Brian Reicher

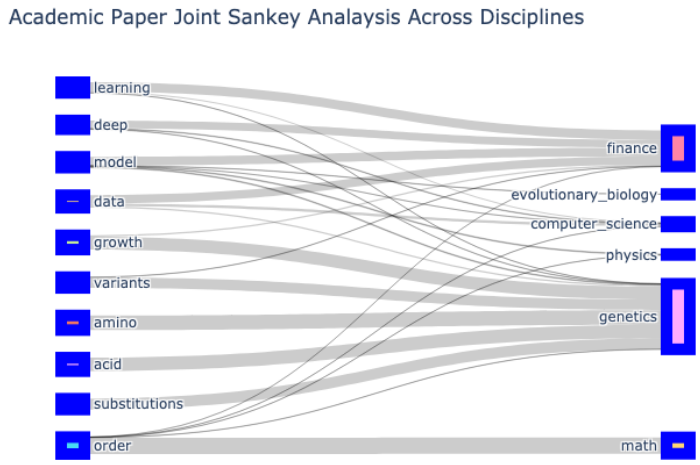
Professor Rachlin

DS3500

27 October 2022

NLP Analysis of Academic Research Papers

For my NLP framework analysis, I decided to observe patterns/disparities in text across scientific research papers in different disciplines. As a recently published author myself, I wished to compare my publication to see how varying fields of science interact with one another. Papers in the fields of physics, math, computer science, computational finance, evolutionary biology, and genetics were selected for this analysis. I first generated a sankey diagram linking each paper's shared 10-most common words:

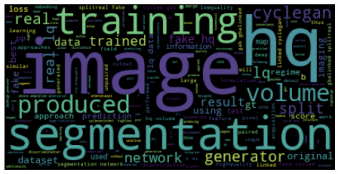


Here, I notice the shared relations through almost all papers regarding computing words like data, learning, and model. This use of software-oriented language across fields makes sense, given that technological terminology like such are inherently far-reaching in today's academic

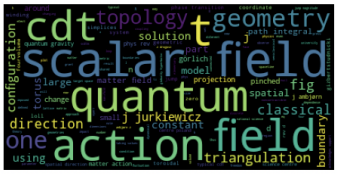
world. This goes to show the interdisciplinary nature of science, specifically how without technology, cs/physics/bio/genetics would lack significant resources. Additionally, I find it curious that a field like mathematics shares no commonalities with such other papers, which leads me to infer the isolation and vast differences which pure math holds from many other sciences.

I next looked to generate topic-based word clouds, as follows:

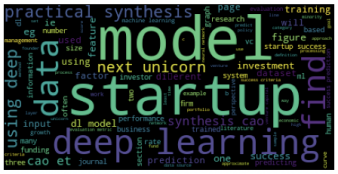
computer_science wordcloud!



physics wordcloud!

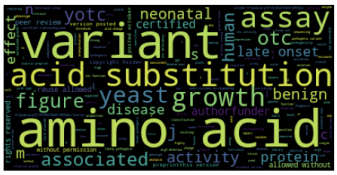


finance wordcloud!

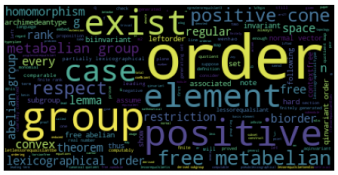


evolutionary_biology wordcloud!

genetics wordcloud!

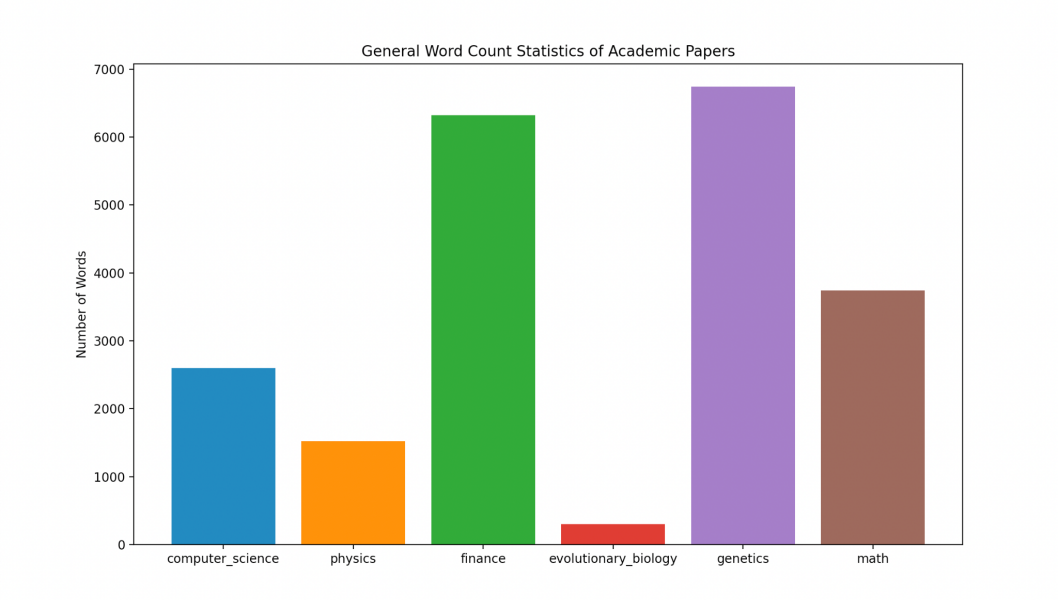


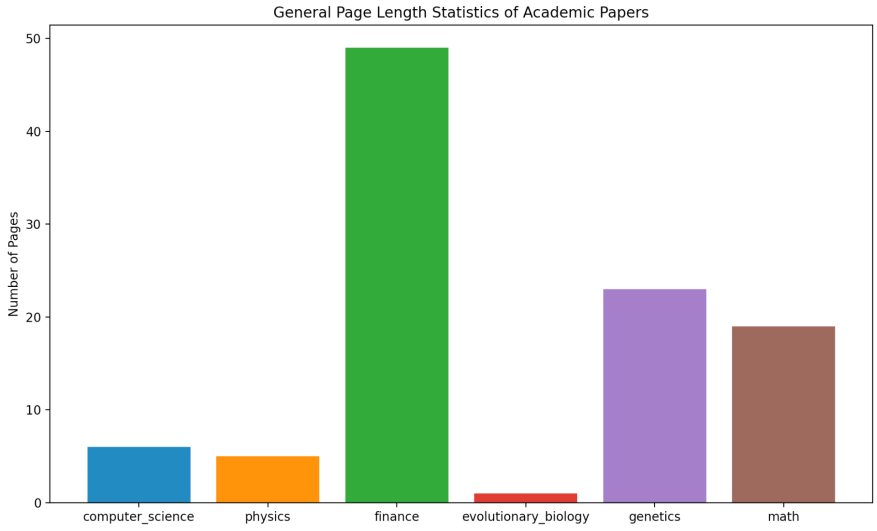
math wordcloud!



Here, we are able to see words that discipline-specific keywords (i.e. amino acids in the genetics word cloud, group & order in the mathematics word cloud) display most prominently. In the CS image, we see words like “training,” “volume,” and “real” that all denote topics in machine learning, which we can infer this paper is about. Also, we can still see tech-based words rise up in biological, physical, and financial papers. These all endorse the suggestions from the previous sankey diagram, and lead us to understand the critical importance of technology across different disciplines.

Next, I hoped to look generally at some basic summary statistics for each given field's paper, specifically the number of pages and words.

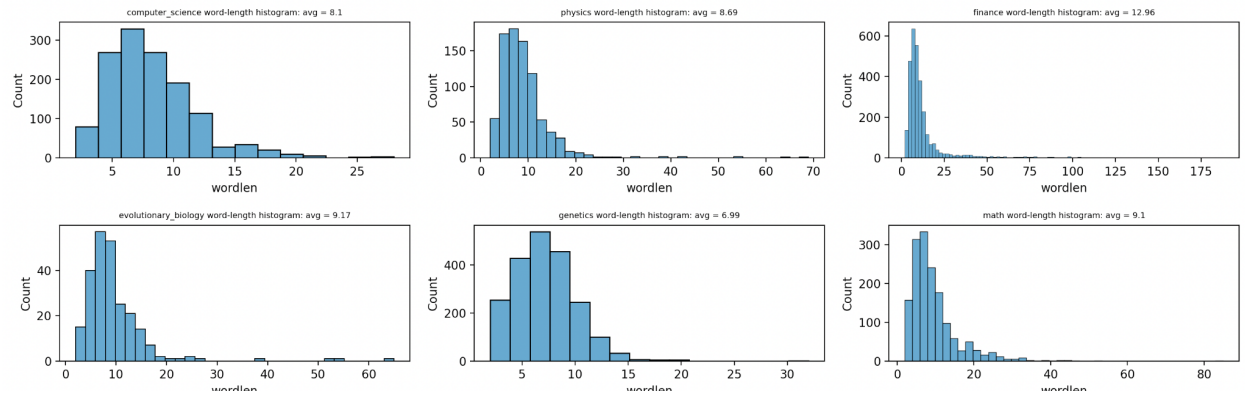




Albeit from a small and random sample size, we can see the finance paper stands out from other fields. Despite having a total word count which is within range of several other papers (i.e. genetics math, cs), we can see that the overall number of pages in the finance paper is drastically higher (almost 50, opposed to the second highest value of 25 pages). This most likely is because

many finance/economics papers contain multitudes of figures, tables, and graphs of data. Although mathematically based, finance papers likely contain less proof-based figures/supplements which would take up less space and contribute as words rather than images. We can see this in the physics and math papers, which still contain between 15000-30000 words, but both are less than 10 pages long.

For my final visualization, I wanted to look at overall paper word length counts and paper averages. Here, I created individual histograms for each file:



We can see that the highest average word count for a paper came from the finance histogram, at almost 13. This might be due to the fact that a financial paper holds long lists of stock tickers/investment logs that might be interpreted as massive strings of words. We can also see that the genetics paper holds the lowest average word length. We can attribute this to the fact that the names of genomes are typically small strings of character that would occur often in this paper and consequently lower the overall average.

References:

1. Ambjørn, J., et al. “Matter-Driven Change of Spacetime Topology.” *ArXiv.org*, 14 Sept. 2021, https://arxiv.org/abs/2103.00198.
2. Barroso, Gustavo Valadares, and Julien Y Dutheil. “Mutation Rate Variation Shapes Genome-Wide Diversity in Drosophila Melanogaster.” *BioRxiv*, Cold Spring Harbor Laboratory, 1 Jan. 2022, https://www.biorxiv.org/content/10.1101/2021.09.16.460667v2.
3. Cao, Lele, et al. “Using Deep Learning to Find the next Unicorn: A Practical Synthesis.” *ArXiv.org*, 18 Oct. 2022, https://arxiv.org/abs/2210.14195.
4. Cao, Lele, et al. “Using Deep Learning to Find the next Unicorn: A Practical Synthesis.” *ArXiv.org*, 18 Oct. 2022, https://arxiv.org/abs/2210.14195.
5. Lo, Russell S, et al. “The Functional Impact of 1,570 SNP-Accessible Missense Variants in Human OTC.” *BioRxiv*, Cold Spring Harbor Laboratory, 1 Jan. 2022, https://www.biorxiv.org/content/10.1101/2022.10.26.513893v1.
6. Rhoades, Jeff L, et al. *On the Topic of Unpaired Image and Segmentation Enhancement of X-Ray Holographic Nanotomography for Connectomics*. International Symposium on Biomedical Imaging, 24 Oct. 2022.
7. Wang, Wenhao. “Orders on Free Metabelian Groups.” *ArXiv.org*, 26 Oct. 2022, https://arxiv.org/abs/2210.14630.