Referencing in Scientific Publishing over the years

By Jan Rombouts, student number 439738

Introduction

This report is part of the final assignment of the course Programming 3, where the aim was to work with large datasets and learn to handle multicore processes. Therefore, analysing ALL of PubMed's literature would be a suitable way to show what I learned. A total of 1063 XML files each with 30k articles and their details were downloaded from the NCBI database and made available for analysis. Parsing the XML files took a lot more time than expected, so I did not have time left to create a graph network. As such, I only managed to answer the following, more descriptive questions about the data:

1. How many co-authors does a publication have on average?
2. How many co-authors does a publication have on average per year?
3. How many papers are published per year?
4. How many references does a publication have on average per year?
5. What is the distribution of the primary languages papers are published in?
6. How many articles does an author have as primary author on average?

Methods

Parser

The first challenge was to extract only the required data from each file. I first tried to parse it from scratch myself but then I found that BioPython has its own methods to do this. The biggest challenge came from the reference used in each article. These were not formatted identically everywhere (that would be too easy right!). Some had the PMID of the referenced paper, and some only the authors and title. The PMIDs were easily extracted but I had to be creative to handle all the unique ways the authors were cited. Finally, I parsed each file using multiprocessing and created a Pandas DF for each file, and saved this as a CSV. I tried to save it as a pickle but then I could not load it again (*Error: __new__ missing 2 positional arguments: "tag" and "attributes"*).

Results

1. How many co-authors does a publication have on average?

On average, a paper has 3.2 co-authors. I calculated this by adding up all the co-authors for all papers and dividing by the total number of papers.

2. How many co-authors does a publication have on average per year?

Figure 1 displays the average number of co-authors since 1800. As seen in the figure, the number of co-authors per paper drastically increased since the 1950s, going from ~1 to ~4.5 in the 2010s.

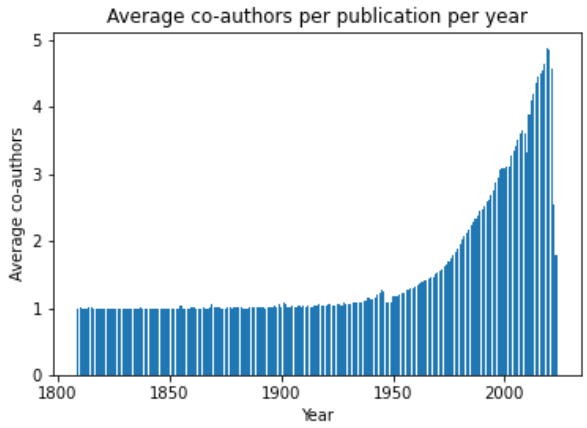

Figure 1

3. How many papers are published per year?

Similar to the number of co-authors, the total number of papers per year also increased after the 1950s.

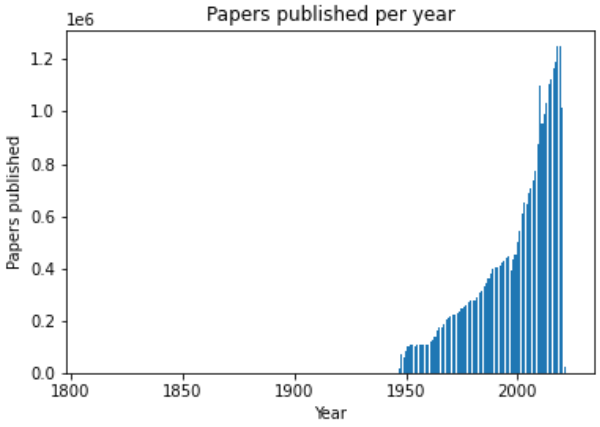

Figure 2

4. How many references does a publication have on average per year?

The number of references per paper was quite high in the 1820s at around 25 and then decreased until 1910. Interestingly, there is very high variability in the references in this period from 1820 to 1910, so perhaps I made a mistake in parsing, where the older references did not parse correctly. Additionally, the number of papers in these years was low (seen in question 3), which further explains the higher variability. Again, after the 1950s, the average number of references increased up to now.

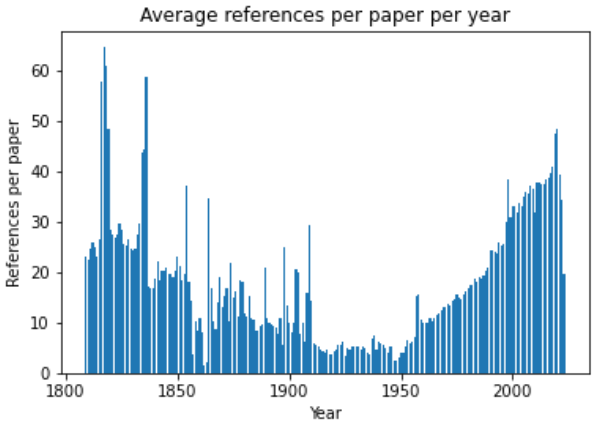

Figure 3

5. What is the distribution of the primary languages papers are published in?

English is the most published primary language for papers with 84%. After that papers are mostly published in German, French, and Russian with ~2%. Japanese, Spanish, Chilean, and Italian are primary languages for ~1% of the papers. All other languages are only used as the primary language in ~3% of all papers.

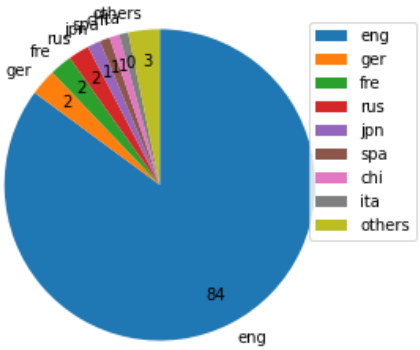

Figure 4

6. How many articles does an author have as primary author on average?

An author has 6.6 publications as a primary author on average.

Conclusions

The questions I answered in this assignment give a general overview of the papers available on PubMed since 1800. The average number of co-authors for a paper is 3.2, and each author has 6.6 papers as the main author. Interestingly, the number of co-authors was ~1 until 1950 and then started increasing.

It was still under 3.2 until ~1990 but increased up to ~4.8 in 2020. Similarly, the number of published papers was <100k per year until 1950 but increased up to 400k in 2000 and 1200k in 2020. The number of references per paper also increased ~4-fold from 1950 to 2020 from 10 to 40.

Overall, publishing scientific papers has seen a large increase in productivity since the 1950s. This goes with an increase in published papers, as well as more scientists working on the same paper. Furthermore, the references each paper uses also increased around fourfold since 1950. Further research could focus specifically on the differences in characteristics of scientific publishing before and after 1950, which could explain the increase in co-authors and references.