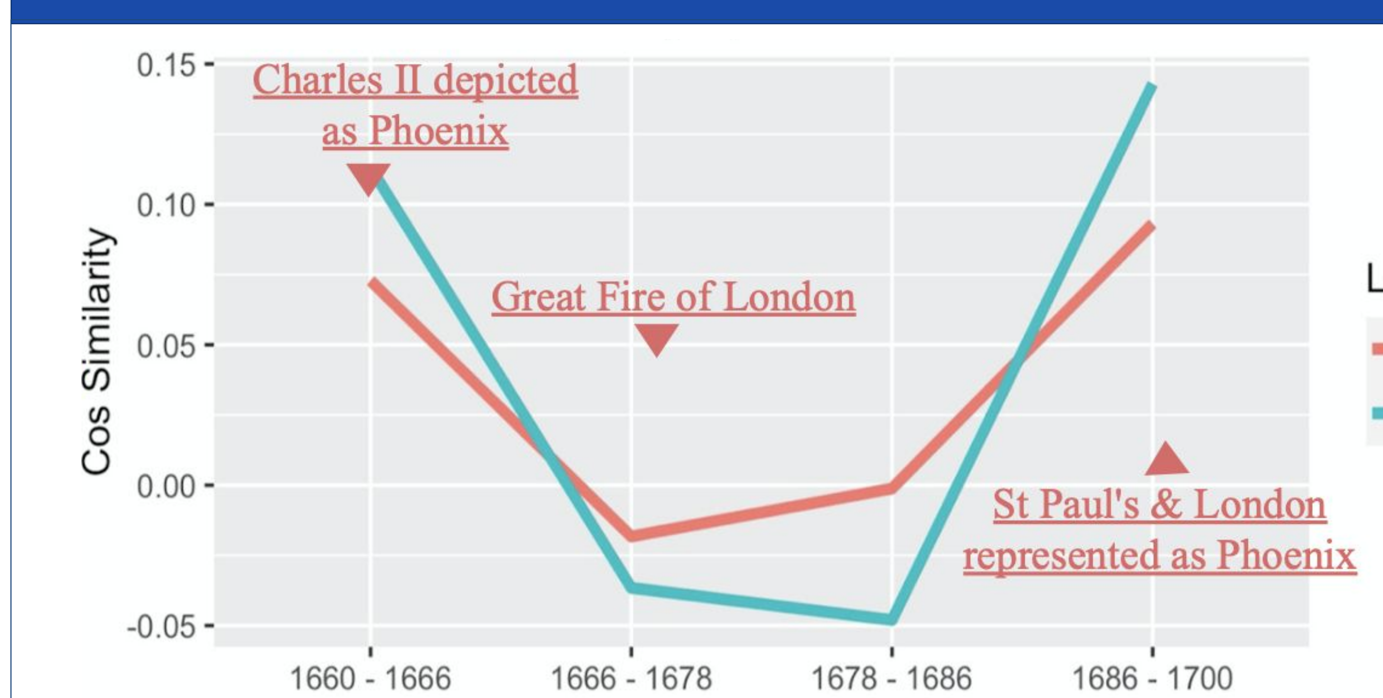
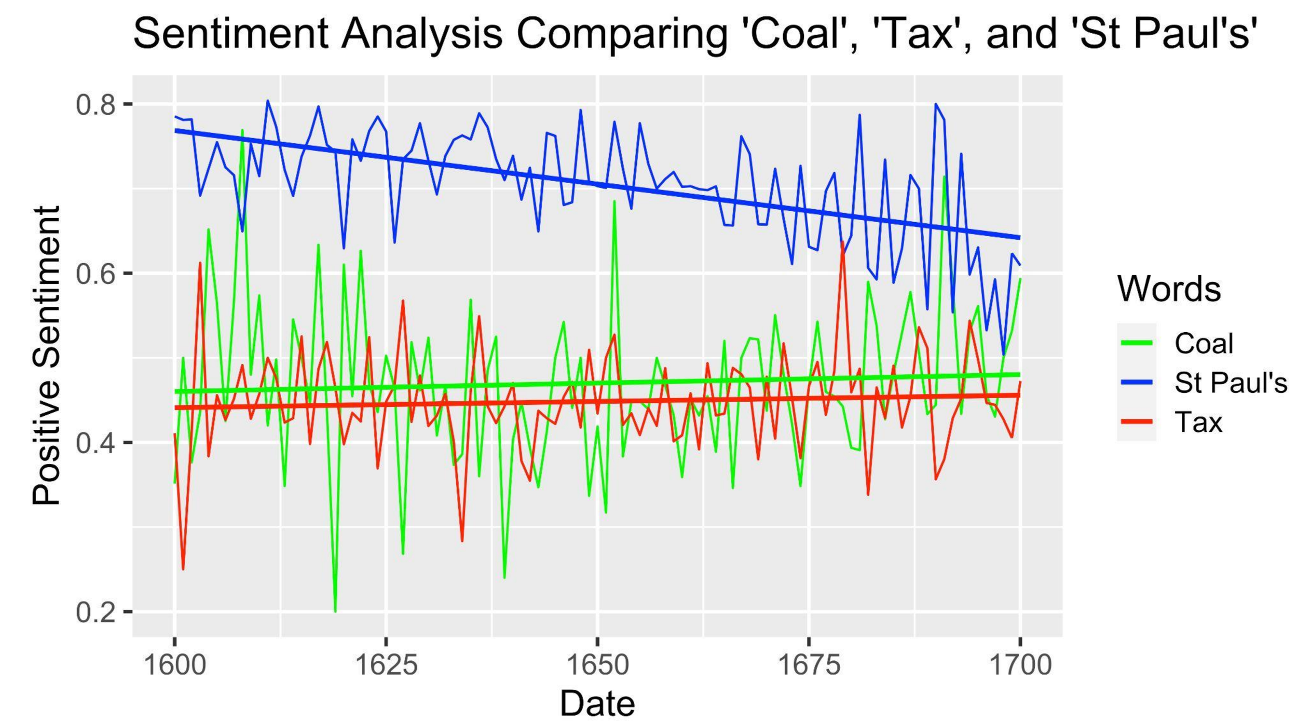


Figure 2: Cos-Similarity between "Phoenix", "Christ", and "King"

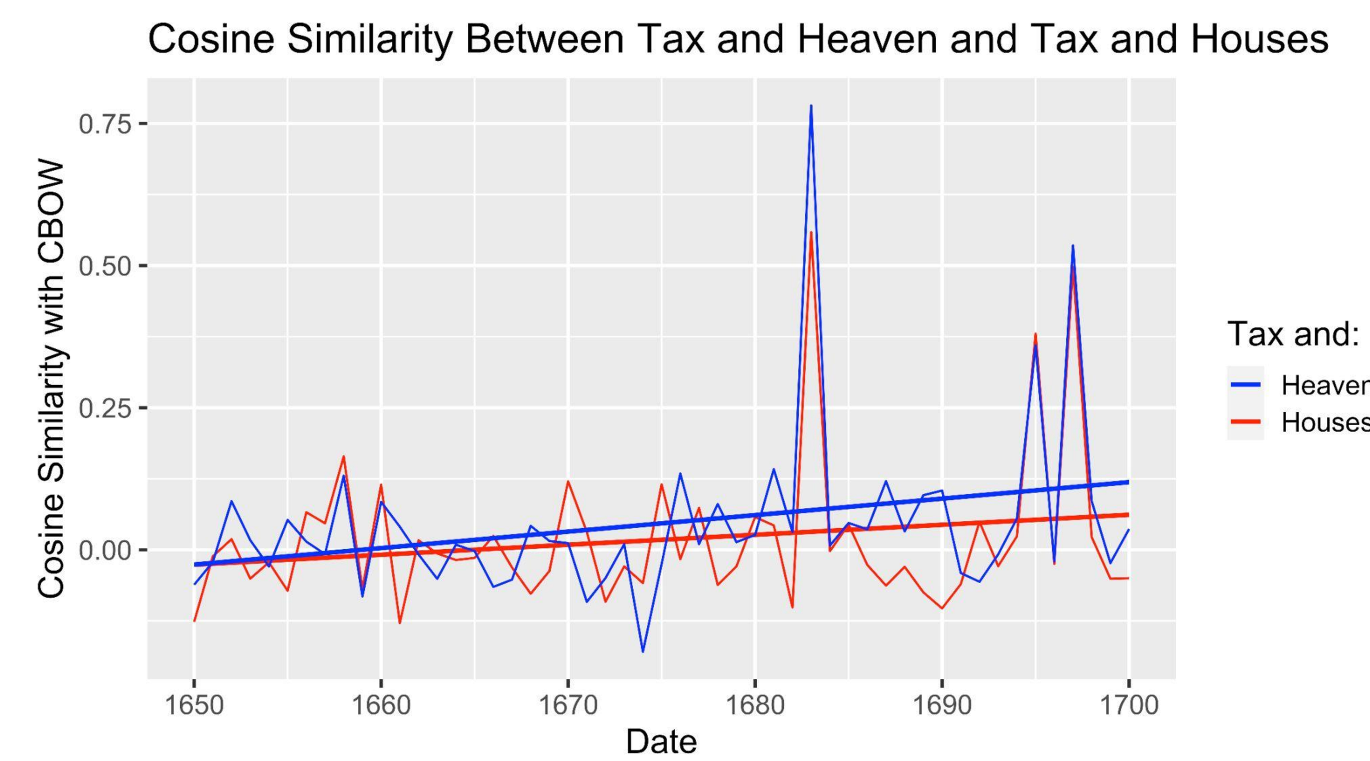
The phoenix, a metaphor used throughout London's reconstruction, became viewed with increasing political motives over pragmatic ones.

A utopian society was initially intended to be built for the people, prioritizing pragmatic initiatives over symbolic ones. However, the purpose of reconstruction became more politically oriented with the approval of more costly projects.

Micro-Analysis: Coal Taxes, Sentiment Analysis, Cosine Similarity



1. The 1666 Fire and the reconstruction might have generated backlash, possibly due to coal, as indicated by a decrease in positive sentiment of words related to St Paul's, especially after 1650.
2. The slight increase in positive sentiment of words related to coal and tax could imply that writers started depicting coal and taxes in a more positive manner in response to backlash.



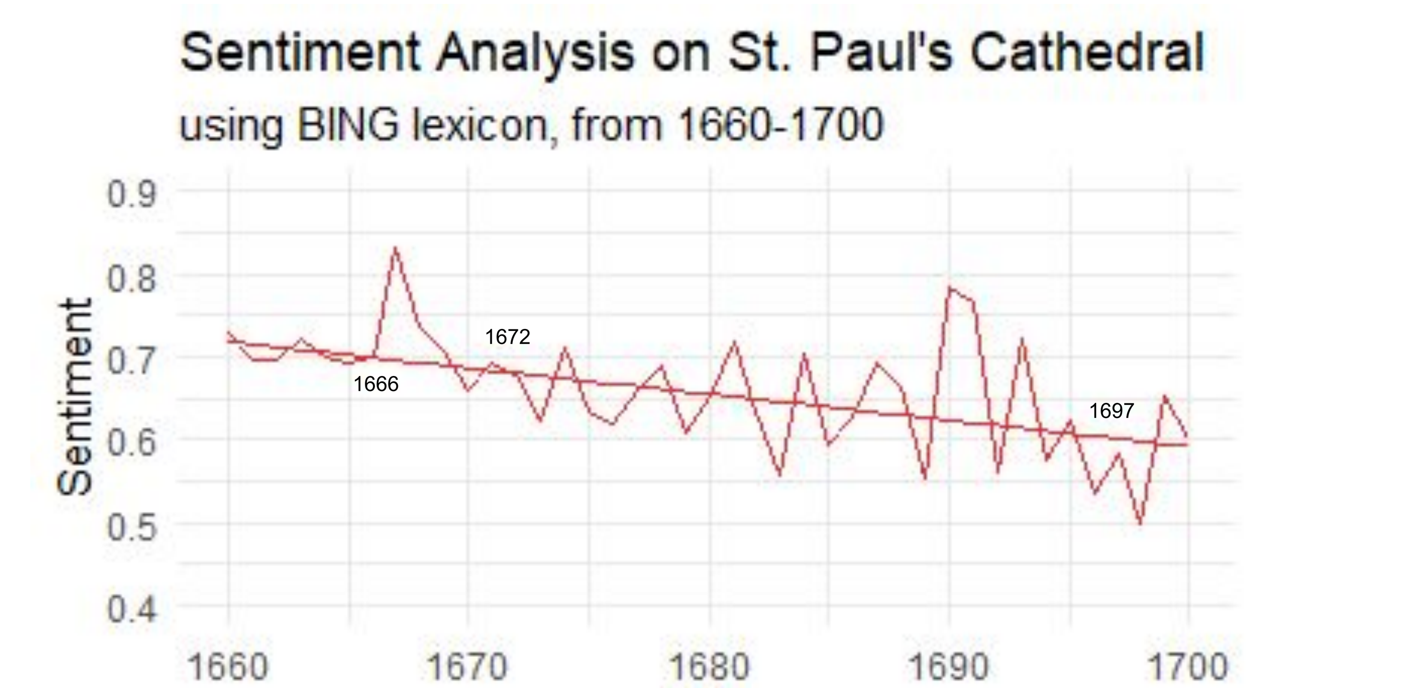
1. An increase in cosine similarity between houses & taxes and heaven & taxes, implies an emphasis on economic pragmatism within contemporaneous literature and a portrayal of taxes as a vehicle for the greater good.

Limitations and Next Steps

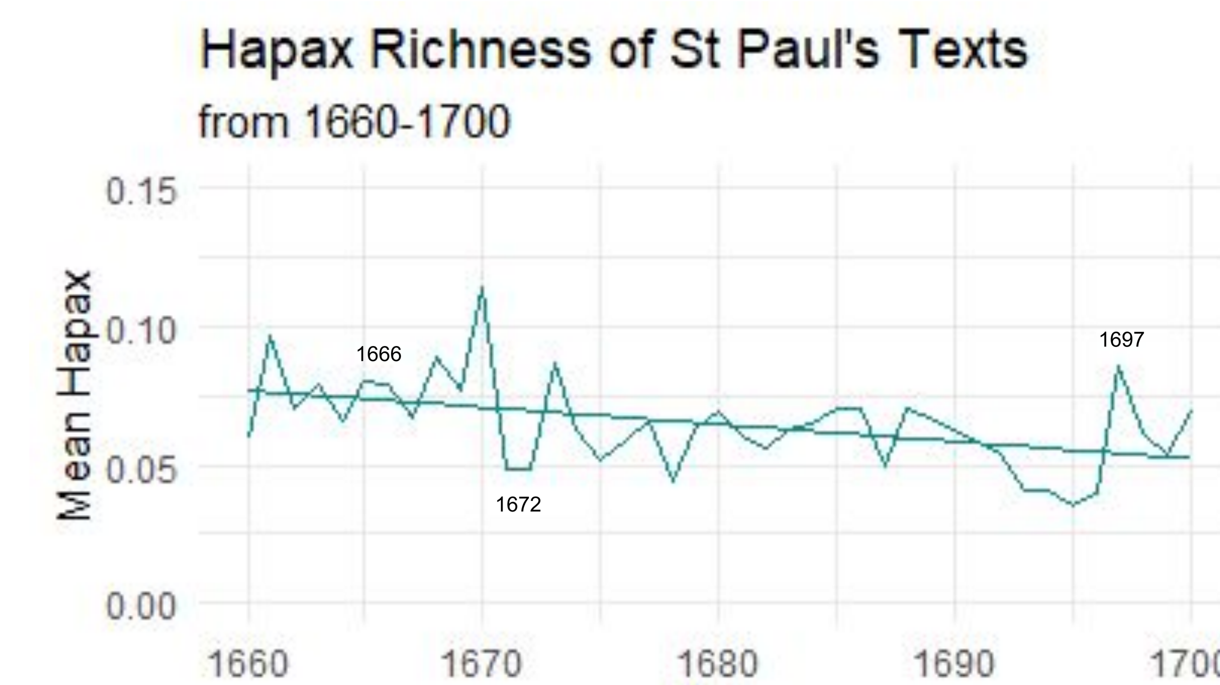
Limitations:
- Printed texts from the working class were often less preserved and underrepresented in our data set, which means it may not be completely representative of the population.
- Using VARD to normalize spelling did not guarantee all spelling variations to be normalized, thus hapax richness could be higher than anticipated and sentiment analysis could be skewed due to the pre-made dataset being unable to recognize certain words and assign a sentiment value.
- Unsupervised Learning as texts are not labeled with genre.

Next Steps:
- Since our analysis results indicate correlation rather than causation, we plan to investigate causation among various variables to form a more concrete understanding of seventeenth-century England.

Micro-Analysis: St Paul's Cathedral, Sentiment Analysis, and Hapax Richness



1. Our initial results indicate that changes in sentiment reflect a few major historical events relevant to the reconstruction (i.e. Great Fire, coal taxes, consecration)



1. Though overall trends are similar, sentiment and hapax richness have an inverse relationship during key historical events, which can connect to the genre of the text (i.e. poetry as a form of protest relates to positive hapax richness and negative sentiment)
2. Hapax richness and sentiment analysis can be used for a future testable hypothesis relating to genreship

References

Our project website (or scan QR code): https://sites.duke.edu/reconstructingutopia/
Our Github Repository: https://github.com/leona-lu/Reconstructing_London

Jockers, Matthew. *Text Analysis with R for Students of Literature*. Springer. 2014.

Silge, Julia and David Robinson. "Sentiment analysis with tidy data". *Text Mining with Tidy Approach*.

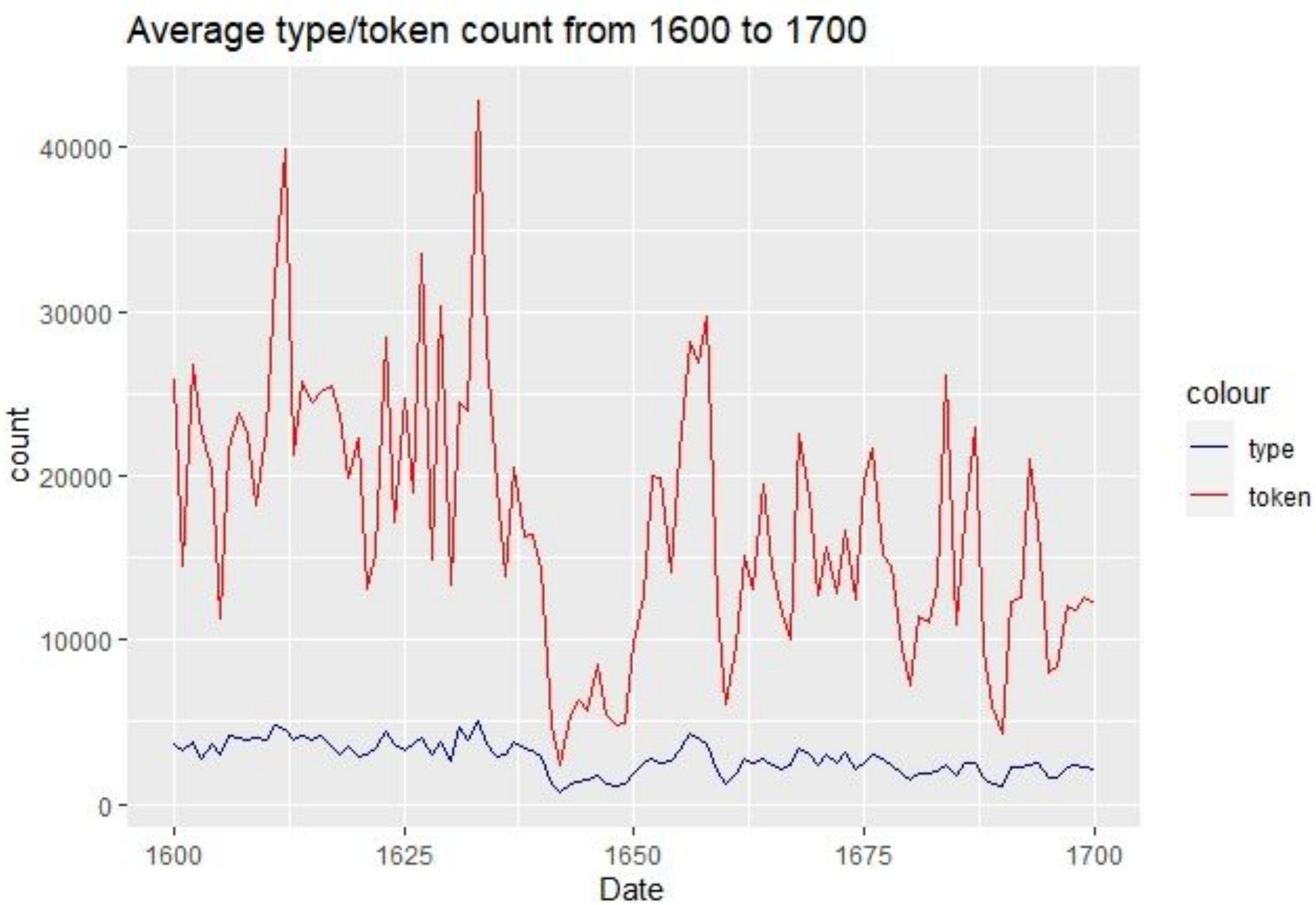
Kulshrestha, Ria. "NLP 101: Word2Vec — Skip-gram and CBOW". *Towards Data Science*. 2019.

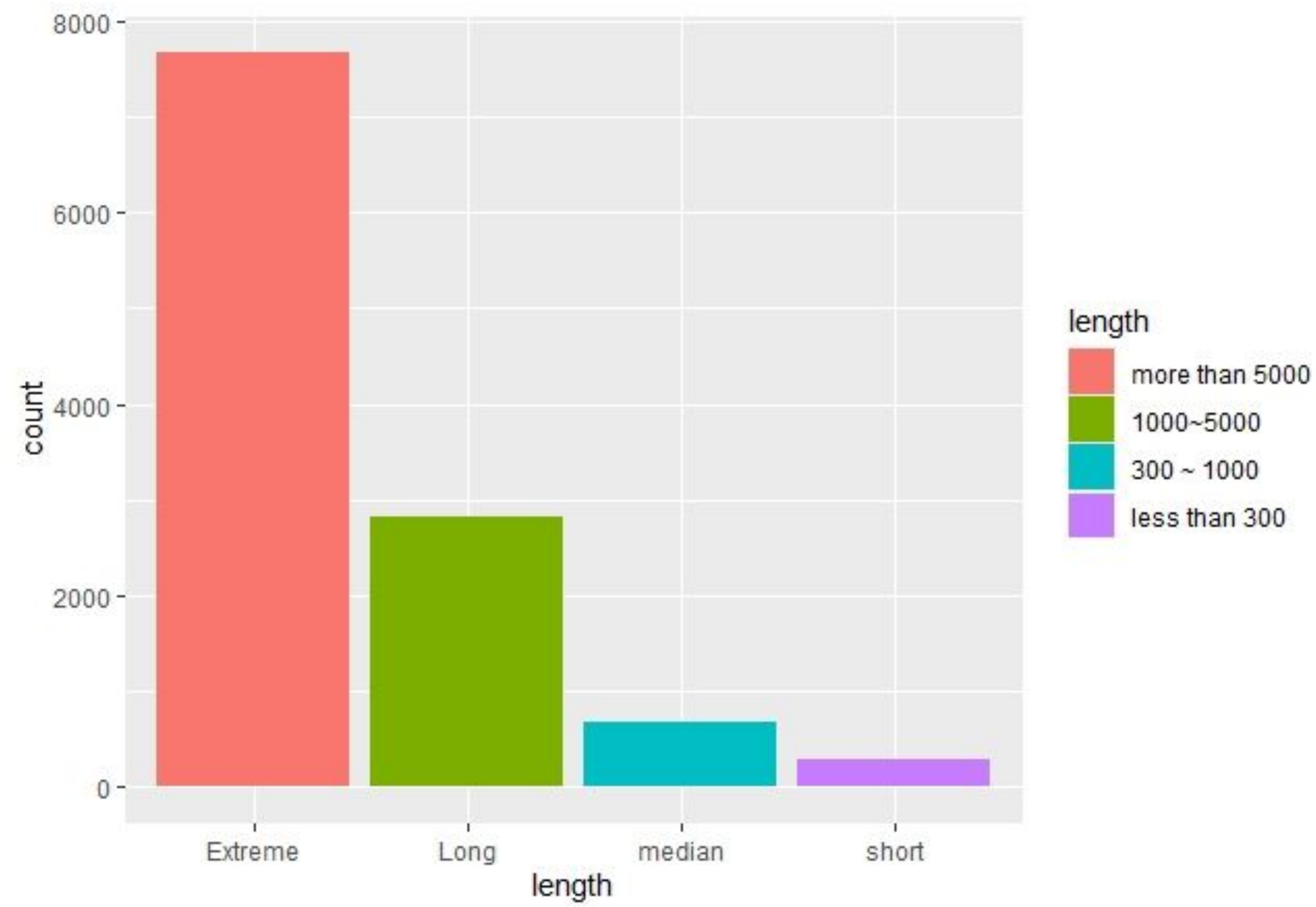
Prabhakaran, Selva. "Cosine Similarity – Understanding the math and how it works (with python codes)". *Machine Learning +*. 2018.

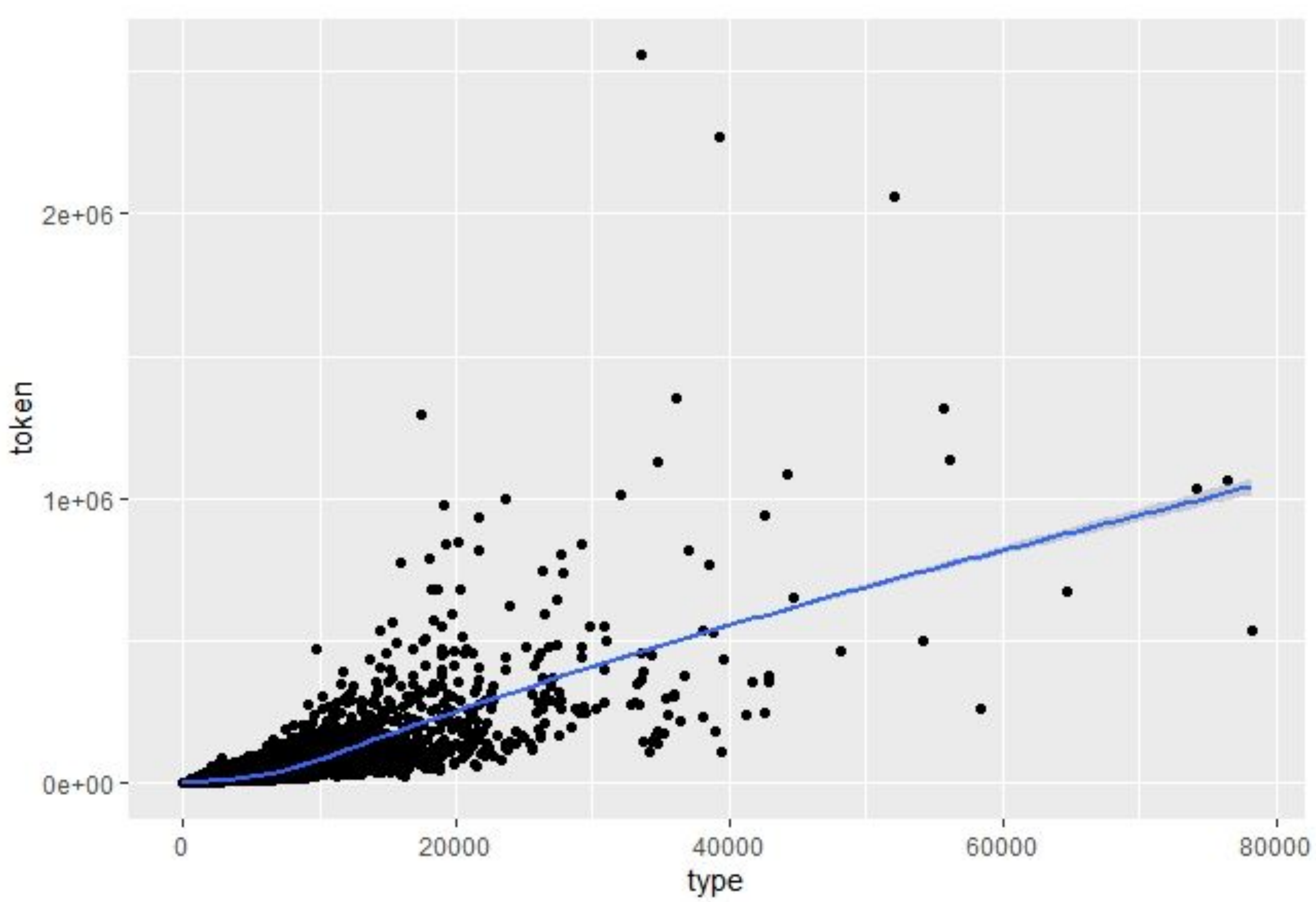
Socher, Richard. "CS 224D: Deep Learning for NLP", Lecture Notes(https://cs224d.stanford.edu/lecture_notes/notes2.pdf)

Charu C. Aggarwal. 2018. Neural Networks and Deep Learning: A Textbook (1st. ed.). Springer Publishing Company, Incorporated.

Macro-Analysis: Basic EDA of Text







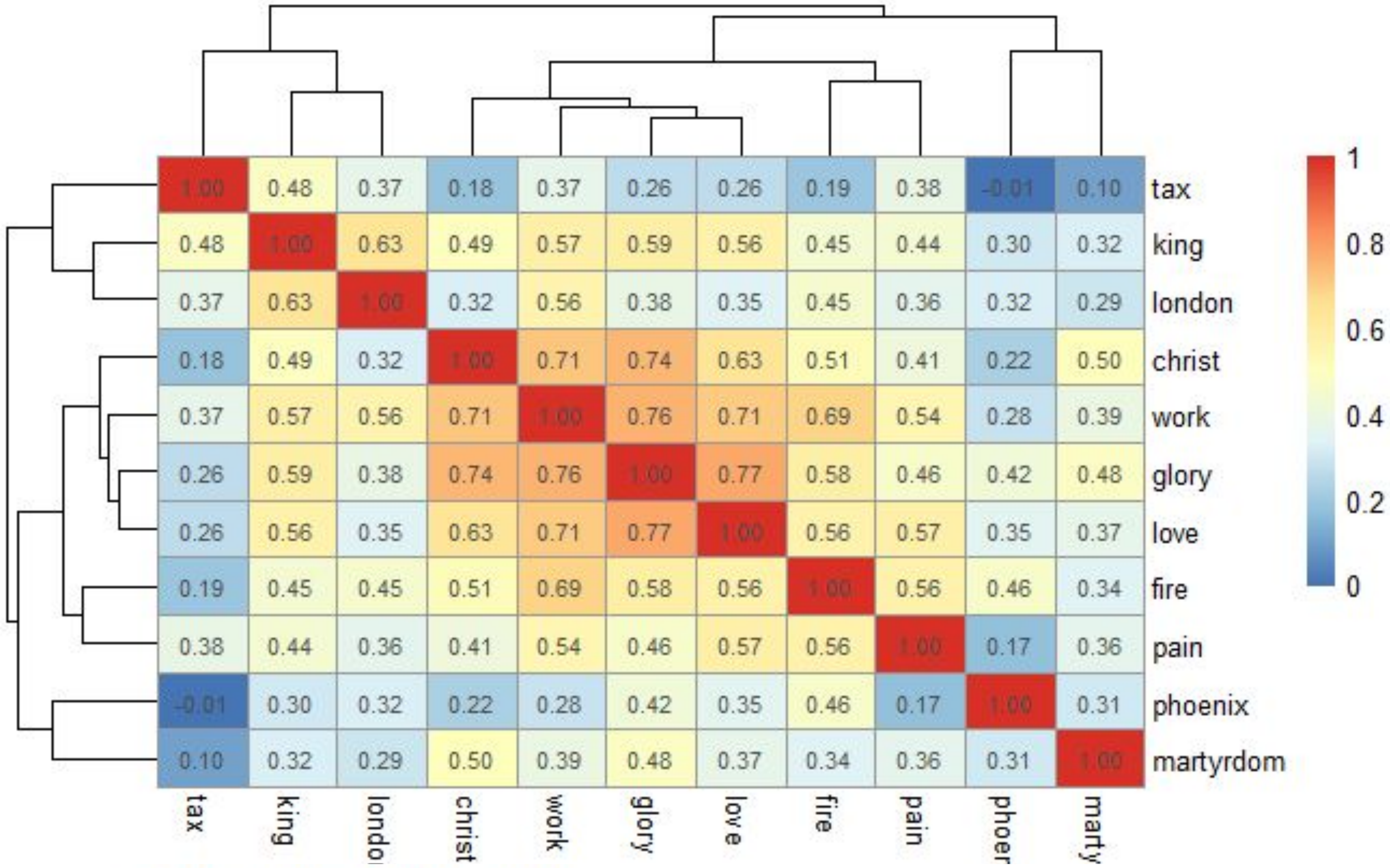
Findings :

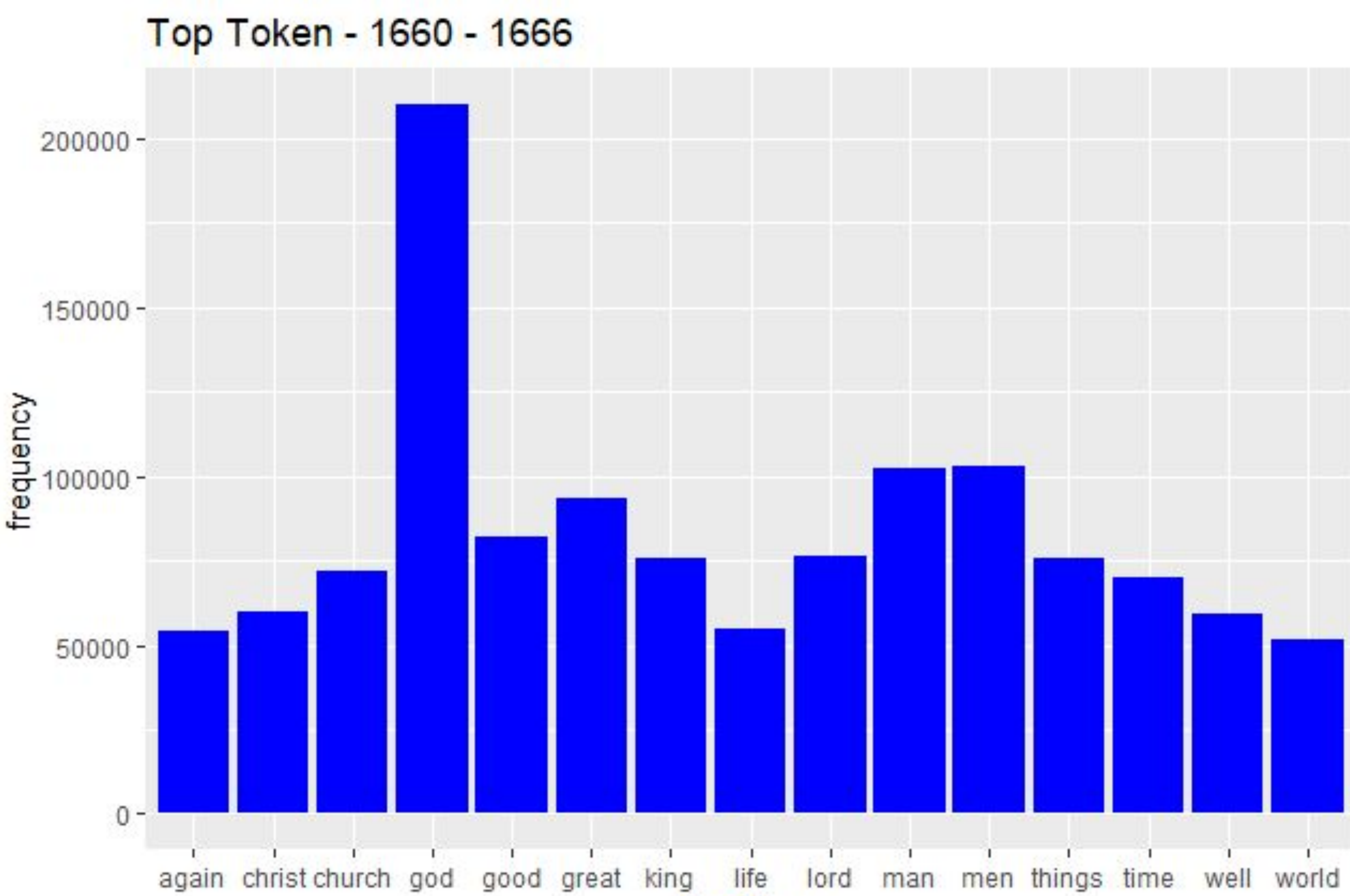
- Texts from EEBO-TCP Database are predominately long: implication that we mostly analyze books and pamphlets instead of poetry

- Token count increase with Type count (displaying a linear relationship)
  - lexical richness is correlate with text genre/length

- Pre-1650 (English civil war) displayed a higher type/token count compared to post-1650
  - Political instability causes decrease in #type/token and the return of monarchy causes increase # type/token

Macro-Analysis: Word Embedding





What is Word Embedding

- Word vectalization: similar word has similar distance

Why GloVe

- Utilize enter corpus instead of just within the context window (Word2Vec): producing better result semantically and syntactically for larger dictionary
- Count base instead of predictive based (developed based on Word2Vec model)

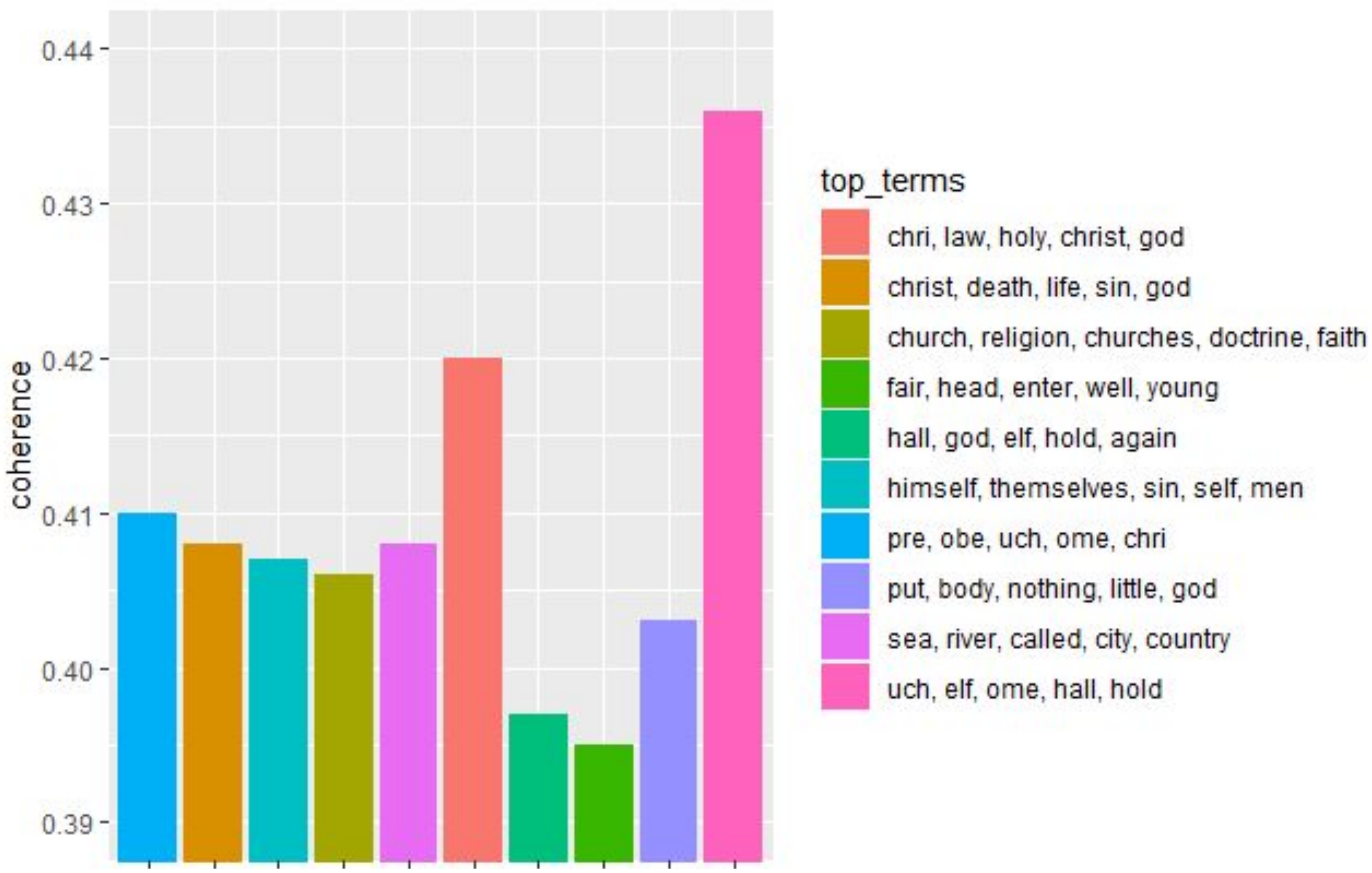
Limitations

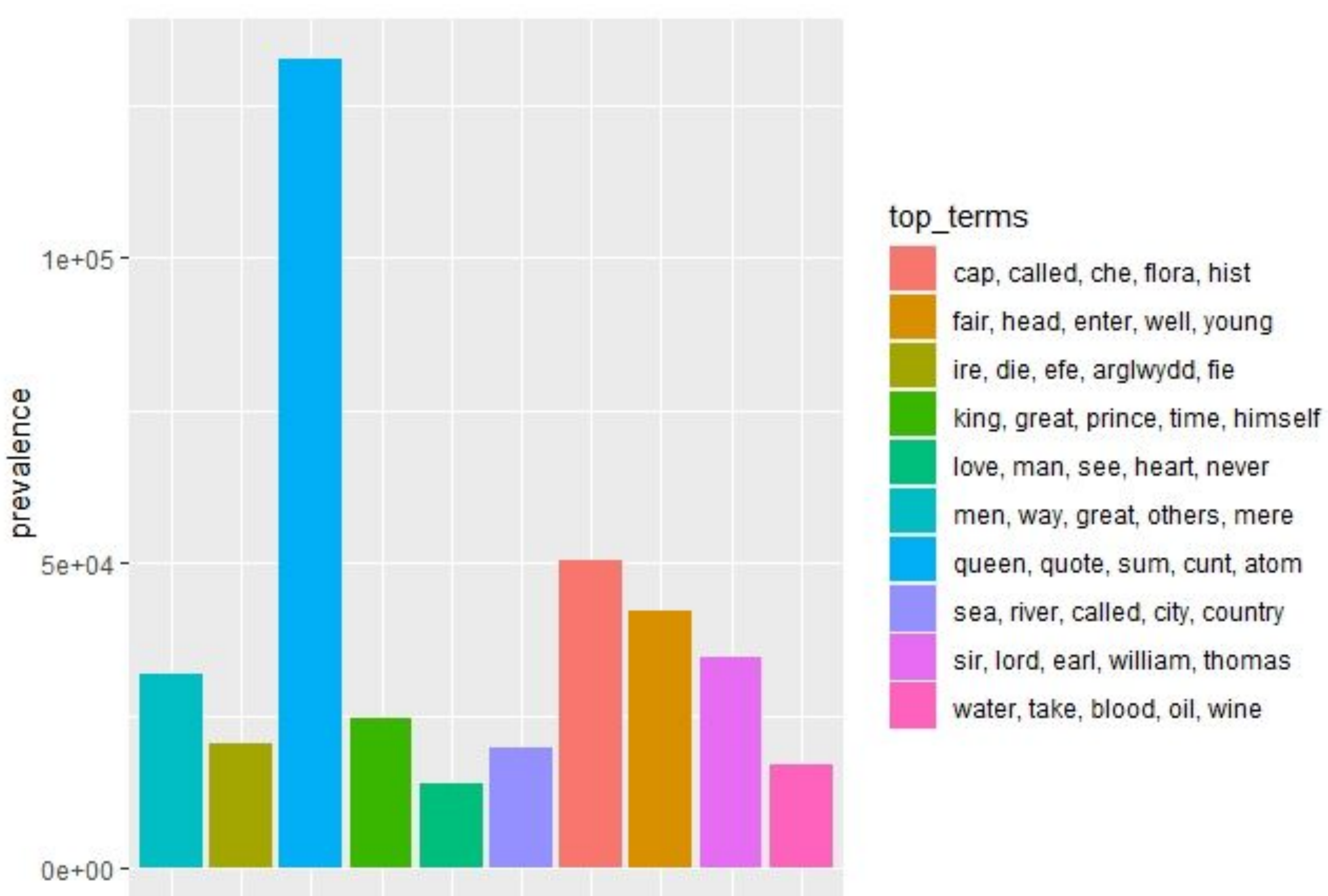
- We ran 5 iterations as Word Embedding Models are extremely computational expensive. We will produce better model with 1000 iterations
- Genre of text highly correlated with the top token and correlation. Labeling of author/audience will help with classification and decreasing bias in our model

Next Step

- Map phrasal embedding value on top of word value
- Increase iterations and classify corpus into different genre to decrease bias and increase model performance
- Experiment and track result for different Word Embedding algorithms

Macro-Analysis: Topic Modeling (Latent Dirichlet Algorithm)





What is Topic Modeling?

- Use Model to discover the abstract topics that occur in a collection of documents

Why LDA Model?

- Distributional hypothesis: similar topics make use of similar words

- Statistical mixture hypothesis: documents talk about several topics for which a statistical distribution can be determined (Dirichlet Distribution)

Prevalence V.S. Coherence

- Coherence: the probabilistic coherence of each topic (how associated words are in a topic)
- Prevalence: Statistical distribution of topics

Limitations

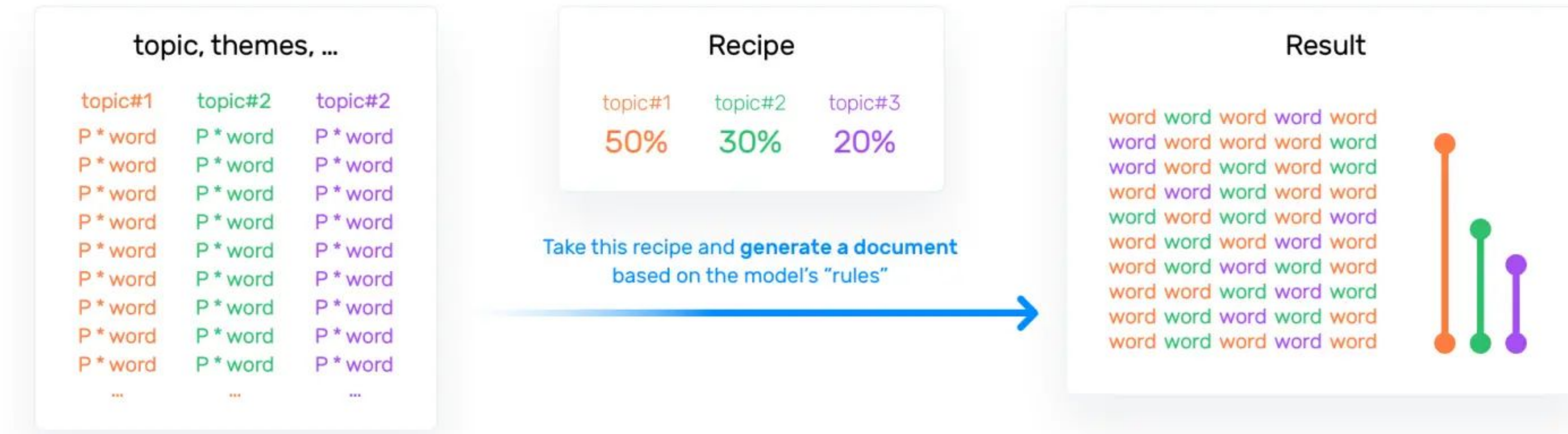
- Dirichlet topic distribution cannot capture correlations
- Fixed number of topics to be harvested

Findings

- Popular topic include:
  - Religion: christ,holy,god,church,faith,...
  - Livelihood: men, self, sin,...
  - Monarchy: queen, king, prince, great,...
  - Environment : sea, city, country, river, ...

Macro-Analysis: Algorithm Flow Sheet

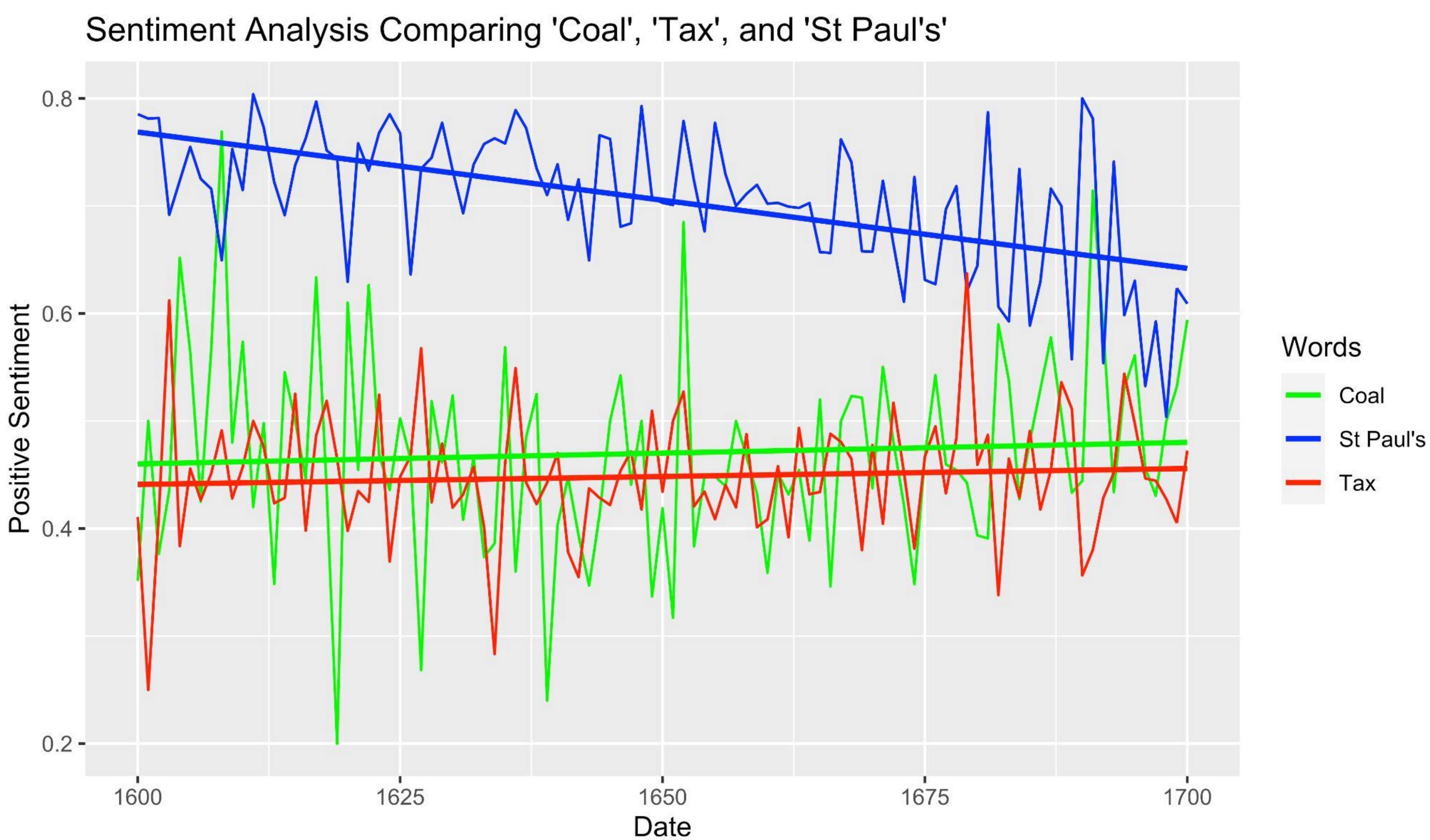
LDA Topic Modeling



GloVe Co-Occurrence Matrix

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Case Study: Coal Tax



Context

- Coal taxes were passed in 1667 and 1670 to help pay for the reconstruction of London after the Great Fire of 1666
- Backlash against coal dues took place from 1687 to 1697 given that there were to be two more decades of coal taxes being distributed to the already expensive project.

Our Focus

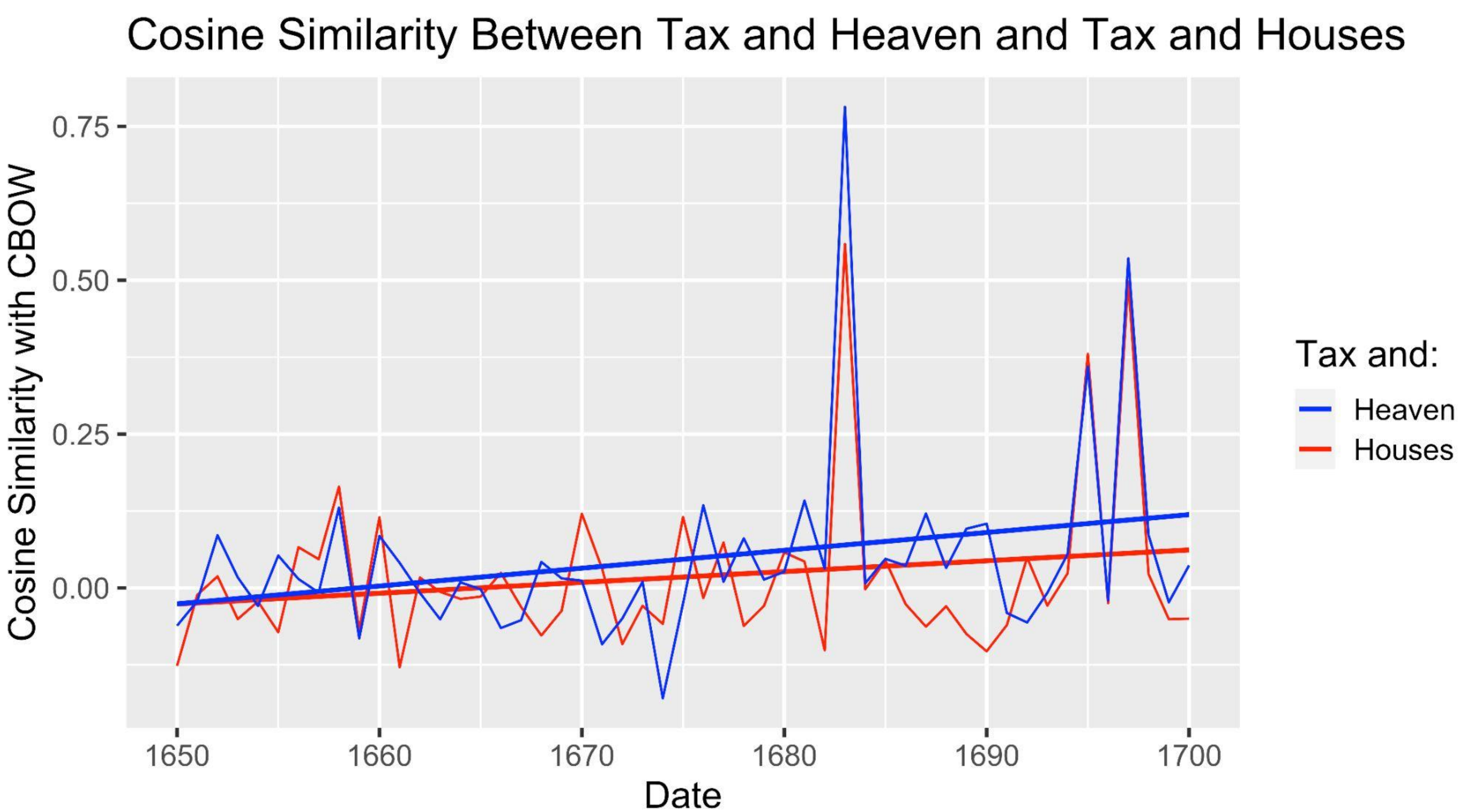
- How did the upper/ruling class respond to the backlash against coal taxes?
- How does this reflect the relationship between the upper and lower classes?
- How does this contribute to the formation of a utopia?

	word	sentiment
1	delight	positive
2	saint	positive
3	sin	negative
4	quarrel	negative
5	ready	positive
6	worthy	positive
7	condemned	negative

Sentiment Analysis

- The use of natural processing language (NPL) and text analytics to extract and quantify emotion and subjective information
- The BING dataset determines whether a word has positive or negative sentiment
- Positive sentiment = # positive words / # total words

Case Study: Coal Tax



Cosine Similarity

- After assigning a vector to each word, we performed cosine similarity between 2 different words.
- Cosine similarity is the calculation of the similarity between two n-dimensional vectors by looking for a cosine value from the angle between the two.
- A cosine similarity of +1 means that two words are perfectly correlated, 0 means that they are not correlated, and -1 means that they are strongly opposite.

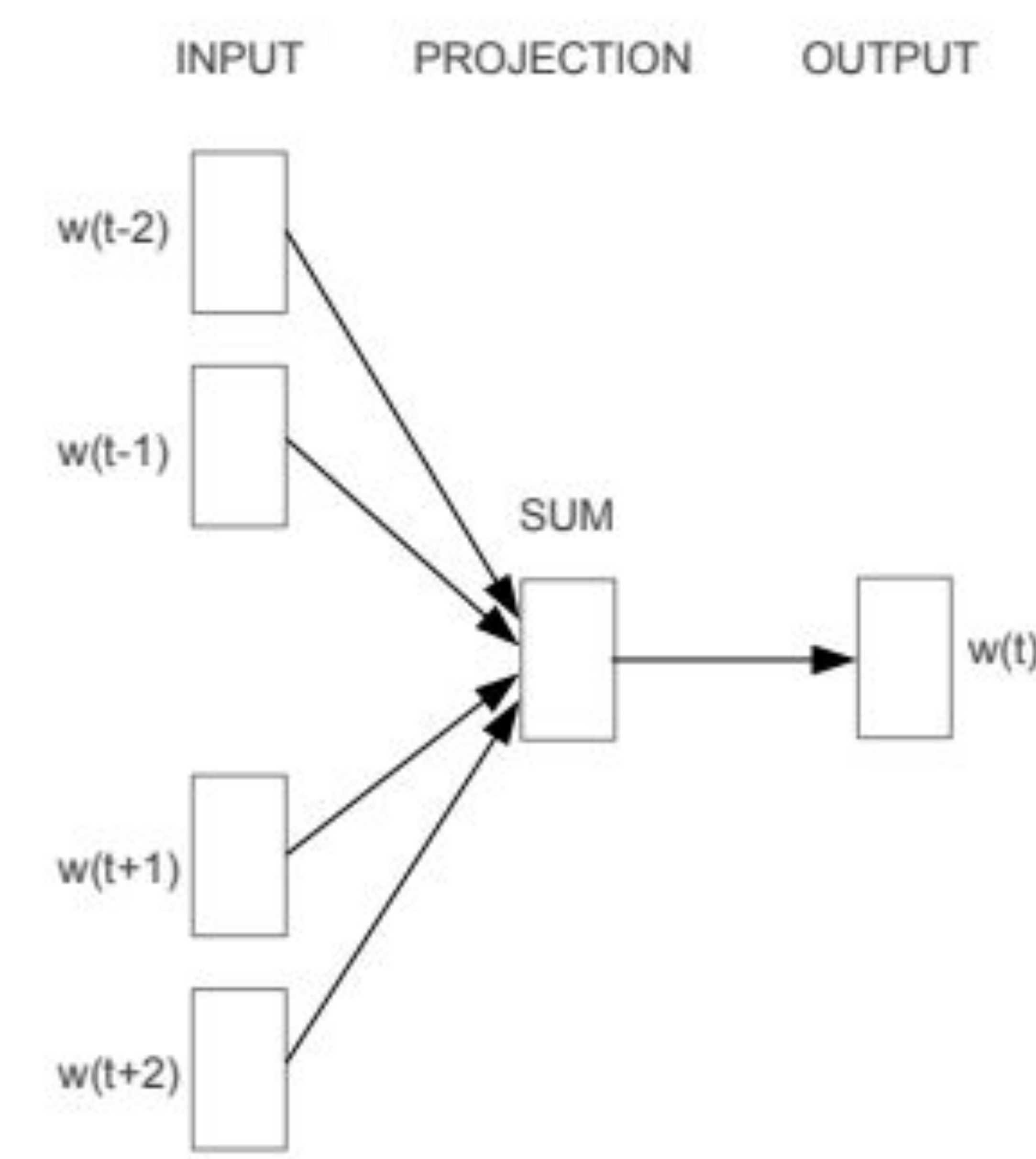
$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Word2Vec

- Word2Vec represents words as vectors based on several features, such as window size and vector dimensions.
- Similar words tend to have the same vector values and are grouped in the same block.

CBOW

- The method we used to calculate vector values was Continuous Bag of Words (CBOW)
- Surrounding words are combined to predict the word in the middle.



CBOW