Reuters Write-up

Question: How does complexity vary by authors and categories?

I wanted to analyze our text data to understand how complexity of a writing style varies by author. I also investigated if this writing style was influenced at all by the category of text.

Approach: Analyzing writing styles through complexity measures and clustering

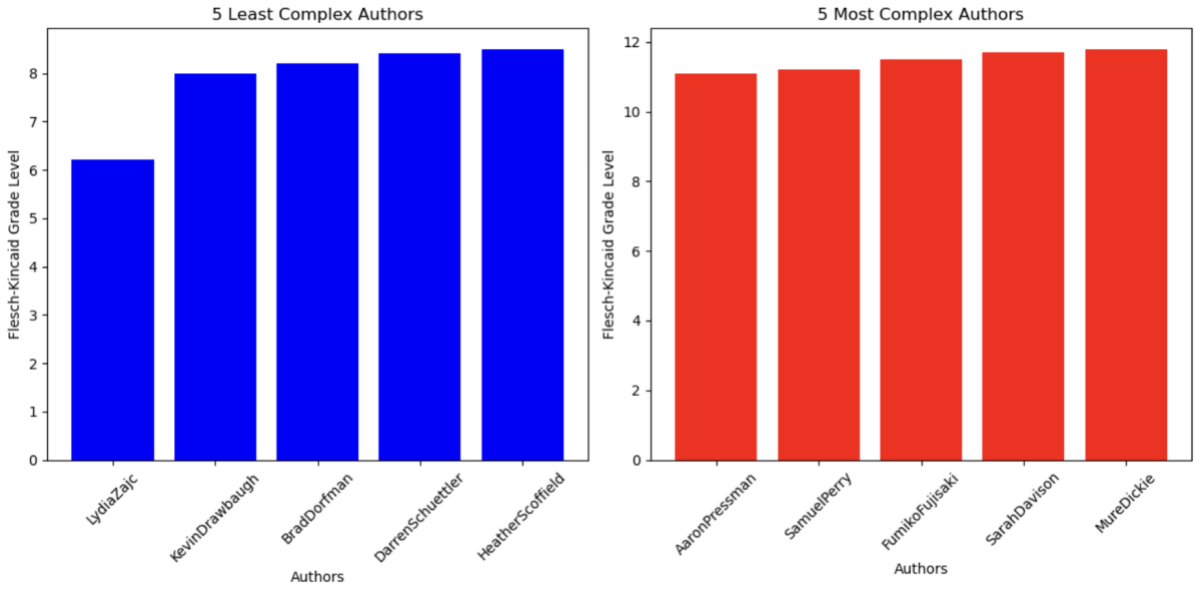
I processed the data to group each text file by author and gathered various complexity metrics (Flesch Kincaid Grade, TTR, Average Sentence length, Dale-Chall, Lexical Diversity, etc.) for each text file. I then averaged these scores across all text files, for each author, to get an average complexity score. Lastly, I did K-means clustering to further analyze complexity metrics.

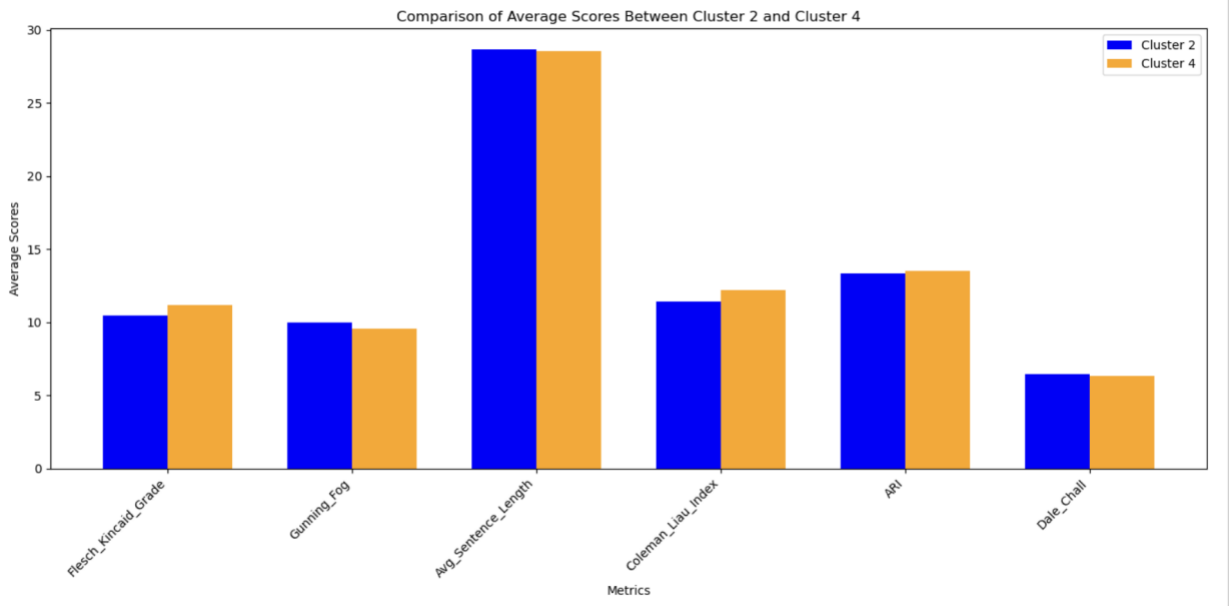
For my categories, I iterated through each text file, added the key words per category, and then saved the top 5 categories with the most key words (there are better ways to do this, but this was more simple and computational efficient method). I then took the average complexity scores per category and then did K-means clustering.

Results: Large readability range for authors and unique category clusters

Author Complexity:
When analyzing author complexity, we see a large range of writing complexity levels. Our least-complex author writes at a 6th grade level while our most complex author writes at a 12th grade level. We also see that our clusters differ on some complexity metrics, in interesting ways. There can be a 2-grade level gap between clusters, and yet the lower-cluster may have a slightly higher Dale-Chall score (a measure of the quantity of difficult words in a text). This highlights the unique nature of writing styles, that more complex writing doesn't always involve more difficult words, or longer sentences.

Top 5 Least- Complex Authors and Top 5 Most-Complex Authors

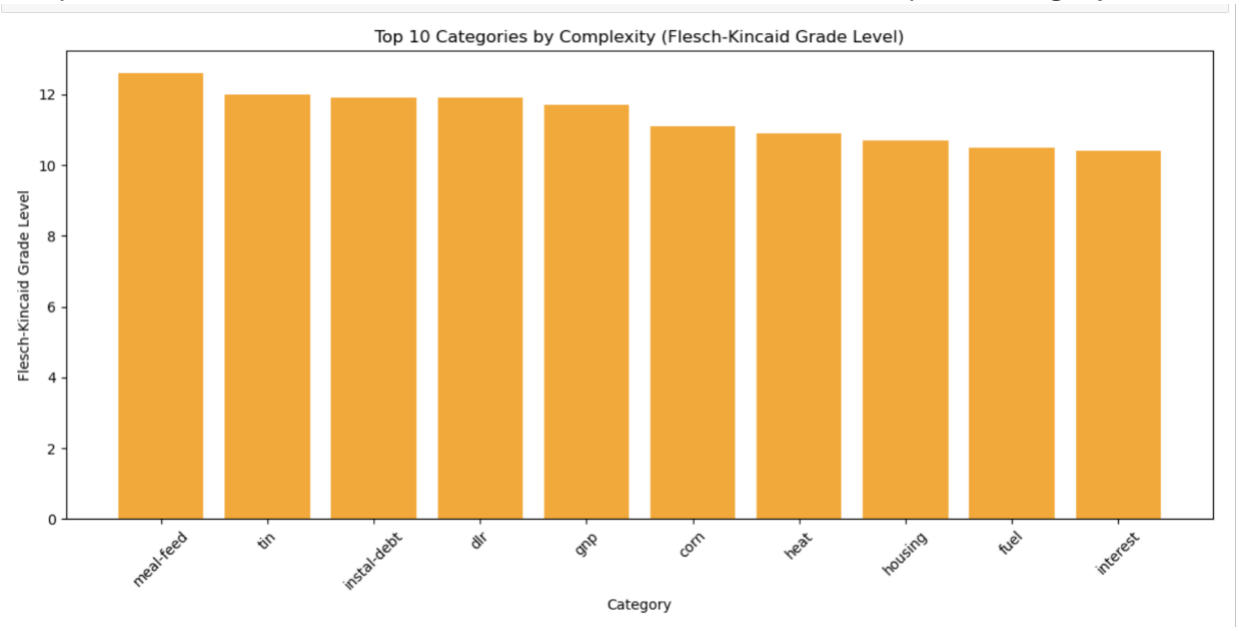


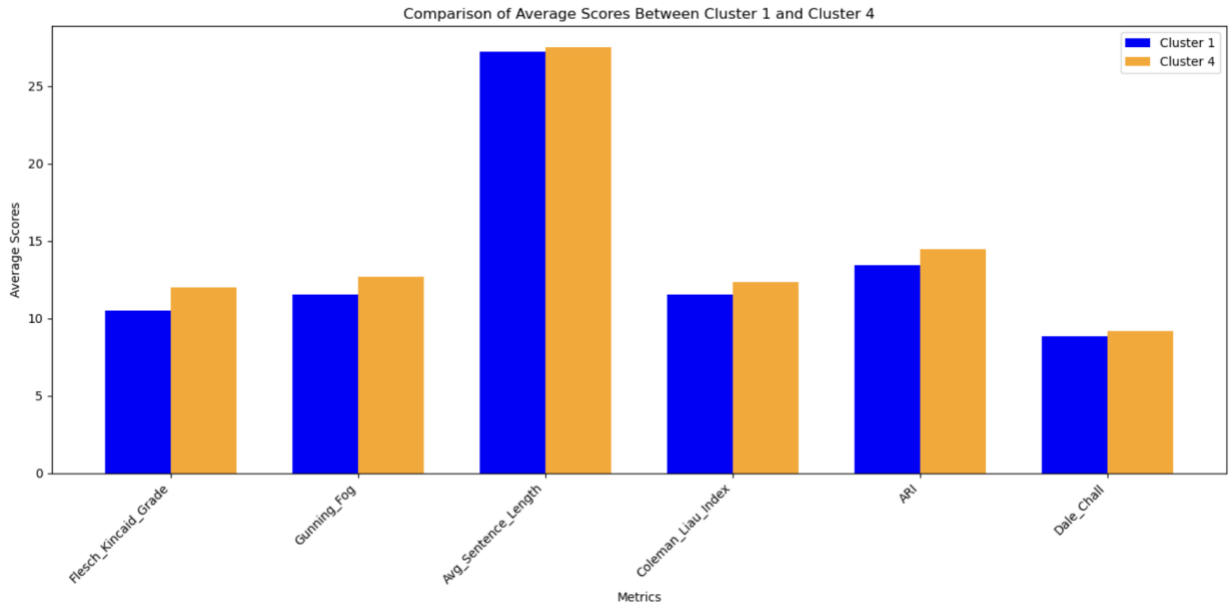


Category Complexity:

Our category method provides less-conclusive results, as we see our categories with the highest complexity scores are 'meal-feed', 'tin', and 'install debt'. Our clustering analysis shows that we have a variety of categories in our clusters, as we see a cluster with a good

mix of finance and agriculture categories. This could mean that these fields are highly complex, or authors that tend to write about these fields favor complex writing styles.





Edit Metadata

```
category_metrics[category_metrics['Cluster']==2]
```

	Category	Flesch_Reading_Ease	Flesch_Kincaid_Grade	Gunning_Fog	TTR	Avg_Sentence_Length	Syllables_per_Word	Coleman_Liau_Index	ARI	Dale_
0	acq	60.95	9.4	7.87	0.040353	27.176390	1.335715	11.49	12.2	
1	alum	62.17	8.9	8.22	0.141985	26.348355	1.299824	11.02	11.5	
2	barley	60.65	9.5	8.79	0.124467	26.276916	1.327409	11.37	12.2	
3	bop	62.88	8.7	8.35	0.185457	29.257971	1.299188	11.02	11.2	
4	carcass	61.87	9.1	8.79	0.185116	27.498708	1.323811	10.79	11.3	
5	cocoa	59.64	9.9	10.27	0.264948	23.882096	1.312854	11.78	13.0	
7	copper	59.84	9.8	8.37	0.043983	27.298793	1.346261	11.55	12.7	
9	cpu	54.73	9.7	8.63	0.207465	25.977901	1.350489	11.66	11.6	
10	crude	65.22	7.8	8.00	0.292202	21.813084	1.331620	10.90	10.1	
12	earn	61.36	9.2	9.23	0.201485	26.707692	1.316628	11.26	11.8	
15	gold	70.43	7.8	9.29	0.459236	23.071429	1.297214	9.80	10.3	

Conclusion: Segment Reuter's readers and assess individual author's complexity scores

If Reuter's is trying to expand their customer base, they can segment readers by category and complexity level. They then can assess if a large gap exists between the complexity score of current articles and the complexity score desired by readers. Reuters can train authors to write less-complex articles and capture a larger customer pool.