Report

NLP and Knowledge Graph-Based Analysis of SWECRIS and ERC Database

Capstone- Odd Semester 2023

Guide: Dr. N Arul Murugan

Group 5: Abhishek Acharya (MT22004)

Project Aim

In continuation of the Independent Project done during the Summer Semester, there is a need to delve further into the SweCris Database and conduct an in-depth analysis of the available data. The primary objective is to extract meaningful insights from Knowledge Graphs related to diseases, viruses, and infections. This involves extracting the words preceding and following these key terms to draw comprehensive conclusions.

The analysis will include the calculation of word frequencies associated with each keyword, followed by the assignment of weights based on these frequencies. Subsequently, the top-weighted words for each keyword will be identified and plotted to visually represent significant associations within the data. This gives the Frequency Word Graph.

Upon completion of the SweCris Database analysis, the research will extend to the European Research Council (ERC) database. Similar methodologies will be applied, encompassing keyword extraction, frequency calculations, weight assignments, and visualization of top-weighted words. This parallel analysis aims to derive comparable insights from both databases.

The final phase involves a comparative analysis of the results obtained from the SweCris and ERC databases. By juxtaposing and scrutinizing the outcomes of the analyses, a holistic understanding of the commonalities and distinctions between the two datasets will be attained, contributing to a more comprehensive overview of the research landscape in the realm of trending projects, diseases, viruses, and infections.

Databases Used

1. SweCris Database
2. ERC Database

Tools Used

1. Python Libraries: csv, requests, pandas, matplotlib, nltk, collections, string, networkx
2. Google Colab

Google Colab Notebooks

1. https://colab.research.google.com/drive/1ffOY7Yw6tnZNnHm_T_8szeRPUrgJiomm?usp=sharing
2. https://colab.research.google.com/drive/1osIf5akBiqfsV-pI2f44mkhwWj04G4VE
3. https://colab.research.google.com/drive/1iIIuuDCQ0ayrB-sKM9yy2EwBMVZ6uR7A
4. https://colab.research.google.com/drive/1XuQvQ_RihuC7ll8MMaZY9ESUwWy6MFwu?usp=sharing

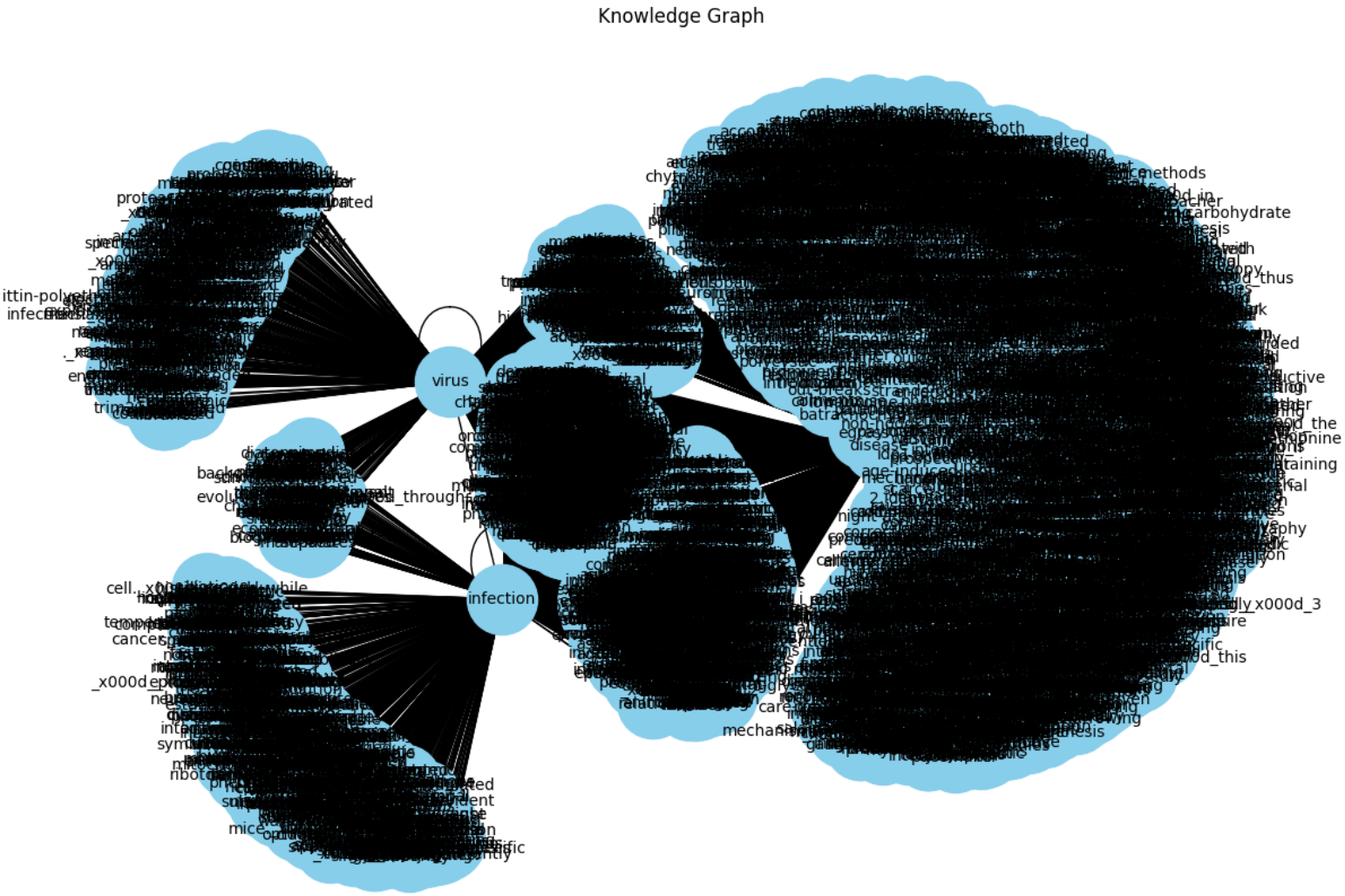
Work Done

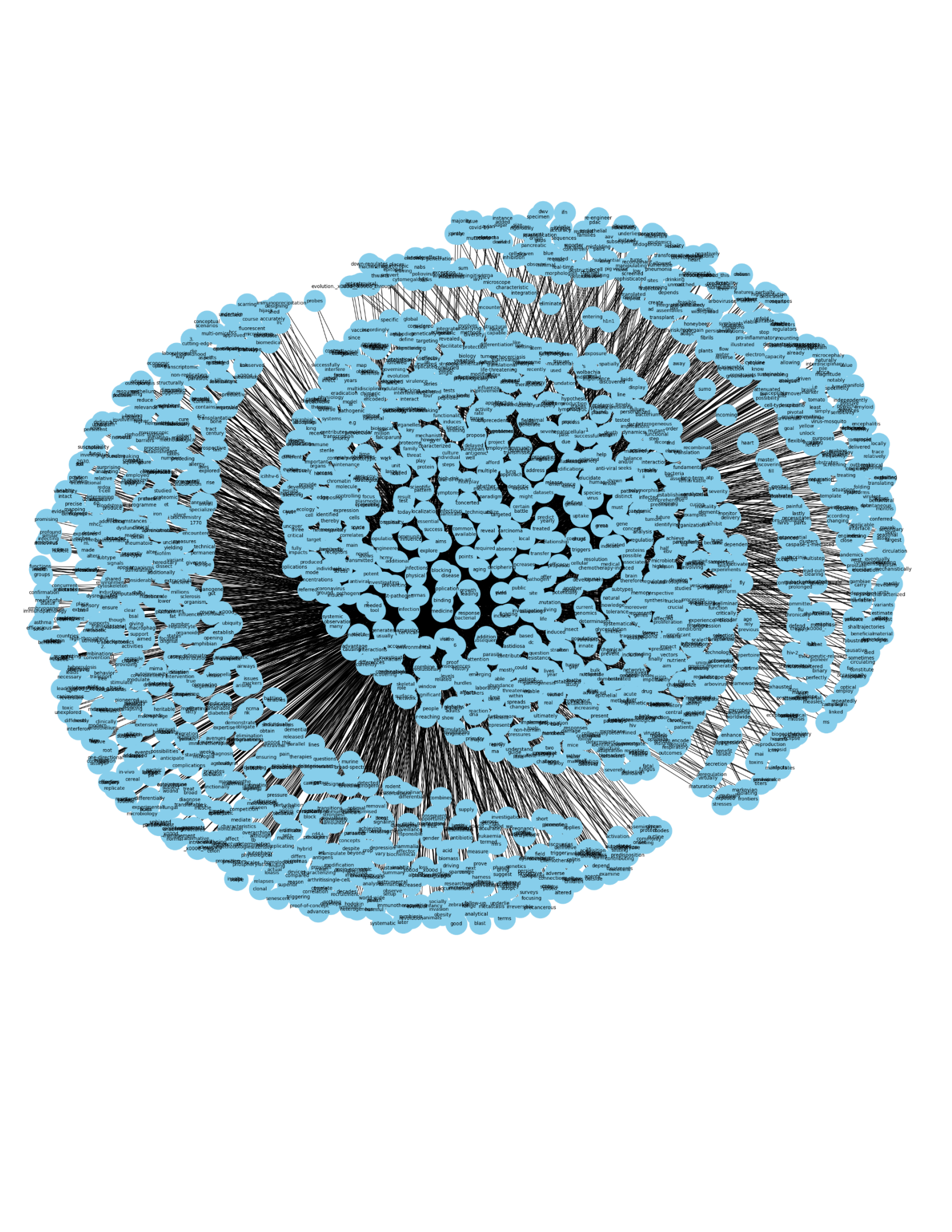
1. A thorough analysis of the SWECRIS database.
2. Conclusions about viruses, illnesses, and diseases are derived from Knowledge Graphs by examining the words that come before and after them.
3. Computed the database's word graph frequencies.
4. A thorough analysis of the ERC database was conducted, and all of the previous SWECRIS database activities were completed.
5. Analyzed and compared the two databases' respective results.

Results

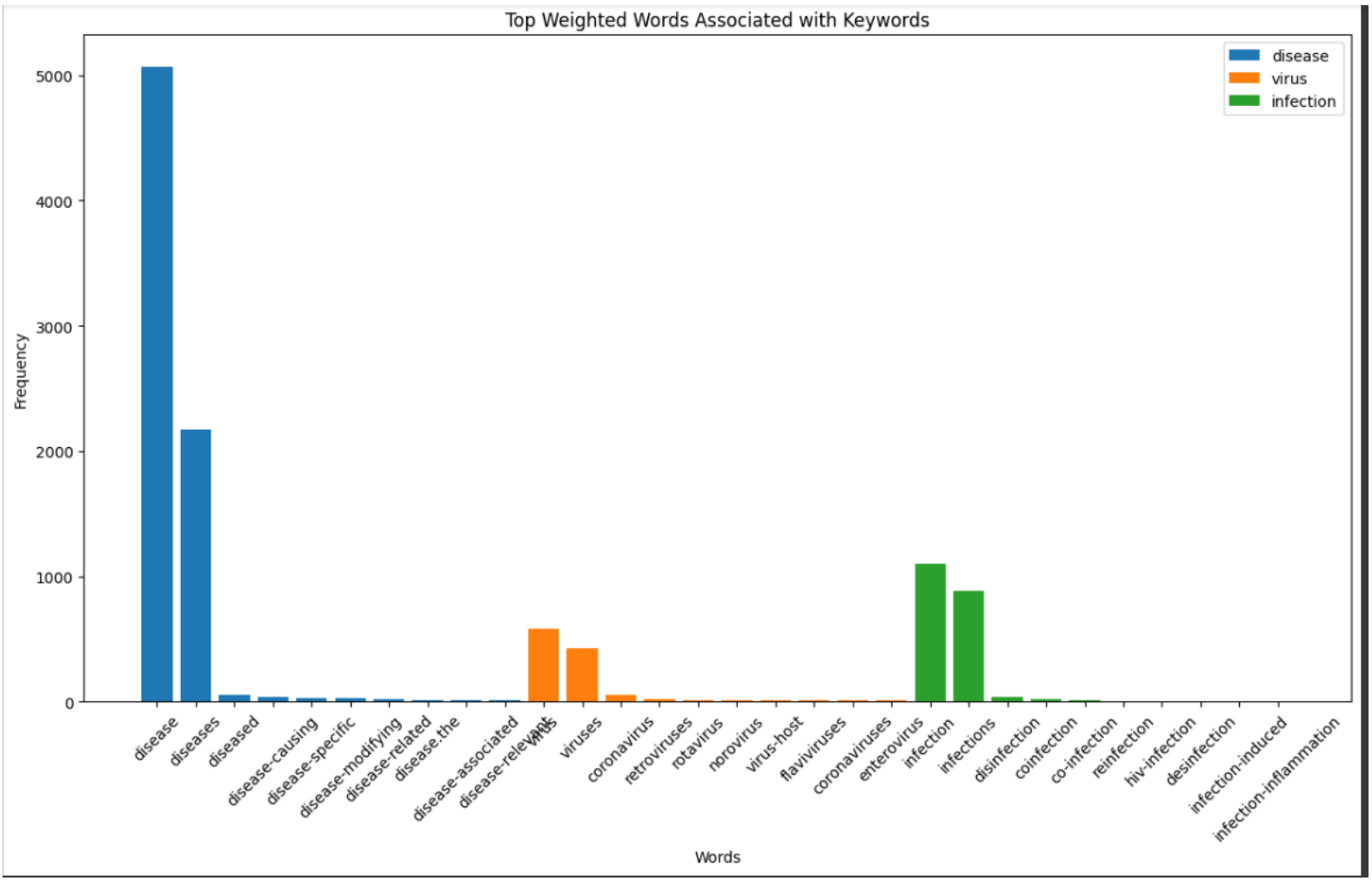
[As some results are long, they have been clipped here. The full results are available in the
Jupyter Notebooks itself.]

1. Knowledge Graph (Natural and after reducing Density)

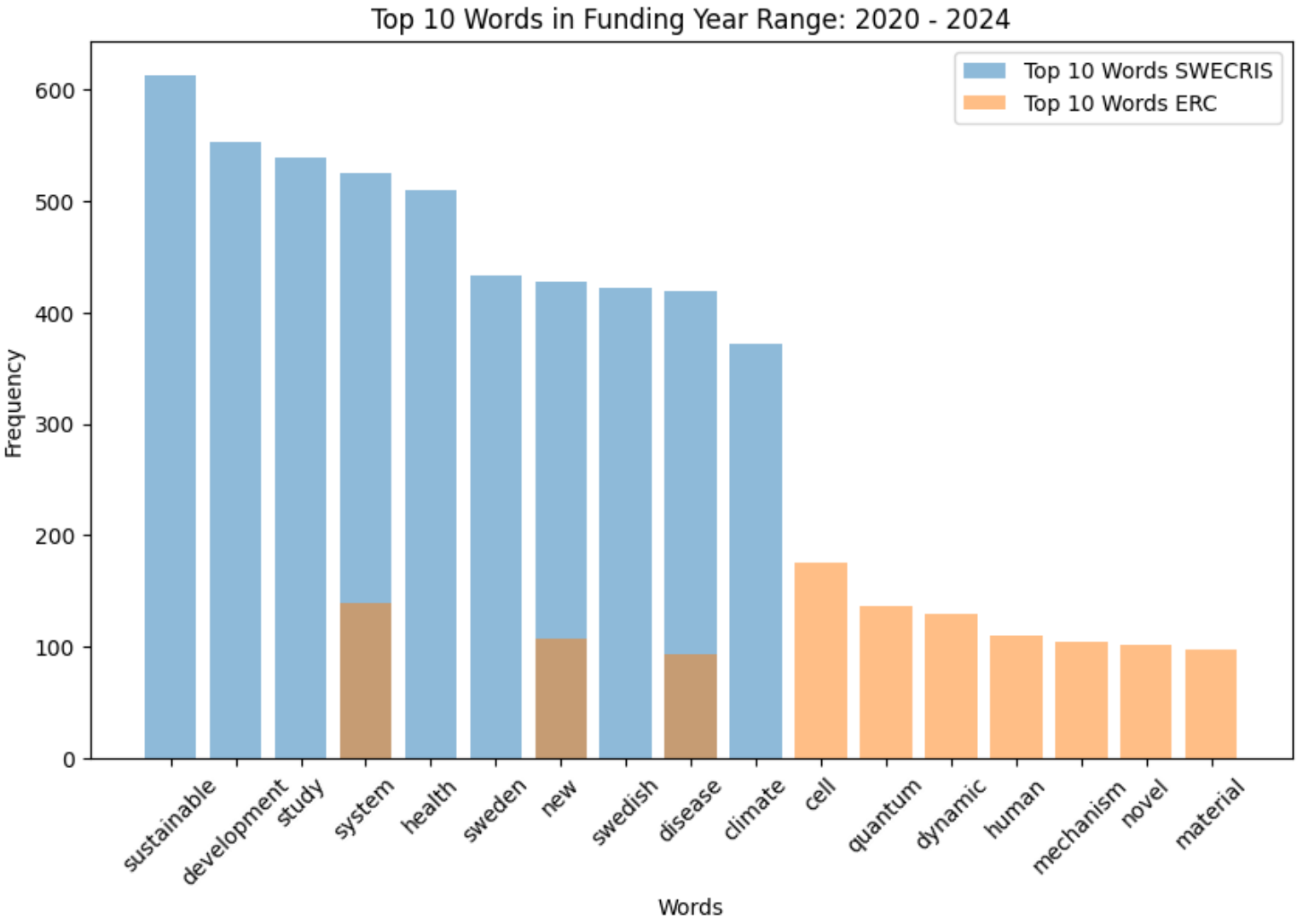


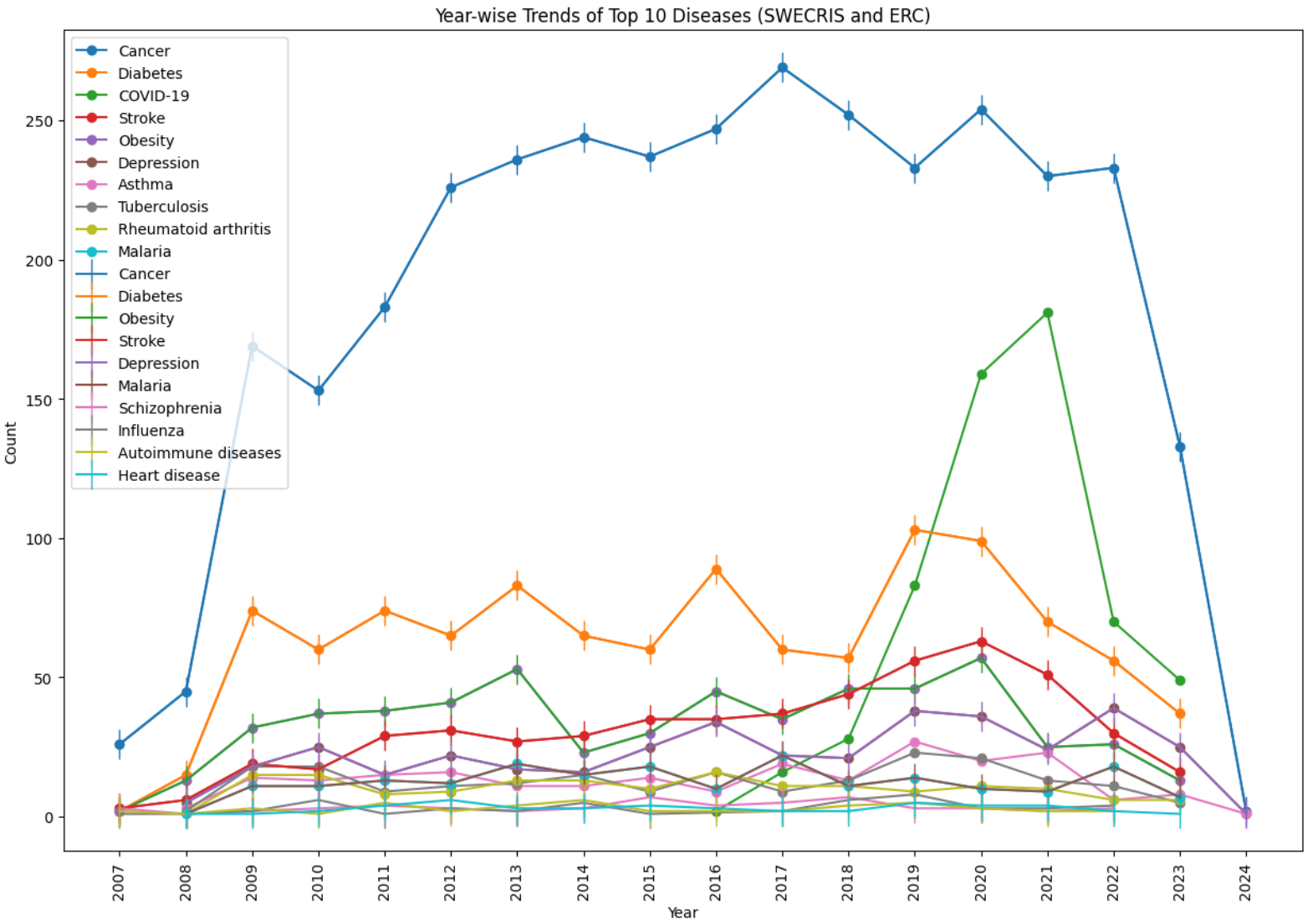


2. Frequency word Graph



3. Analysis between SweCris and ERC





Challenges faced in SERB PRISM Database

The accessibility of the SERB PRISM Database posed a challenge, necessitating the use of web scraping techniques involving Selenium Automation and BeautifulSoup.

After successfully testing these techniques on a comparable Census Government Database, attempts to implement the same procedures on the SERB PRISM Site encountered obstacles. The primary impediments include the dynamic nature of the site, where drop-down options are only accessible after a preceding option has been selected. Consequently, attempting to interact with these options in the code results in a "no such element is present" error.

Furthermore, the absence of an ID for the submit button complicates the process of submitting queries to retrieve the desired tables. Additionally, certain options lead to tables spanning multiple pages, introducing challenges in efficiently fetching the complete data.

Another complexity arises from the variability in available program options across different years, leading to instances where tables may be blank or nonexistent.

Navigating through these intricacies requires a nuanced approach to web scraping, acknowledging the unique characteristics of the SERB PRISM Database. Addressing these challenges will contribute to the successful extraction and interpretation of valuable data from the platform.

Future Work and Conclusion

A concerted effort is imperative to extract the entirety of the SERB PRISM Database utilizing Selenium Automation and subsequently subject it to a comprehensive analysis akin to that performed on the SWECRIS and ERC Databases.

The ensuing step involves a meticulous comparative examination of the three databases to discern prevailing trends related to top diseases and their respective frequency word graphs. This endeavor seeks to unravel insights into the nature of projects being undertaken globally, with a particular focus on Europe and India. The findings are anticipated to serve as a valuable resource for future researchers, aiding them in selecting projects within specific domains of interest.

Moreover, the in-depth disease analysis, encompassing Knowledge Graphs and Frequency Word Graphs, promises to shed light on predominant diseases worldwide, specifically in the regions of Europe and India. This scrutiny is poised to unveil discernible patterns, potentially offering a strategic understanding of disease trends. Such insights hold the potential not only to enhance our comprehension of prevailing diseases but also to discern patterns that may inform strategies for preemptive measures against emerging health threats. This holistic approach contributes to the broader endeavor of safeguarding global health and fostering advancements in research and disease prevention.