Natural Language Processing Library

Sreevatsa Nukala

DS3500

1 Introduction

For this project, I am working alone and am exploring a topic that I, personally, have much interest in. I am passionate about bioinformatics but seeing as this project is related to textual analysis, I am deciding to look at how CRISPR, a recent biotechnology used for genome editing, is portrayed in formal scientific papers as well as in articles. I will explore the differences in papers and articles, as they both discuss the same topic, yet are attempting to reach differing audiences. I will investigate various other relationships and discrepancies in the texts that could be useful to a holistic understanding of the topic and the writing regarding the topic. Along with these comparisons, this project will investigate the methodology of natural language processing analyses, in order to uncover the best practices and techniques. All the code for my library and analyses can be found here: GitHub Repository

2 Data Sources

I will analyze and compare 3 research papers (1) (2) (3)and 3 articles (4) (5) (6) about CRISPR-based gene editing. To find research papers to use as texts, I form a basic PubMed query for related papers and pick the most highly cited and recent papers. I use both criteria, as I want to make sure the texts are relevant to today's conversation on CRISPR and have had some measured impact on the community. I am able to download .pdf files of these research papers, so I will have to create a custom parser to process these files. For articles regarding CRISPR, I use a a simple Google search of the topic and pick the first 3 articles I see. Here, I rely on Google for relevant and significant articles, as their search engine automatically sorts by both factors. For the articles, I have .txt files using an online text extractor. While I'll still have to make a custom parser for .txt files, it will be much more reliable and simple compared to a .pdf file parser.

3 Insights

My first visualization in Fig. 1 is the most simple and yet provides useful information into the properties of the selected texts. This visualization illustrates a comparison between the number of words in each text. This number correlates to the original, uncleaned versions of the text. It shows that the number of words in the research papers is far higher than in the articles, possibly meaning that more information and details about CRISPR are provided in the papers than in the articles, as expected.

The second visualization in Fig. 2 consists of 2 subplots, displaying some basic statistics about the texts. However, even these basic visualizations can prove worthy in understanding key differences in the presentation of information between academic papers and online articles. The bar chart for average sentence length demonstrates a pattern of shorter sentences in research papers compared to articles. This result contradicts my initial thought that papers would have substantially longer sentences. Some reasoning behind this could be that the sentences in research papers avoid unnecessary wording, being more straightforward in nature, while articles may contain more flowery language and 'fluffed' sentences. The second subplot for average number of unique words per 1000 words reveals the distinctness of each text. Here, we see only a slightly varying amount of unique character in each text. This could be the product of explaining a scientific topic, as there is more related jargon.

Fig 3 shows a comparison between how each text abides by Heaps' law. Heaps' law means that as more instance text is gathered, there will be diminishing returns in terms of discovery of the full vocabulary from which the distinct terms are drawn. We know from Fig. 2 that the extent of the vocabulary in each text is about the same. Fig 3 helps to reiterate this idea and validate the text processing of the library. I calculated Heaps' law for each text after cleaning and tokenizing the text. Seeing that Heaps' law still applies by comparing the points (in various colors) with the regression line (in black), we can be sure that our pre-processing does not totally remove meaning from the texts.

In our sentiment analysis of the original texts in Fig. 4, we can see that neither the articles nor the papers are particularly polar. This could be because the papers and articles discuss both the benefits and issues with CRISPR. Only Article 3 has a polarity above 0.1 and seems to be positive about CRISPR than not. This is not seen in any of the papers, as they seem to be more neutral. The fact that all the texts have a subjectivity over 0.35 is surprising, as I thought because they are discussing a scientific topic, the information would be much more objective. Even the research papers display this behavior. This could stem from the fact that CRISPR is still controversial and all of these texts simply look to contribute to that conversation about the use of gene editing.

Fig. 4's second subplot shows the sentiment analysis but with cleaned versions of the texts. This cleaning includes removing special characters, numbers, stop words, and lemmatizing (or reverting back to their root forms) all the words. The cleaned text shows significantly lower polarity and subjectivity scores for all texts while maintaining the same differences and patterns of sentiment between texts.

Fig. 5 is a Sankey diagram illustrating the relationship between each text and the k most common words shared across texts. We used k=7 in the included visualizations. This Sankey diagram is based on absolute frequency of words and subsequently shows the research papers to have significantly thicker lines (higher count) to the most common words. This can be explained by the enormous gap in number of words between the papers and articles, as shown in Fig. 1. Thus, I created an additional Sankey diagram that calculated the list of most common words based on relative frequency in each text.

This diagram in Fig 6 calculate line thickness as proportion of word appearances in the enirety of that text. This Sankey diagram shows each text to have very similar proportions of usage across the more common words. The list of most common words only differs by one word. The first diagram includes ”dna”, while the second one includes ”human”. Perhaps if we used a larger k value, we would see a difference in the list of most common words. Across both Sankey diagrams, gene seems to be the most common word.

4 Conclusion

The natural language processing framework created during this project was used to analyze 3 academic research papers and 3 online articles regarding CRISPR gene editing. The analyses, while not incredibly complex, provided many insights into the discrepancies between presentation of information in papers and articles. The analyses also investigate the best methodology for textual analyses and natural language processing techniques themselves.

Taking the results of both subplots in Fig. 2 together, we can see that although the average sentence length heavily differs between research papers and articles, they use approximately the same amount of unique words, meaning the same general type of information presentation is the same. In Fig. 4, the comparison between the sentiment analyses of the original and cleaned texts highlights how important things like stop words and context in the form of prefixes and suffixes add to the sentiment of texts. Fig 5 and Fig 6 demonstrate that the most common words across texts don’t correlate to relative frequency when the texts are all about the same topic. Thus, we can make Sankey diagrams based on absolute frequency of words when comparing between texts of the same topic.

While articles on CRSIPR seem to have more flowery language and may be more subjective on the topic, they seem to provide the same general type of information as academic papers. Thus, when looking for information on CRISPR, articles may be a good starting place for basic information before getting into details buried in research papers.

References

[1] R. Barrangou, “The roles of crispr–cas systems in adaptive immunity and beyond,” *Current Opinion in Immunology*, vol. 32, pp. 36–41, 2015, innate immunity. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952791514001563

[2] F. B. Ayanoğlu, A. E. Elçin, and Y. M. Elçin, “Bioethical issues in genome editing by CRISPR-Cas9 technology,” *Turk. J. Biol.*, vol. 44, no. 2, pp. 110–120, Apr. 2020.

[3] P. Hsu, E. Lander, and F. Zhang, "Development and applications of crispr-cas9 for genome engineering," *Cell*, vol. 157, no. 6, pp. 1262–1278, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867414006047

[4] S. Sufian, "The dark side of crispr," *Scientific American*, Feb 2021. [Online]. Available: https://www.scientificamerican.com/article/the-dark-side-of-crispr/

[5] "Full stack genome engineering," *Synthego*. [Online]. Available: https://www.synthego.com/learn/crispr

[6] L. Warneck-Silvestrin, "The promises of crispr genome editing in biomedicine," *Labiotech.eu*, Mar 2021. [Online]. Available: https://www.labiotech.eu/interview/crispr-therapeutics-genome-editing/

Figure 1: Number of Words per Text

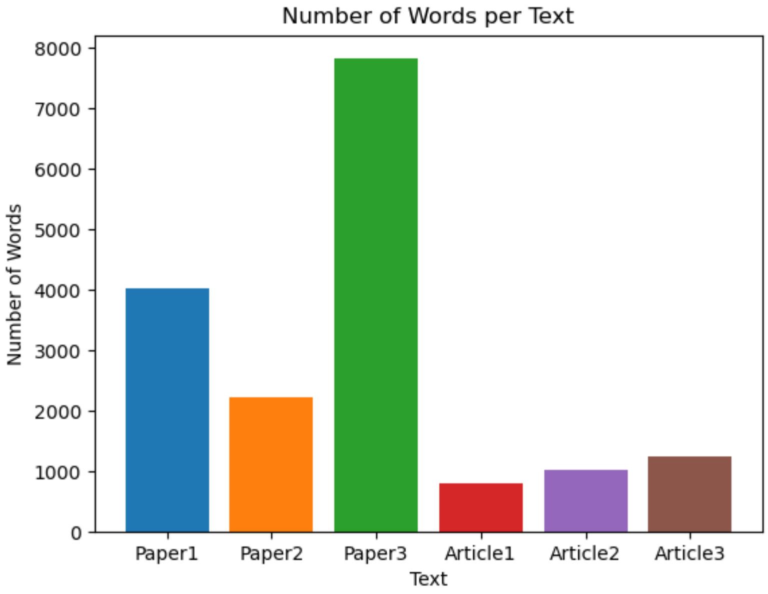


Figure 2: Average Sentence Length and Average Number of Unique Words per 1000 Words Comparisons

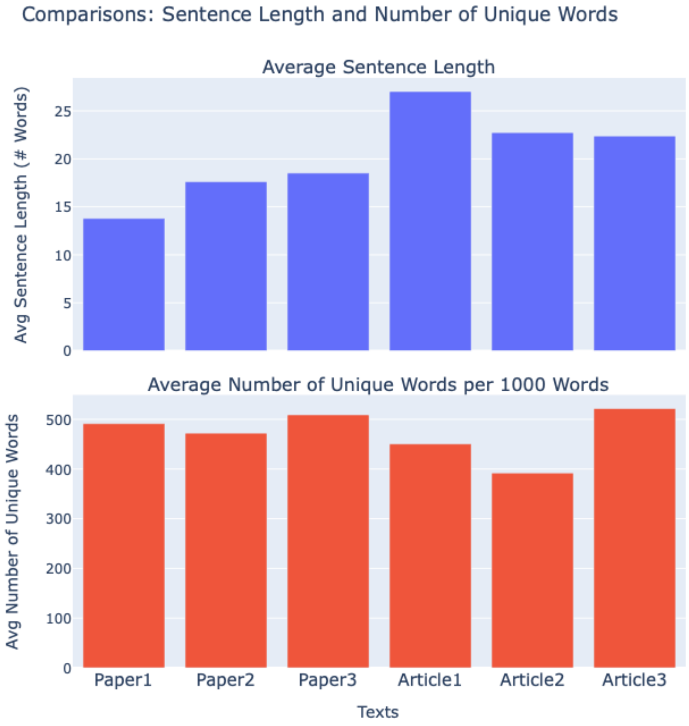


Figure 3: Heaps' Law for Each Text

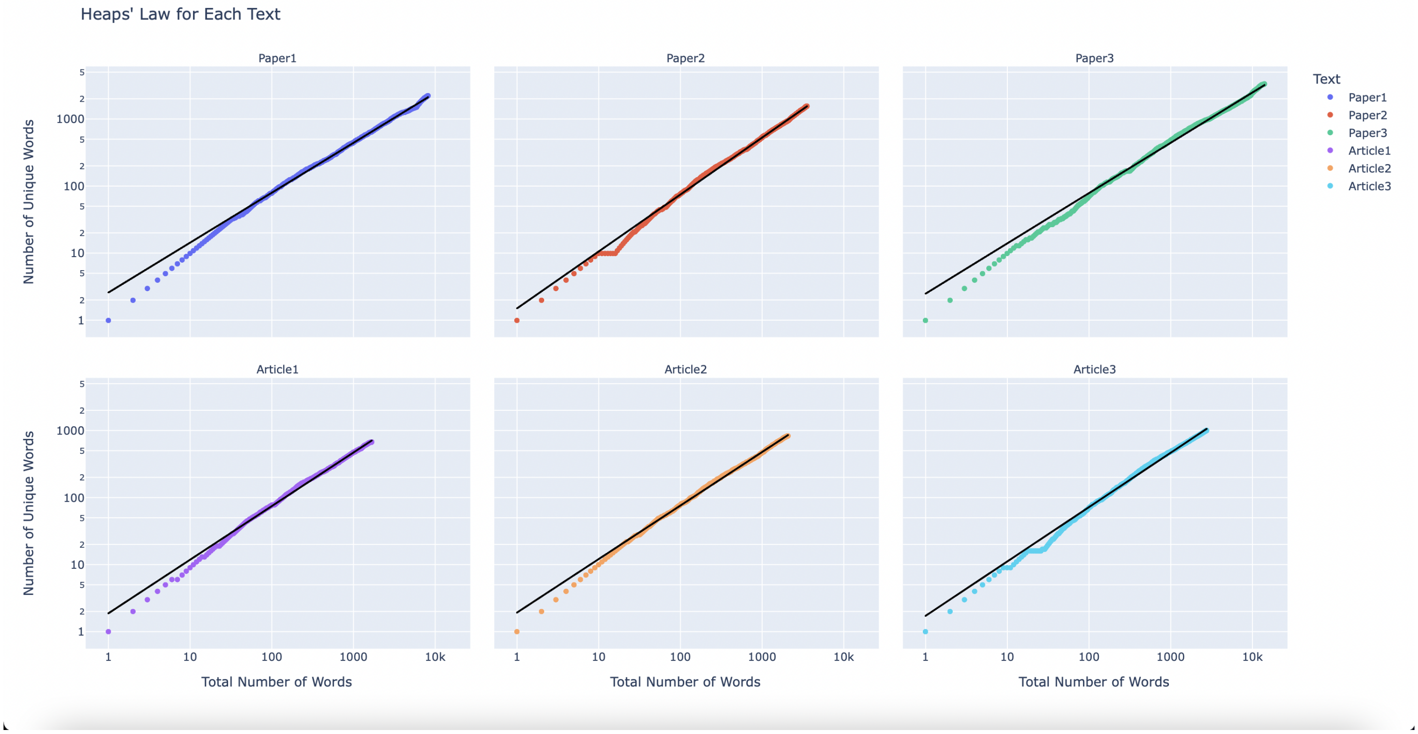


Figure 4: Sentiment Analysis on Original and Cleaned Texts

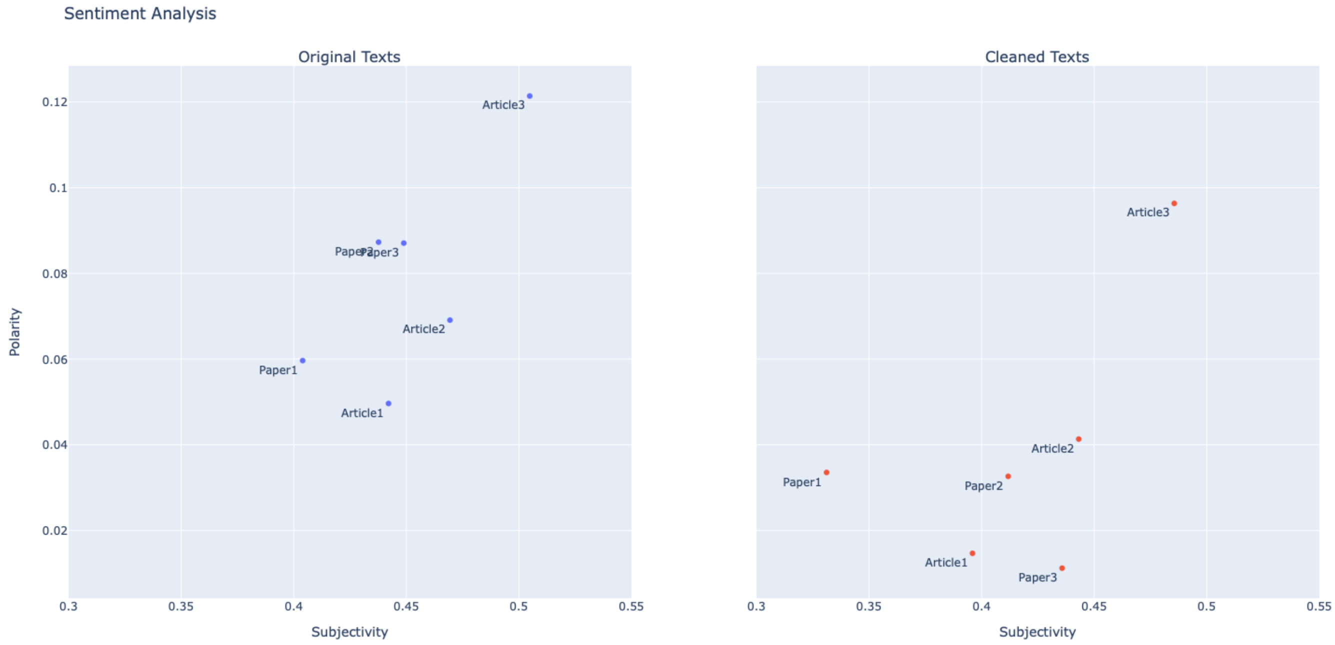


Figure 5: Text to Most Common Words Relationship

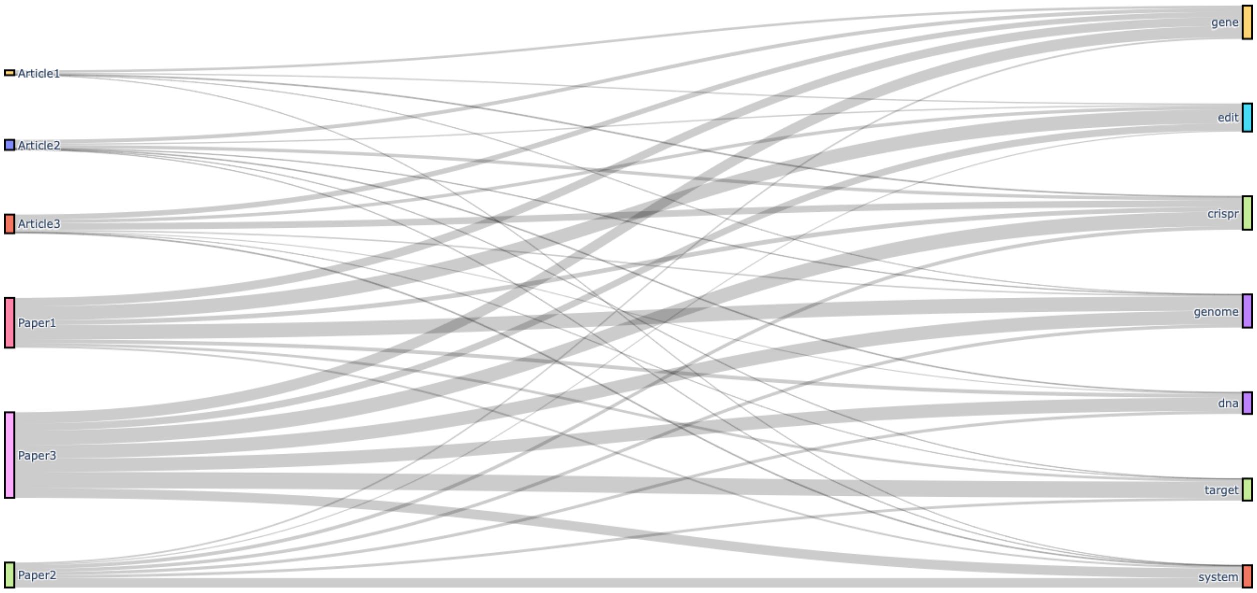
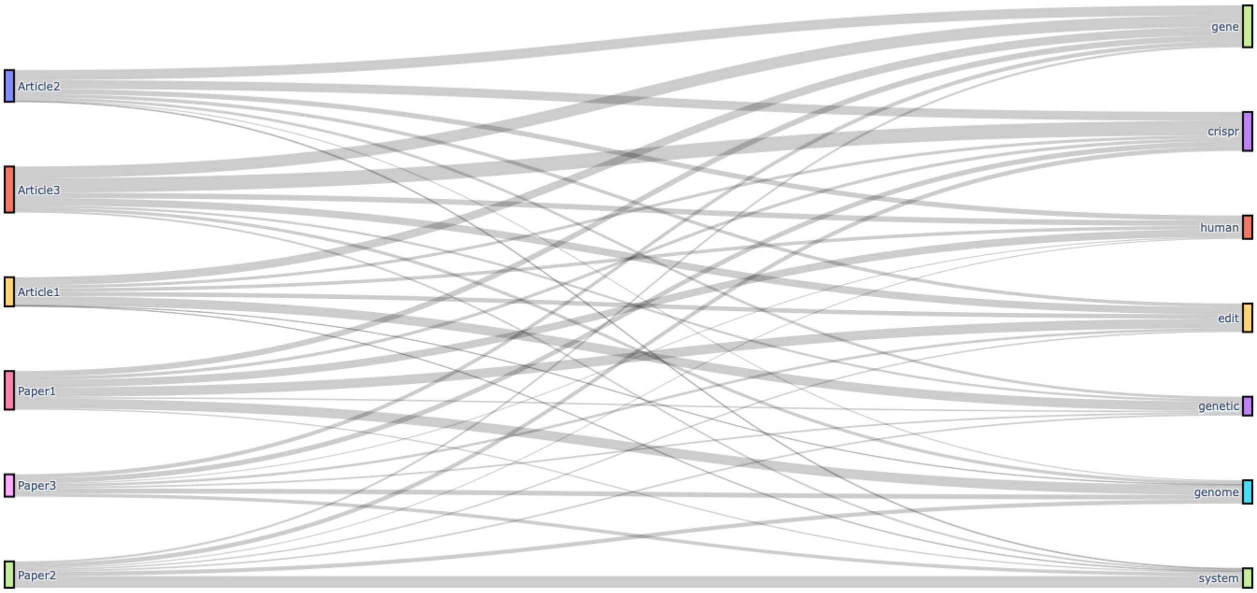


Figure 6: Text to Most Common Words Relationship Based on Relative Frequency


Article2
Article3
Article1
Paper1
Paper3
Paper2
gene
crispr
human
edit
genetic
genome
system