Classifying Newsbooks from London, 1649

By Brand Leng, Claire Liu, Lauren Stampfli
Advisor: Joel Greenhouse

BACKGROUND

Goal: Characterize and Classify Newsbooks in order to identify counterfeit status

Research Question: Is it possible to develop a classification system for these newsbooks in terms of authorship and can this classification help reveal counterfeit status?

Setting: Newsbooks, or newspapers, published around 1649. This period was characterized by political upheaval with the recent conclusion of the English Civil Wars and execution of King Charles I.

- Important political factions included the Royalists, Parliamentarians, Levellers, and the New Model Army

DATA

Original Data:

- 1,180 Cleaned Newsbooks (610 from ~1649)
- Each Newsbook had additional metadata, including publishing date
- A subset of newsbooks had identified ideologies and/or authorship
- Many of the newsbooks had uniquely spelled words, as English was not standardized during this period

Mercurius Data Subset:

- Did not account for the uniquely spelled words
- Preformed analysis using a limited subset of the final features
- Used the upenn_tagset for part-of-speech tags

Processed Data:

- Created a non-exhaustive dictionary by looking at ~20 Newsbooks and mapping uniquely spelled words to what we believed was the most correct standardized spelling
  - This was done for a subset of the uniquely spelled words in these newsbooks
  - The correct spelling was applied only when performing calculations that involved part-of-speech tagging
- Stop words were removed from the texts for all calculations

Creating Dimensions:

- >300 Lexical Features in total
- Sentence Length, Word Length, Type Token Ratio (unique/total words), Lexical Density (content/total words), relative frequencies of top 5 most frequent function and non function words for each newsbook, part-of-speech relative frequencies
  - The part-of-speech was determined using the universal tag set and nltk pos tagger

Dimension Reduction

- PCA analysis was used to reduce dimensionality

METHODS

Working with subset

- Applied methods to a subset of newsbooks that had identified authors
  - Known authors helped us identify important contributing variables

Applied same analysis to full dataset

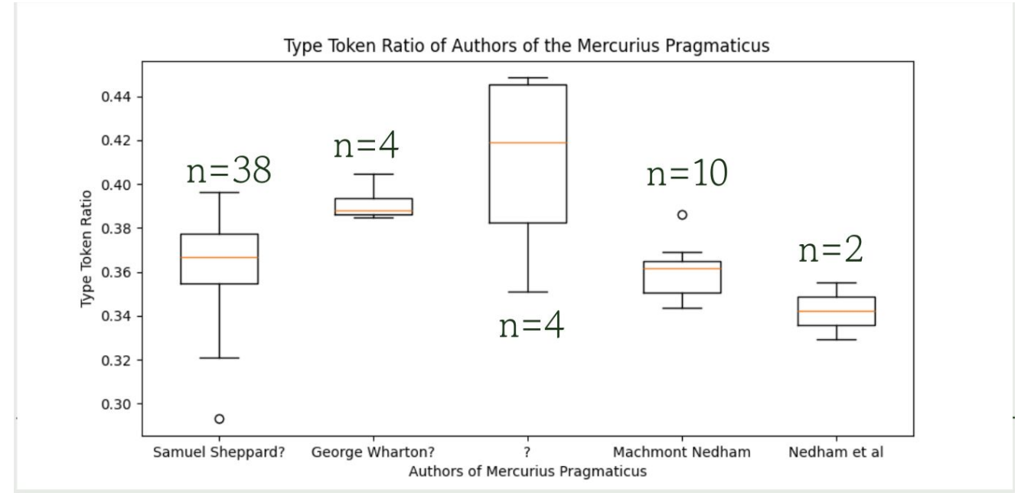
- Increased number of Textual Features and Improved upon old Textual Features
- Utilized Principle Component Analysis to reduce dimensionality in our data
  - Utilized the first 3 components (explaining 15.3% of the variance) in clustering analysis
- Utilized K-Means Clustering to create groupings of Newsbooks
  - Analyzed Newsbooks based on title, ideology, PC3 Dimension

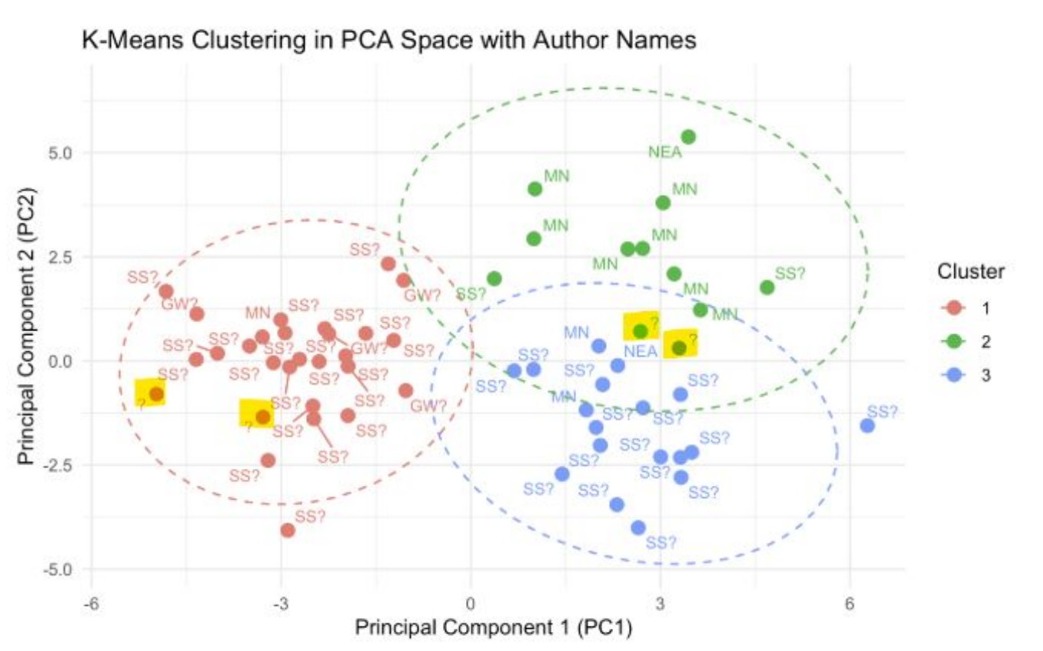
Topic Modeling

- Used Top2Vec to generate topics for individual clusters
  - The topics are represented as a vector of words
  - Top2Vec characterizes clusters by embedding words into a vector space, reducing dimensions, and identifying dense areas with center words to represent topics

ANALYSIS & RESULTS

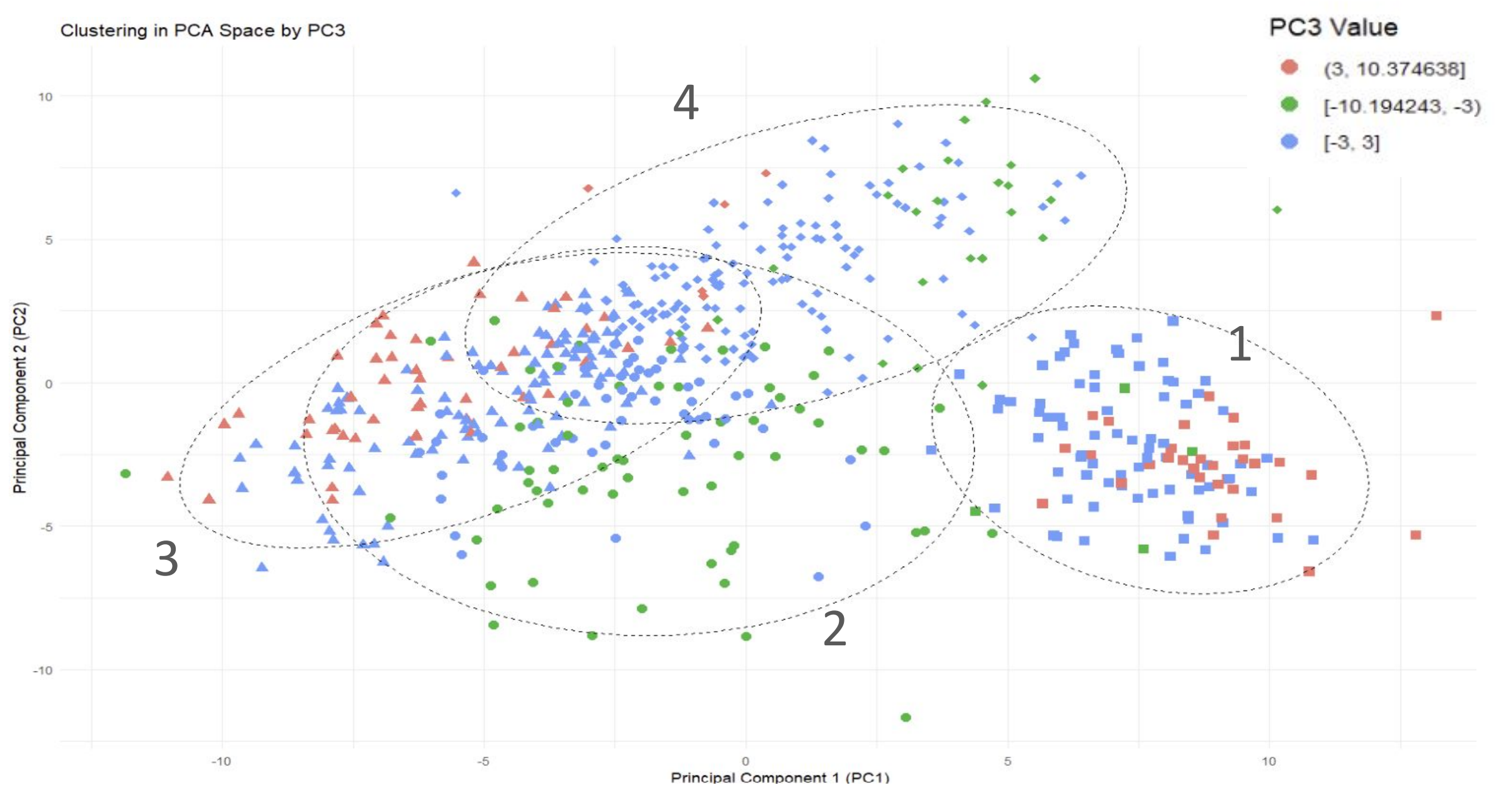
Mercurius Pragmaticus subset

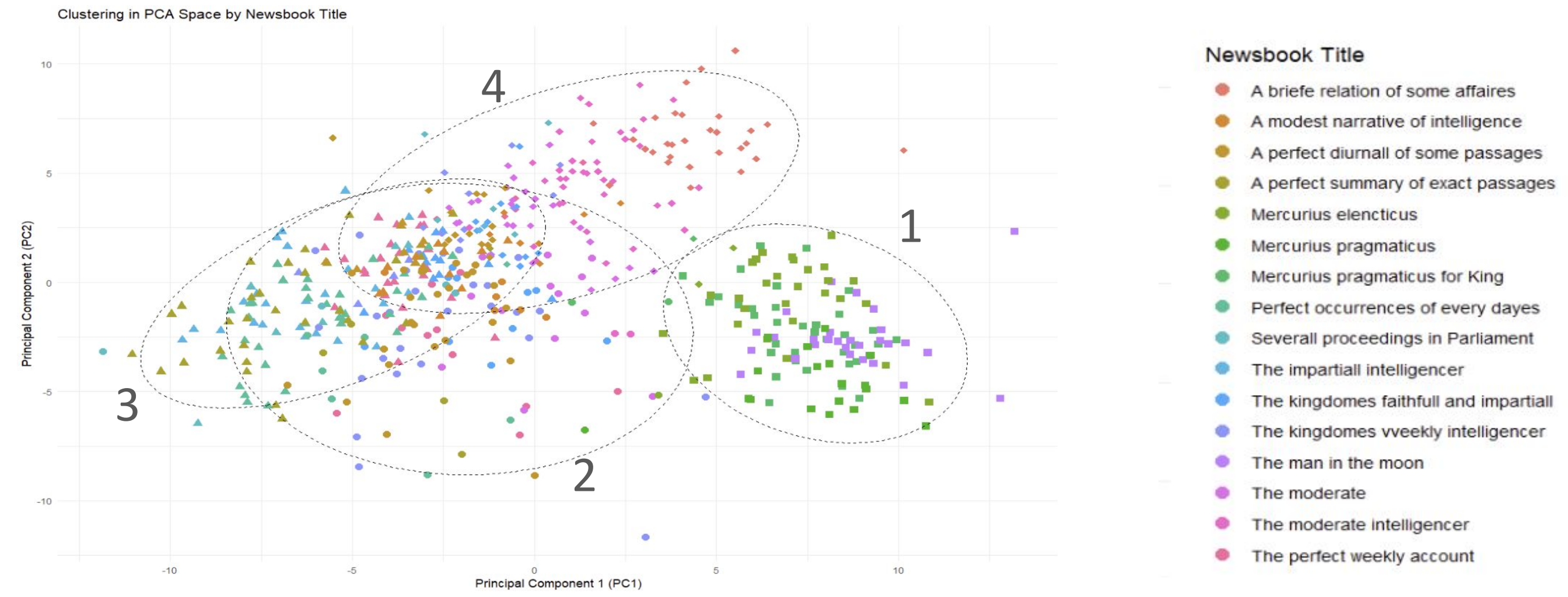




Initial study on the Mercurius Pragmaticus subset allowed us to classify by authorship and identify potentially useful measures for authorship.

K-Means Clustering





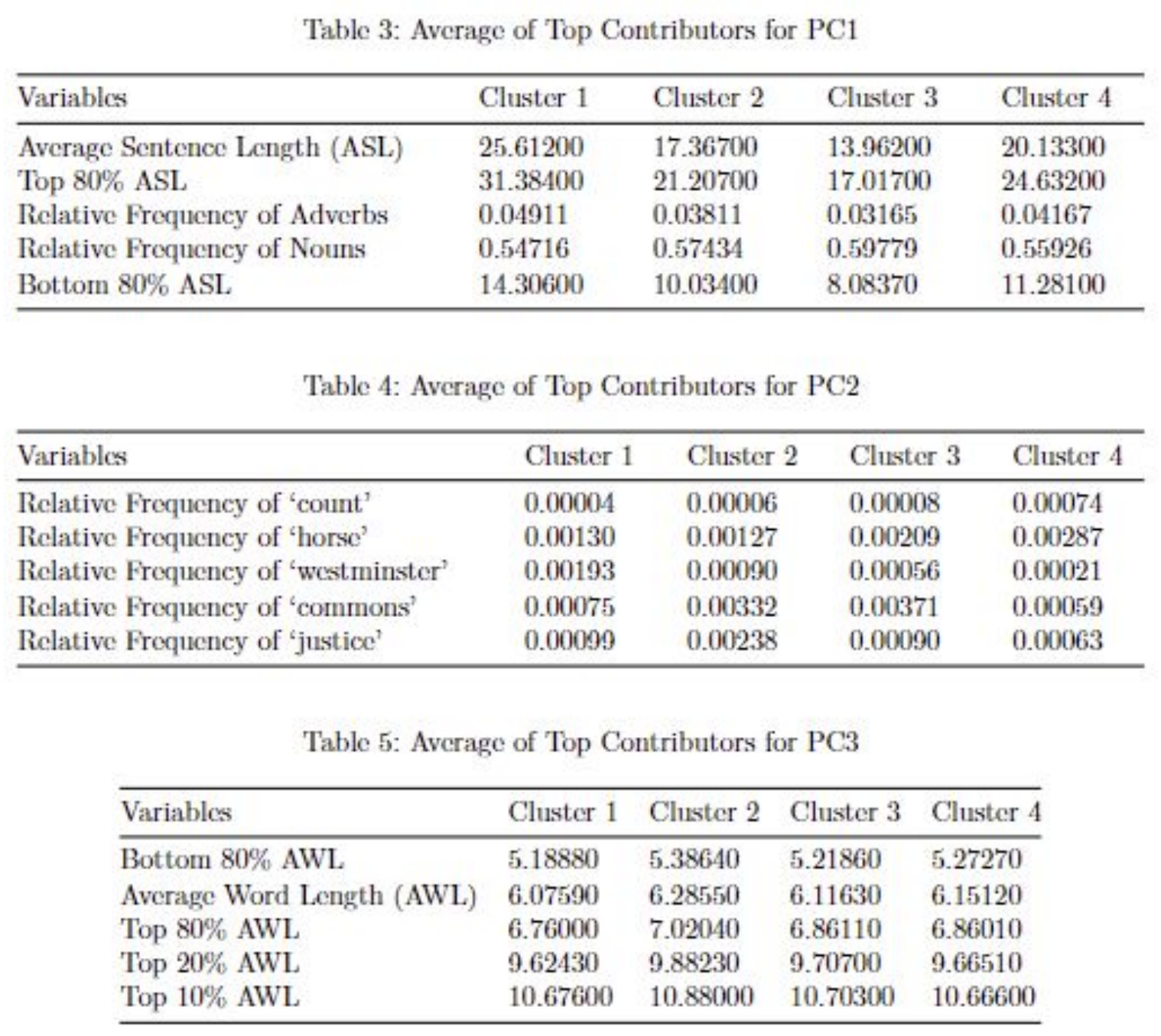


Table 3: Average of Top Contributors for PC1

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Average Sentence Length (ASL)	25.61200	17.36700	13.96200	20.13300
Top 80% ASL	31.38400	21.20700	17.01700	24.63200
Relative Frequency of Adverbs	0.04911	0.03811	0.03165	0.04167
Relative Frequency of Nouns	0.54716	0.57434	0.59779	0.55926
Bottom 80% ASL	14.30600	10.03400	8.08370	11.28100

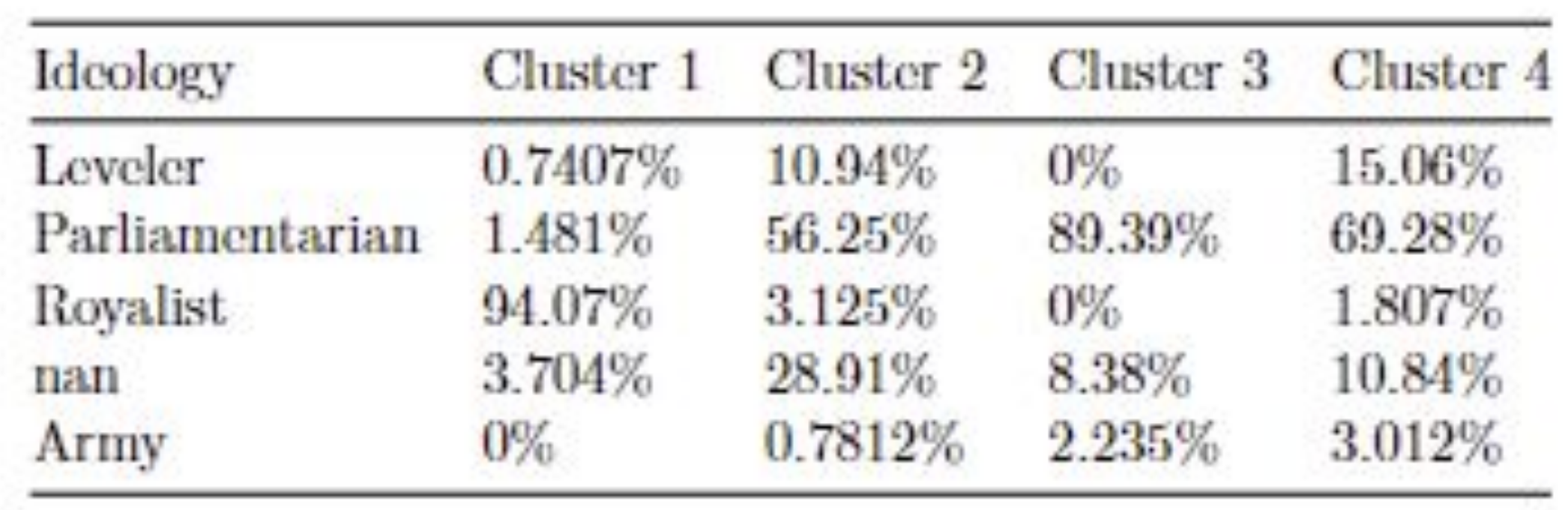
Table 4: Average of Top Contributors for PC2

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Relative Frequency of ‘count’	0.00004	0.00006	0.00008	0.00074
Relative Frequency of ‘horse’	0.00130	0.00127	0.00209	0.00287
Relative Frequency of ‘westminster’	0.00193	0.00090	0.00056	0.00021
Relative Frequency of ‘commons’	0.00075	0.00332	0.00371	0.00059
Relative Frequency of ‘justice’	0.00099	0.00238	0.00090	0.00063

Table 5: Average of Top Contributors for PC3

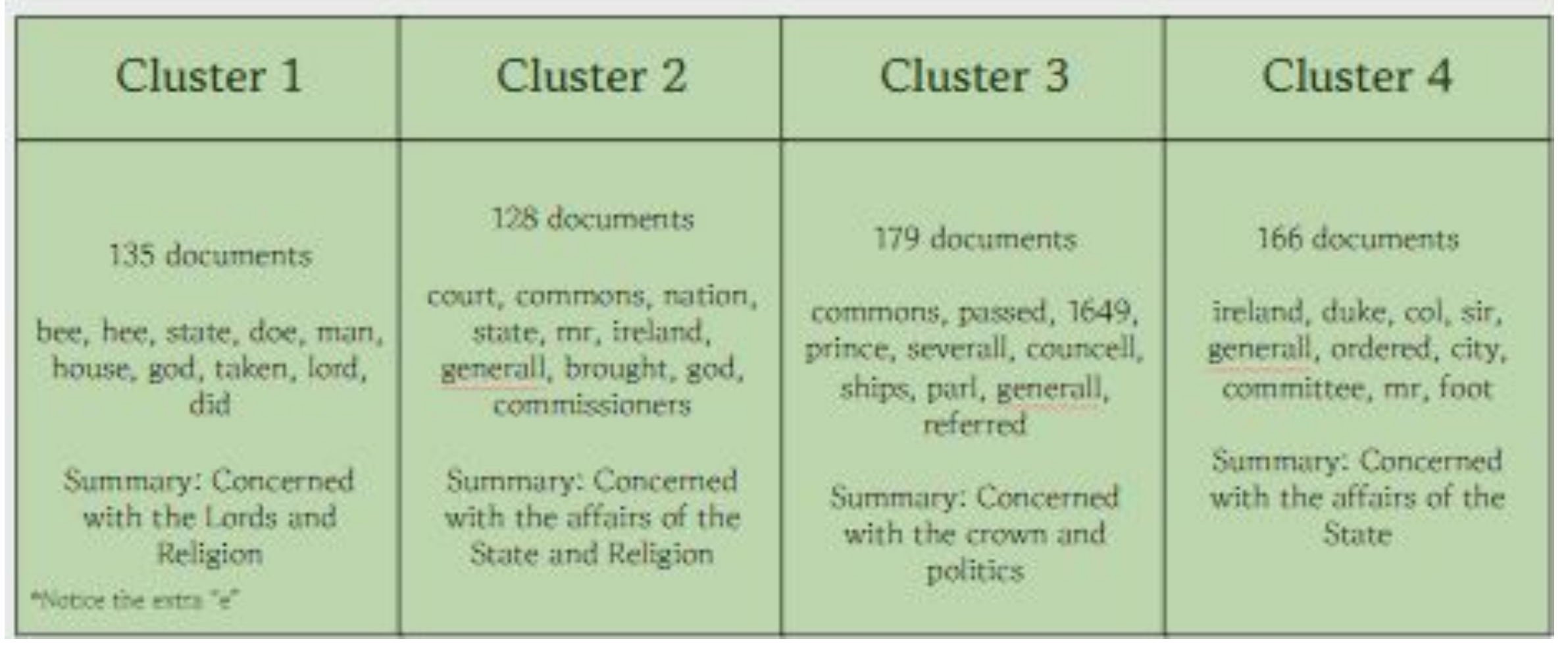
Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Bottom 80% AWL	5.18880	5.38640	5.21860	5.27270
Average Word Length (AWL)	6.07590	6.28550	6.11630	6.15120
Top 80% AWL	6.76000	7.02040	6.86110	6.86010
Top 20% AWL	9.62430	9.88230	9.70700	9.66510
Top 10% AWL	10.67600	10.88000	10.70300	10.66600

Ideological Percentage By Cluster



Ideology	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Leveler	0.7407%	10.94%	0%	15.06%
Parliamentarian	1.481%	56.25%	89.39%	69.28%
Royalist	94.07%	3.125%	0%	1.807%
nan	3.704%	28.91%	8.38%	10.84%
Army	0%	0.7812%	2.235%	3.012%

Cluster Topics From Top2Vec



Cluster 1	Cluster 2	Cluster 3	Cluster 4
135 documents bee, hee, state, doe, man, house, god, taken, lord, did Summary: Concerned with the Lords and Religion *Notice the extra "e"	128 documents court, commons, nation, state, mr, ireland, generall, brought, god, commissioners Summary: Concerned with the affairs of the State and Religion	179 documents commons, passed, 1649, prince, severall, councell, ships, parl, generall, referred Summary: Concerned with the crown and politics	166 documents ireland, duke, col, sir, generall, ordered, city, committee, mr, foot Summary: Concerned with the affairs of the State

CONCLUSION

- The Newsbooks can be clustered based on lexical and grammatical features
  - Potentially interesting features include sentence length, word length, frequency of adverbs and nouns, and the relative frequency of a subset of commonly or uncommonly used words
  - Interestingly, author preference for spelling can also be a potential determining factor for authorship of newsbooks
- Newsbook topics differ, potentially indicating a slight difference based on ideological tilt
- Next Steps:
  - Creating a complete dictionary for uniquely spelled words : standard english words
  - Investigating Author spelling preference as a potential indicator
  - Focusing on individual newsbooks to create smaller clusters to better indicate authorship or non authorship

REFERENCES

Angelov, Dimo. 2020. “Top2Vec: Distributed Representations of Topics.” https://arxiv.org/abs/2008.09470

“Charles I of England.” *Wikipedia*, Wikimedia Foundation, 16 Apr. 2025, en.wikipedia.org/wiki/Charles_I_of_England.

Chunxia Zhang, Xindong Wu, Zhendong Niu, Wei Ding, Authorship identification from unstructured texts, Knowledge-Based Systems, Volume 66, 2014, Pages 99-111, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2014.04.025.

Peacey, Jason. “‘The Counterfeit Silly Curr’: Money, Politics, and the Forging of Royalist Newspapers during the English Civil War.” Huntington Library Quarterly 67, no. 1 (2004): 27–57. https://doi.org/10.1525/hlq.2004.67.1.27.