Ben Heinze, Braxton McCormack, Michael Hagin

Research Question

Has the diction in academic research papers within the computer science community changed after January 2023 at a faster rate than usual as a result of the potential assistance of large language models (LLMs)?

Data Description

- **Instances**: There are 114 PDFs analyzed from 2018 to 2024.
- **Attributes**: Each record in the resultant data tables has the following attributes:
  - word: The stemmed word.
  - frequency: The count of the word in a given year.
  - normalized_frequency: The normalized frequency in each paper, summed into the yearly data.
  - isVerb: Boolean flag if the word is used as a verb.
  - isAdj: Boolean flag if the word is used as an adjective.
- **Missing Values**: The approach uses full counts from the available PDFs, assuming no internal missing values in the data. However, some months have missing articles, which could lead to data gaps.
- **Categorical and Numeric Attributes**: word is categorical (nominal), while frequency and normalized_frequency are numeric. The isVerb and isAdj are categorical (binary).

Pre-processing Techniques

- **Tokenization and Part-of-Speech Tagging**: To identify and classify words.
- **Stemming**: Using PorterStemmer to reduce words to their base forms. This is chosen over lemmatization for its simplicity and speed in processing large text data.
- **Normalization of Frequencies**: To compare frequencies on a similar scale across different years.

Data Mining Techniques

- **Term Frequency**: mention shortcomings and strong suits.
- **Year-to-Year Comparison**: Helps observe shifts in vocabulary usage over time, particularly before and after 2023.

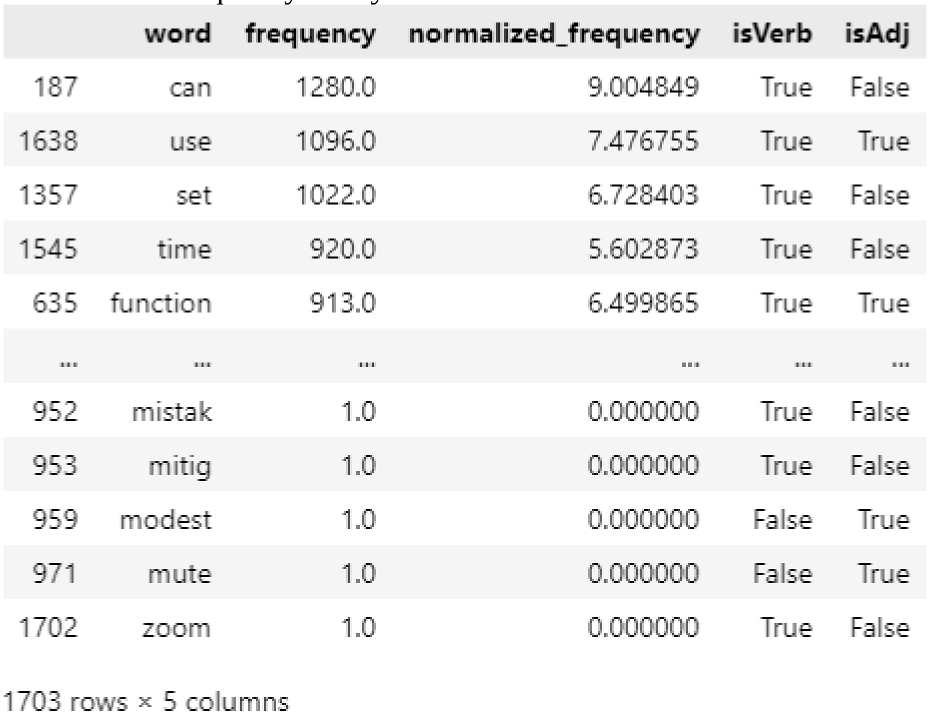
Evaluation Techniques

- **Comparative Analysis**: Direct difference in use of normalized word frequencies over different years to note any significant changes or trends.
- **Visualization**: Scatter plots and frequency tables help visually identify shifts in word usage.

Visualizations and Tables

These tables display the frequency analysis of stemmed words extracted from academic papers published in the Journal of the ACM. It includes each word's overall frequency, normalized frequency to allow comparison across years, and classifications indicating whether the word is predominantly used as a verb (isVerb) or an adjective (isAdj).

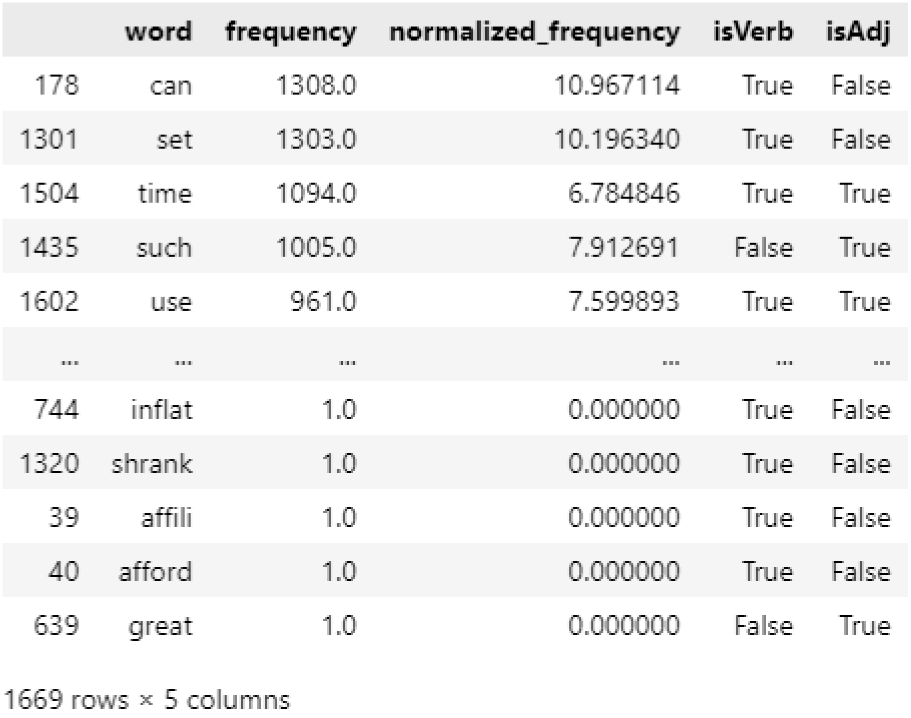
Table 2018: Frequency Analysis of Stemmed Words



	word	frequency	normalized_frequency	isVerb	isAdj
187	can	1280.0	9.004849	True	False
1638	use	1096.0	7.476755	True	True
1357	set	1022.0	6.728403	True	False
1545	time	920.0	5.602873	True	False
635	function	913.0	6.499865	True	True
...	...	...	...	...	...
952	mistak	1.0	0.000000	True	False
953	mitig	1.0	0.000000	True	False
959	modest	1.0	0.000000	False	True
971	mute	1.0	0.000000	False	True
1702	zoom	1.0	0.000000	True	False

1703 rows × 5 columns

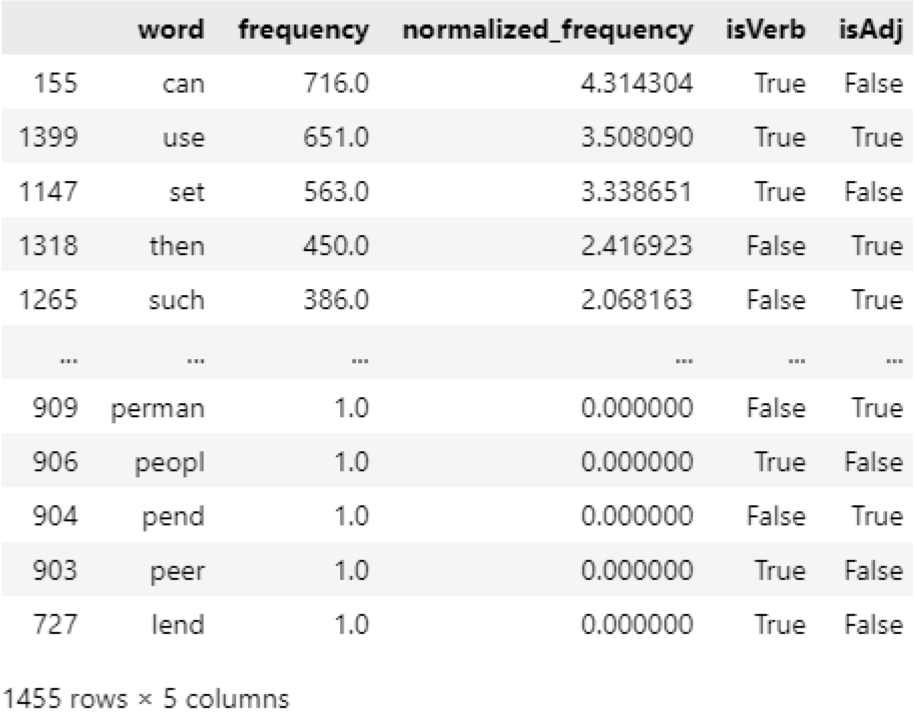
Table 2019: Frequency Analysis of Stemmed Words



	word	frequency	normalized_frequency	isVerb	isAdj
178	can	1308.0	10.967114	True	False
1301	set	1303.0	10.196340	True	False
1504	time	1094.0	6.784846	True	True
1435	such	1005.0	7.912691	False	True
1602	use	961.0	7.599893	True	True
...	...	...	...	...	...
744	inflat	1.0	0.000000	True	False
1320	shrank	1.0	0.000000	True	False
39	affili	1.0	0.000000	True	False
40	afford	1.0	0.000000	True	False
639	great	1.0	0.000000	False	True

1669 rows × 5 columns

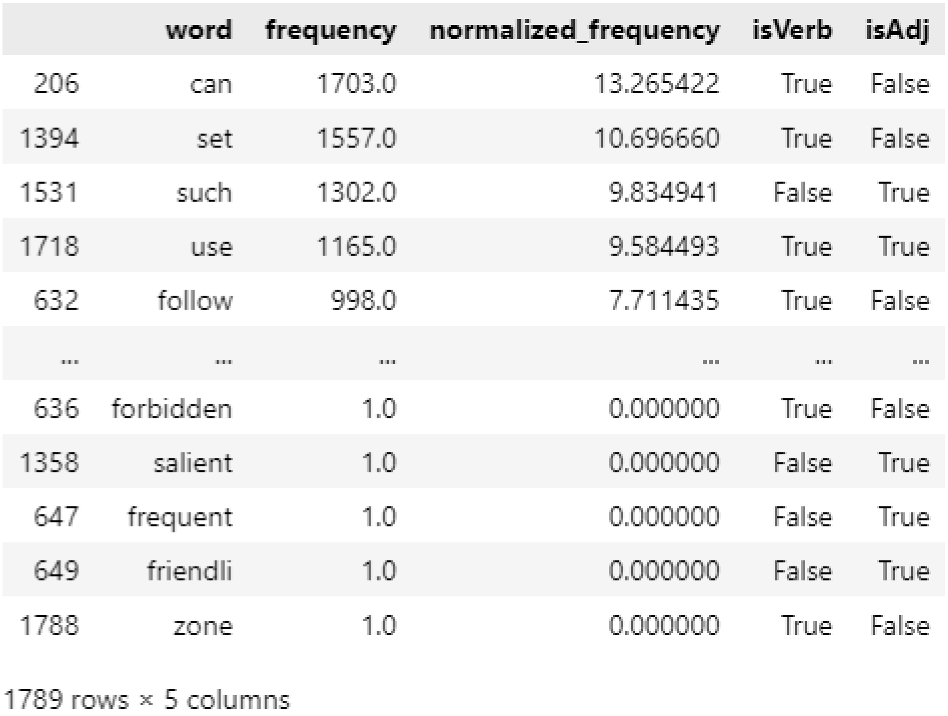
Table 2020: Frequency Analysis of Stemmed Words



	word	frequency	normalized_frequency	isVerb	isAdj
155	can	716.0	4.314304	True	False
1399	use	651.0	3.508090	True	True
1147	set	563.0	3.338651	True	False
1318	then	450.0	2.416923	False	True
1265	such	386.0	2.068163	False	True
...	...	...	...	...	...
909	perman	1.0	0.000000	False	True
906	peopl	1.0	0.000000	True	False
904	pend	1.0	0.000000	False	True
903	peer	1.0	0.000000	True	False
727	lend	1.0	0.000000	True	False

1455 rows × 5 columns

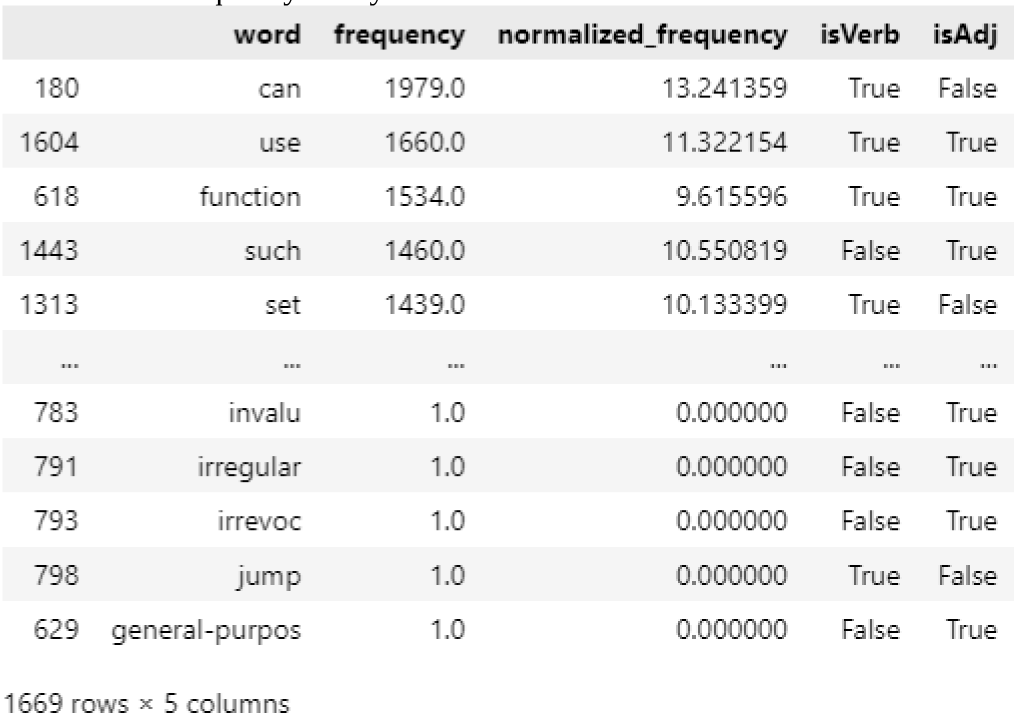
Table 2021: Frequency Analysis of Stemmed Words



	word	frequency	normalized_frequency	isVerb	isAdj
206	can	1703.0	13.265422	True	False
1394	set	1557.0	10.696660	True	False
1531	such	1302.0	9.834941	False	True
1718	use	1165.0	9.584493	True	True
632	follow	998.0	7.711435	True	False
...	...	...	...	...	...
636	forbidden	1.0	0.000000	True	False
1358	salient	1.0	0.000000	False	True
647	frequent	1.0	0.000000	False	True
649	friendli	1.0	0.000000	False	True
1788	zone	1.0	0.000000	True	False

1789 rows × 5 columns

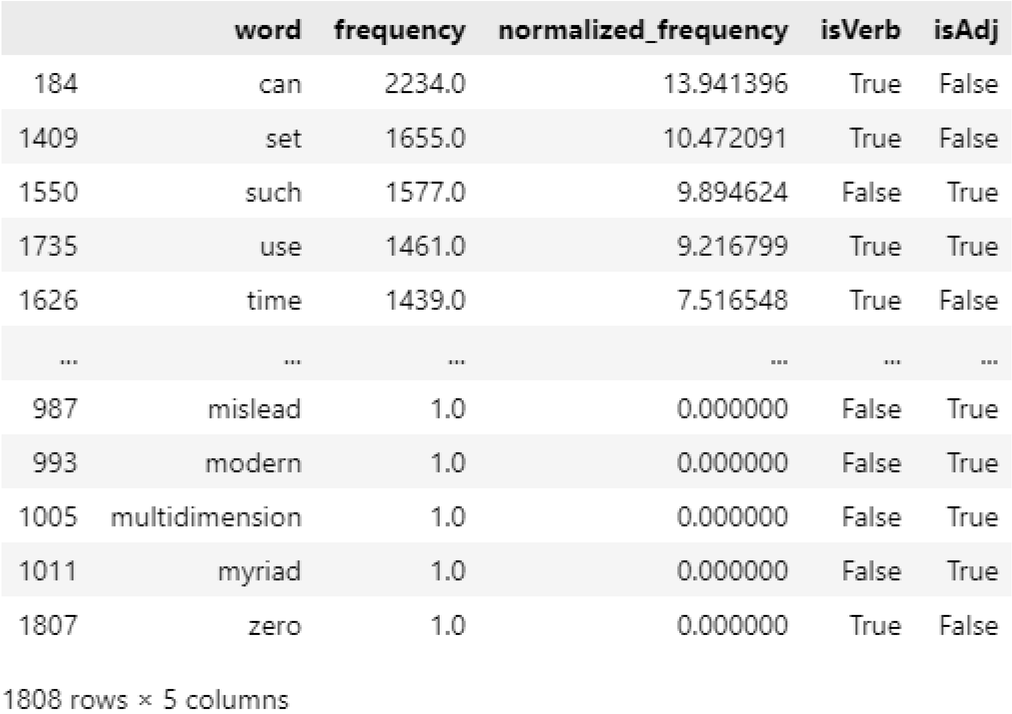
Table 2022: Frequency Analysis of Stemmed Words



	word	frequency	normalized_frequency	isVerb	isAdj
180	can	1979.0	13.241359	True	False
1604	use	1660.0	11.322154	True	True
618	function	1534.0	9.615596	True	True
1443	such	1460.0	10.550819	False	True
1313	set	1439.0	10.133399	True	False
...	...	...	...	...	...
783	invalu	1.0	0.000000	False	True
791	irregular	1.0	0.000000	False	True
793	irrevoc	1.0	0.000000	False	True
798	jump	1.0	0.000000	True	False
629	general-purpos	1.0	0.000000	False	True

1669 rows × 5 columns

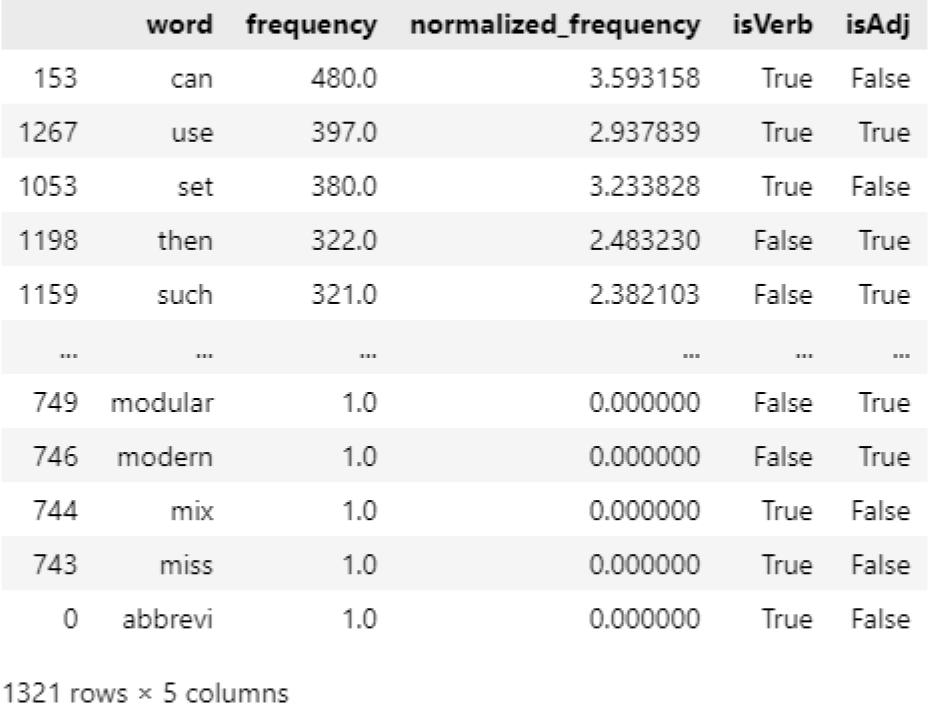
Table 2023: Frequency Analysis of Stemmed Words



	word	frequency	normalized_frequency	isVerb	isAdj
184	can	2234.0	13.941396	True	False
1409	set	1655.0	10.472091	True	False
1550	such	1577.0	9.894624	False	True
1735	use	1461.0	9.216799	True	True
1626	time	1439.0	7.516548	True	False
...	...	...	...	...	...
987	mislead	1.0	0.000000	False	True
993	modern	1.0	0.000000	False	True
1005	multidimension	1.0	0.000000	False	True
1011	myriad	1.0	0.000000	False	True
1807	zero	1.0	0.000000	True	False

1808 rows × 5 columns

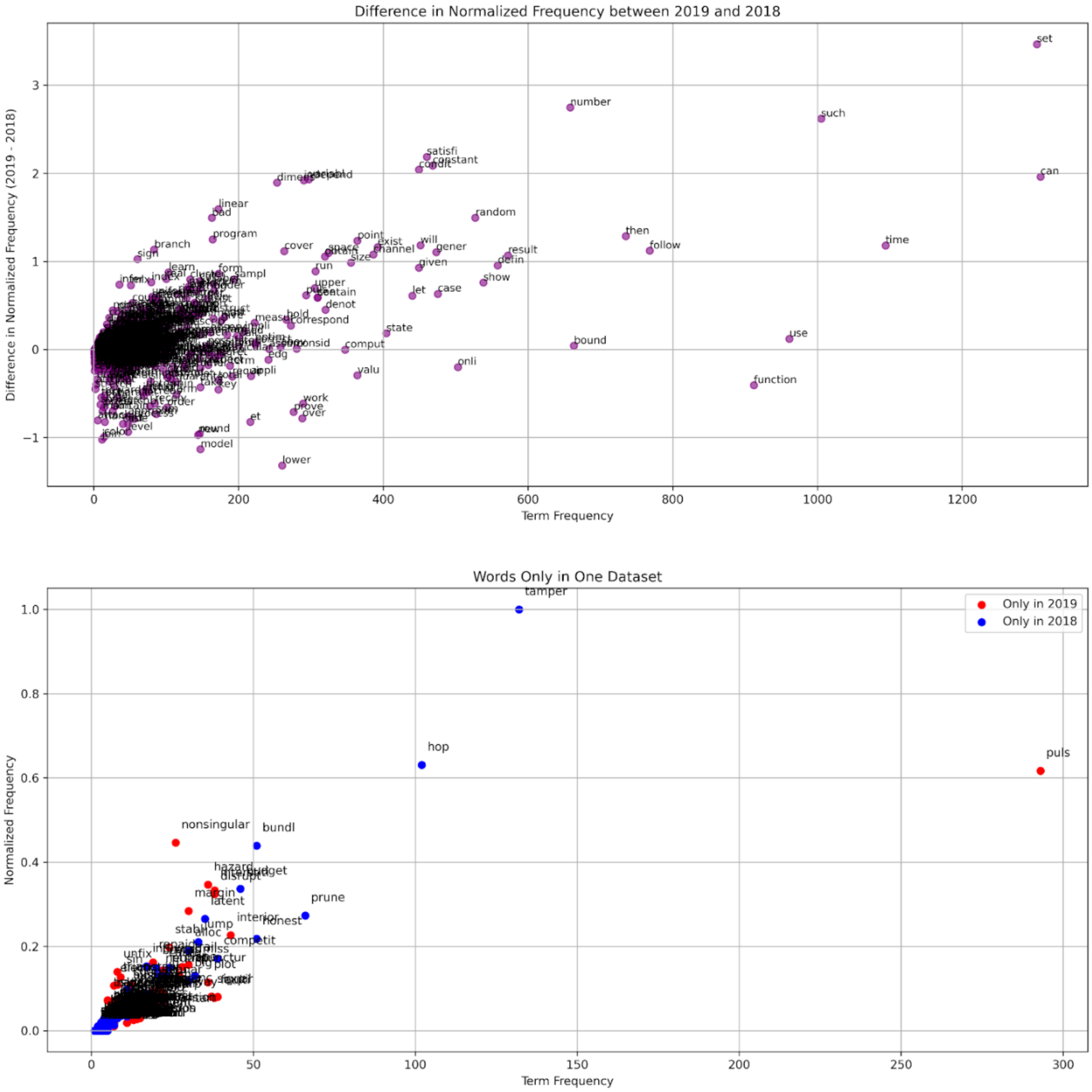
Table 2024: Frequency Analysis of Stemmed Words



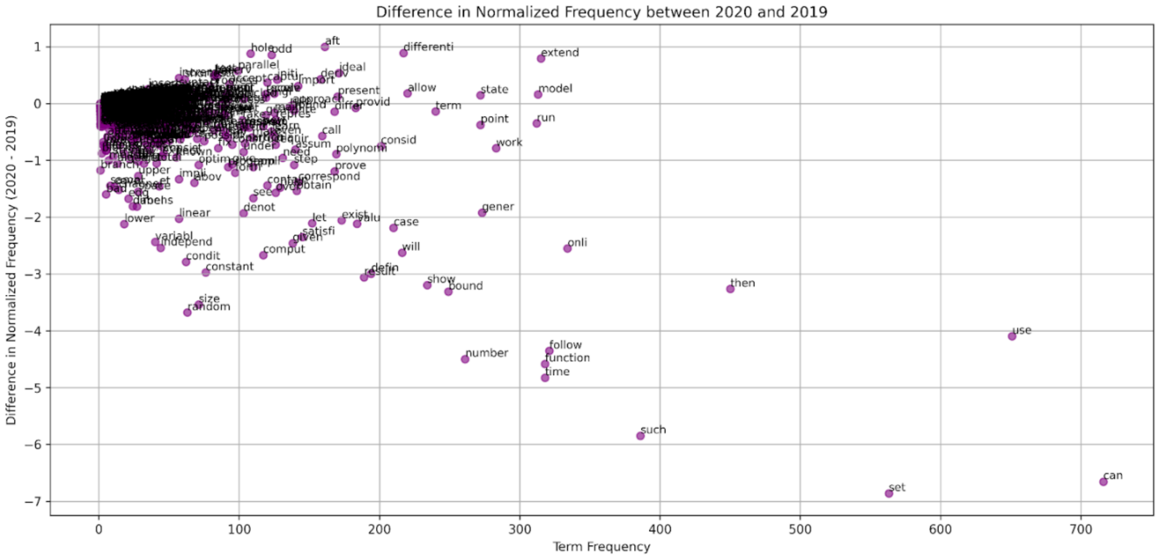
	word	frequency	normalized_frequency	isVerb	isAdj
153	can	480.0	3.593158	True	False
1267	use	397.0	2.937839	True	True
1053	set	380.0	3.233828	True	False
1198	then	322.0	2.483230	False	True
1159	such	321.0	2.382103	False	True
...	...	...	...	...	...
749	modular	1.0	0.000000	False	True
746	modern	1.0	0.000000	False	True
744	mix	1.0	0.000000	True	False
743	miss	1.0	0.000000	True	False
0	abbrevi	1.0	0.000000	True	False

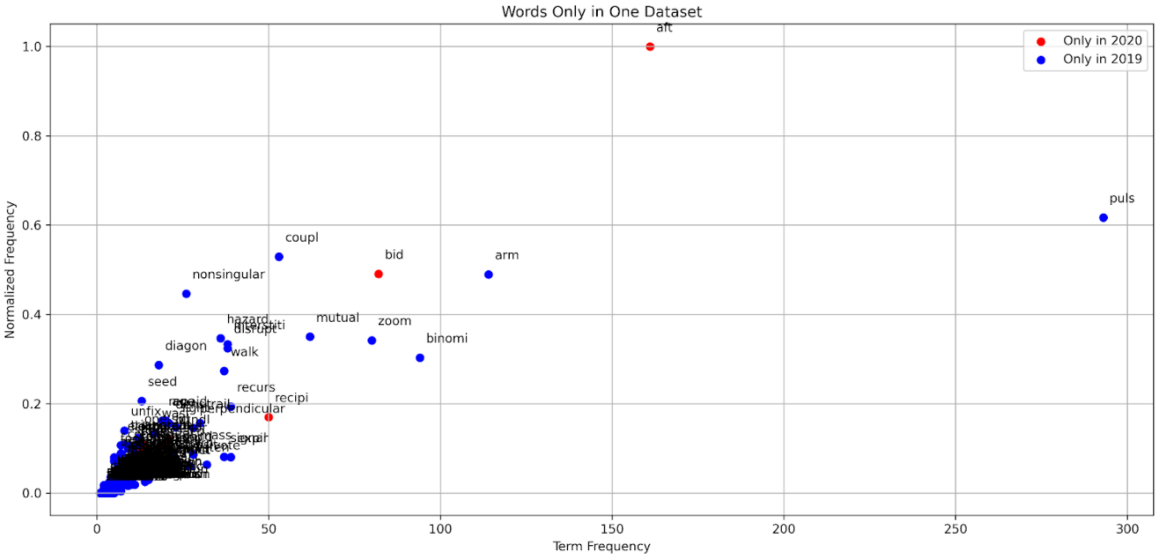
1321 rows × 5 columns

Comparative Graph 2018 vs. 2019: This graph compares changes in normalized word frequencies between papers published in 2018 and 2019.

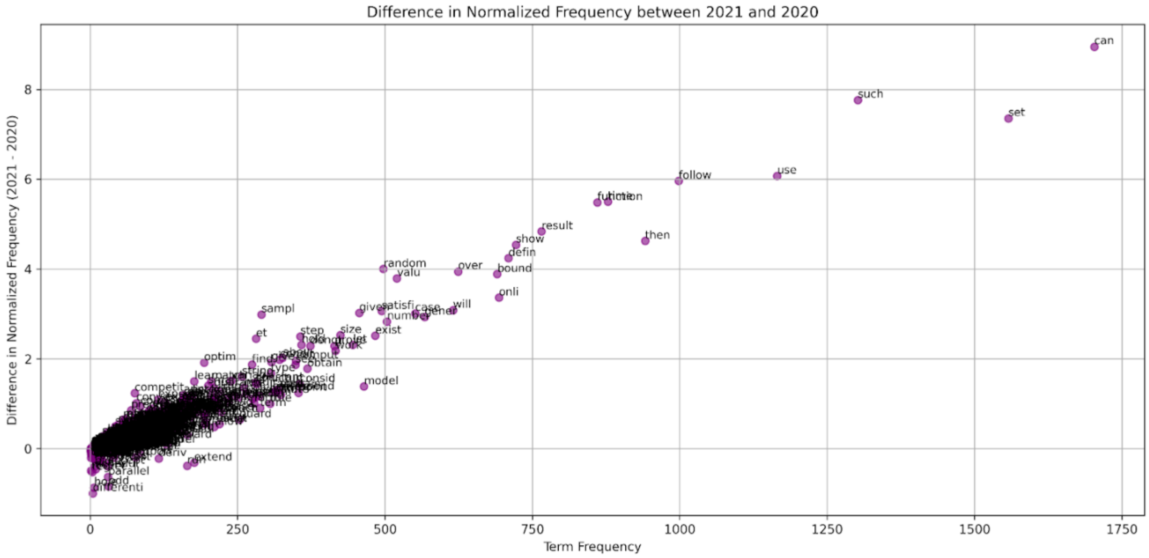


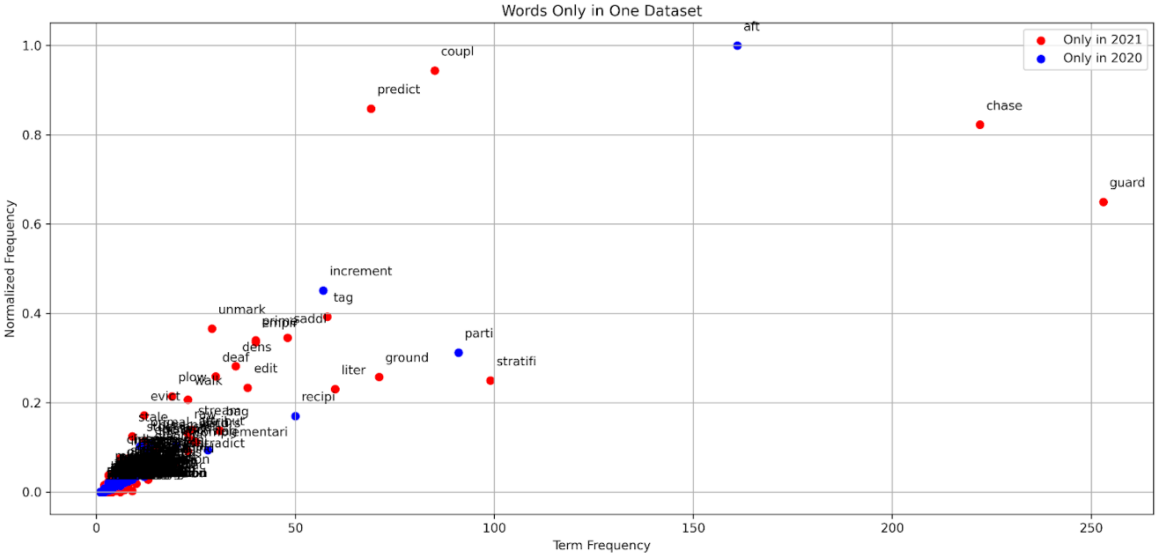
Comparative Graph 2019 vs. 2020: This graph compares changes in normalized word frequencies between papers published in 2019 and 2020.



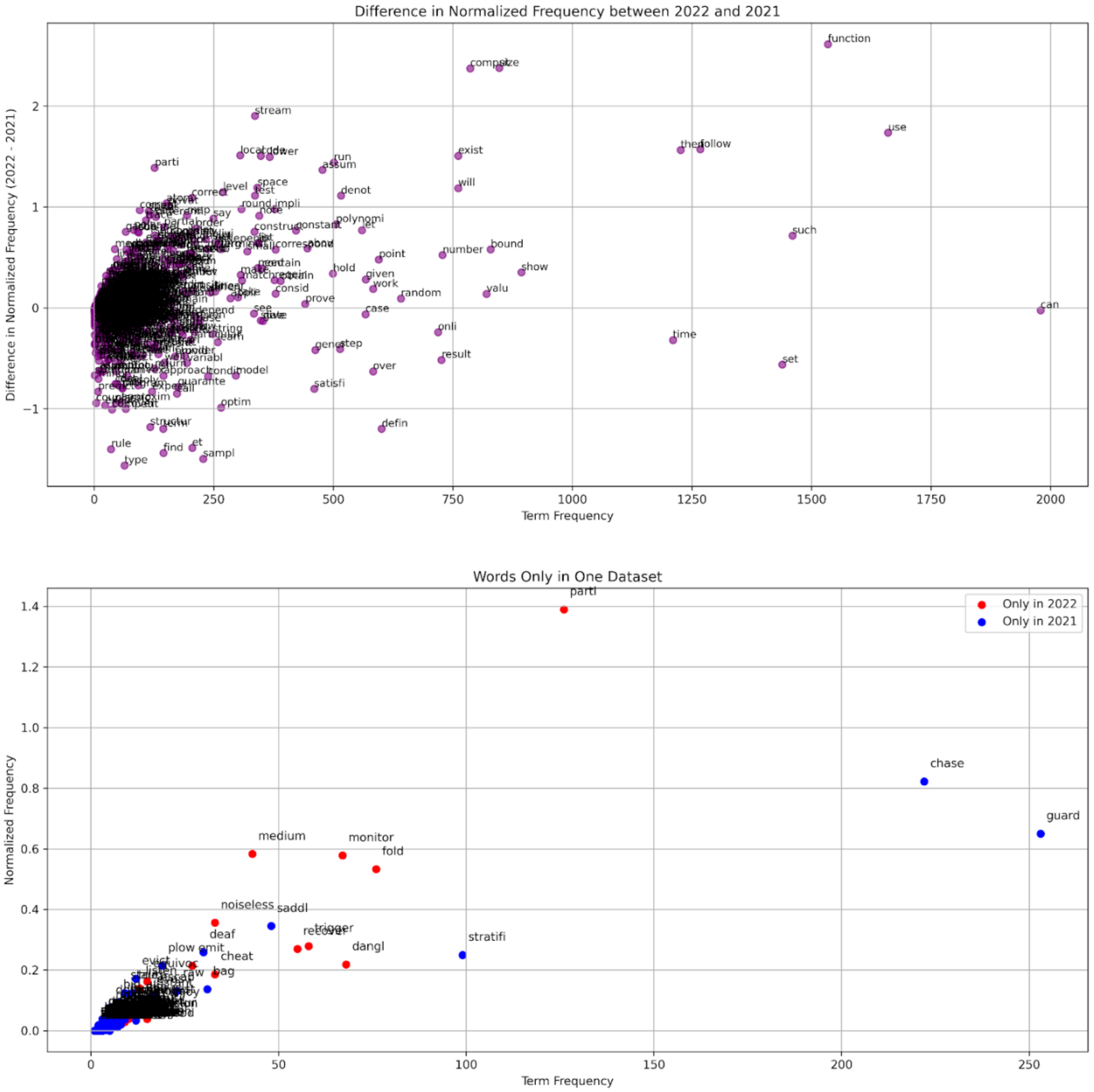


Comparative Graph 2020 vs. 2021: This graph compares changes in normalized word frequencies between papers published in 2020 and 2021.

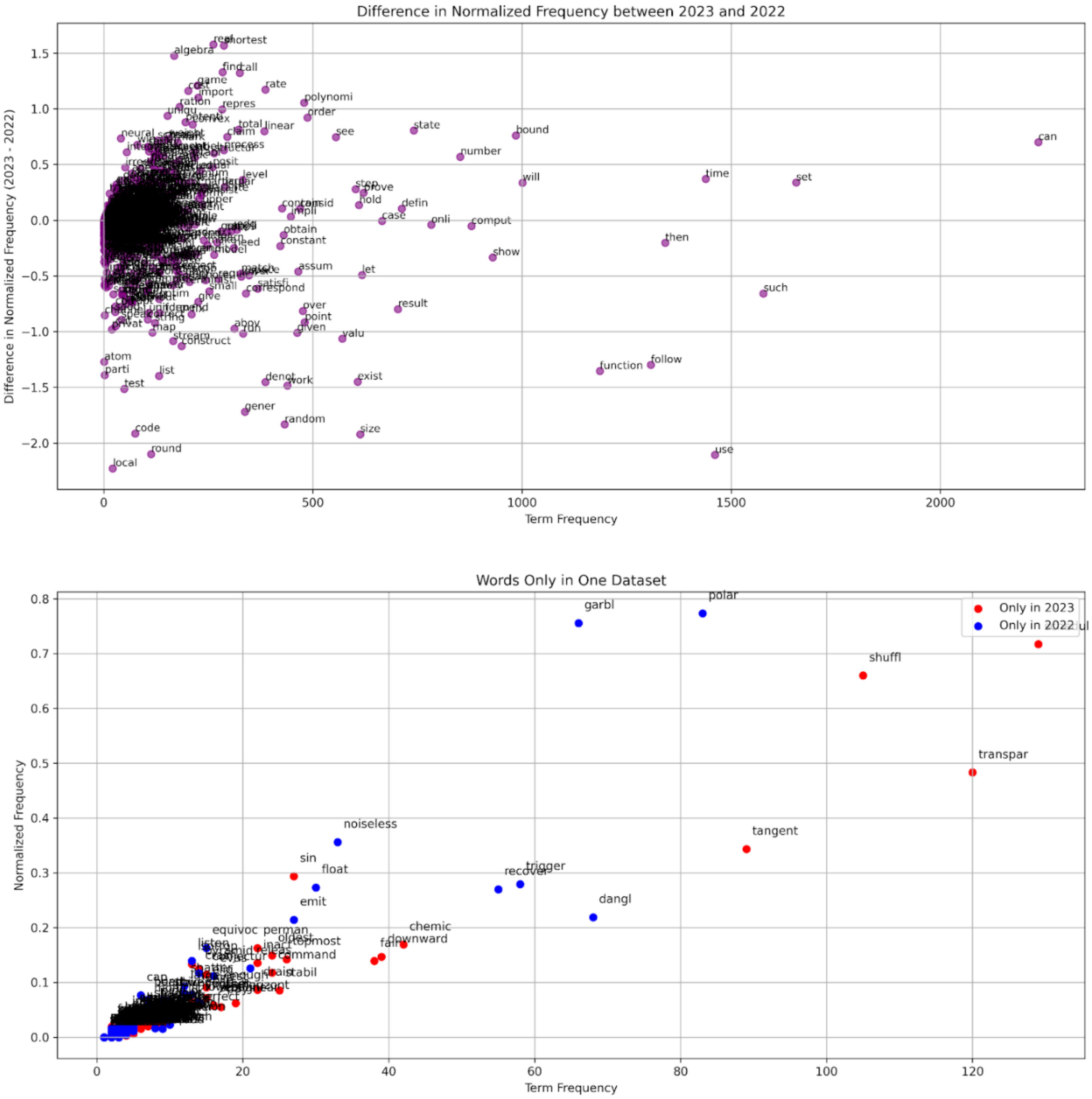

Difference in Normalized Frequency between 2021 and 2020


Words Only in One Dataset

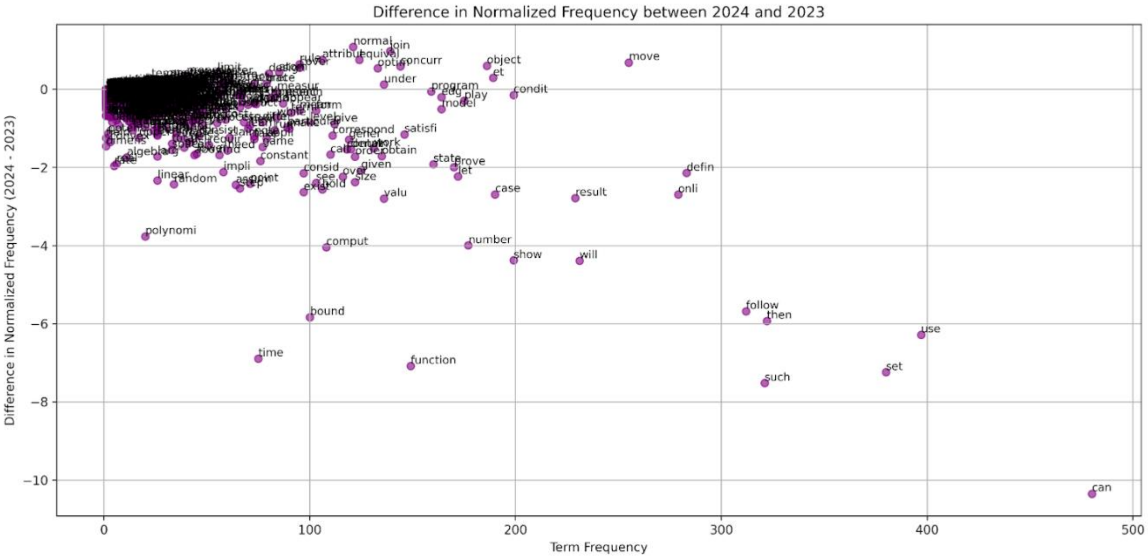
Comparative Graph 2021 vs. 2022: This graph compares changes in normalized word frequencies between papers published in 2021 and 2022.

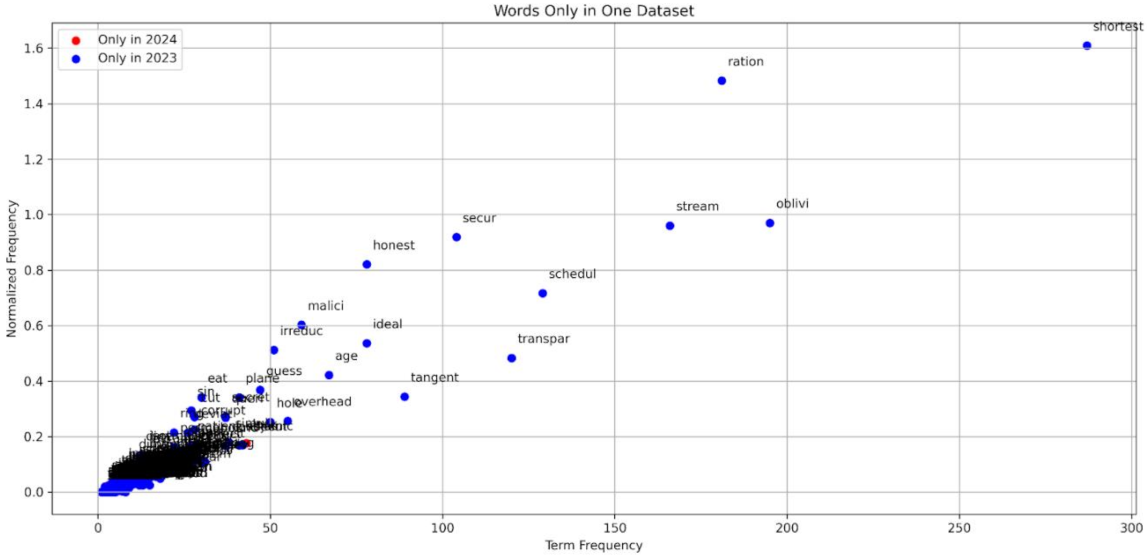


Comparative Graph 2022 vs. 2023: This graph compares changes in normalized word frequencies between papers published in 2022 and 2023.



Comparative Graph 2023 vs. 2024: This graph compares changes in normalized word frequencies between papers published in 2023 and 2024.





Key Findings and Surprises

- **Trends in Differences in Normalized Frequency**: The scatter plots showing differences in normalized frequency between years did not reveal a consistent trend across the dataset. Notably, the difference in normalized frequency decreased from 2019 to 2020 and increased from 2020 to 2021, showing the strongest positive trend observed in the analysis. This fluctuation is likely attributable to the reduced number of articles in 2020, which had the lowest at only seven, compared to at least 17 in other years. This significant drop in publications could be linked to disruptions correlated with the COVID-19 pandemic. The smaller sample size in 2020 might have resulted in less lexical diversity despite the normalization of the data. Interestingly, the Differences in Normalized Frequency between 2022 and 2023 demonstrated a much larger spread of vocabulary, indicating a diverse range of topics and terms used, which contrasts with the more constrained lexical variety in the previous years.

- **Surprise:** The similar trend observed between the years 2023 and 2024 could suggest that the small sample size effect is surfacing again. As of May 2024, only 6 articles have been analyzed, which is comparable to the scarcity in 2020. This is due to how early we still are in the year as only 2 of the 6 issues per year have been released so far. This instance of reduced publications might again be impacting the diversity and representativeness of the linguistic analysis.

Contributions and Future Work

- **Contributions:** This analysis provides an insightful examination of how linguistic patterns in academic computer science research have changed year over year, particularly in response to external factors like the pandemic and possibly the advent of new technologies like LLMs. It highlights how significant global events such as COVID-19 can directly impact academic output and subsequently, the lexical diversity observed in scholarly articles.
- **Future Work**: To refine our understanding of linguistic changes in academic research, expanding the dataset would be crucial. Specifically, analyzing articles beyond 2024 could clarify whether the trends noted in 2023 and 2024 are temporary fluctuations or part of a longer-term shift. Additionally, including a wider array of journals and articles would address the limitations posed by the current small sample size, enhancing the robustness of our findings. A larger dataset would not only provide a more comprehensive view of the field's evolution but also allow for more detailed statistical analyses, potentially revealing subtler trends and patterns that smaller datasets might miss.

Conclusion

This study highlights how significant events can influence the language used in academic papers. The advent of large language models and the impact of the COVID-19 pandemic have both noticeably affected the vocabulary in computer science research. Notably, the Differences in Normalized Frequency between 2022 and 2023 saw a much larger spread of vocabulary, suggesting a dynamic shift or broadening in topics and terminologies used within the field. However, a deeper understanding of these effects requires further investigation. The changes observed around 2020 and 2024 indicate that major events like the pandemic potentially affected academic writing. This suggests that ongoing monitoring of these trends is essential to fully grasp how academic discourse evolves over time.