Assignment 1: Elementary wrestling with data

a) Retrieving Journal Articles From PubMed

Searched PubMed using GUI. Process returned 6,176 results for 'histone' + '1980 - 1990' + 'English Articles' + Publication Type 'Journal Article'.

b) Query Approach for PubMed and Scopus

I used two different PubMed queries to obtain histone related results. Both times, the query returned 6,176 results.

(((histone) AND (("1980/01/01"[Date - Publication] : "1990/12/31"[Date - Publication]))) AND (english[Language])) AND ("journal article"[Publication Type])

(("histon"[All Fields] OR "histones"[MeSH Terms] OR "histones"[All Fields] OR "histone"[All Fields] OR "histonic"[All Fields] OR "histons"[All Fields]) AND (1980:1990[pdat]) AND "journal article"[Publication Type] AND ("english"[Language])

As an alternative, Entrez API was also considered. The results gave an overview from different databases, which would prove to be helpful should we need to select a particular data source.

Scopus Search Results:

I liked the interface of the Scopus GUI, and thought it was intuitive. Although the updated (currently in Beta) version of the new Scopus interface seems to draw heavily from the PubMed GUI.

I used the following query in Scopus:

TITLE-ABS-KEY (*histone*) AND PUBYEAR >*1979* AND PUBYEAR <*1991* AND (LIMIT-TO (DOCTYPE , "*ar*")) AND (LIMIT-TO (LANGUAGE , "*English*")) AND (LIMIT-TO (SRCTYPE , "*j*"))

The search query returned 6,402 results, which was more than the results obtained from the PubMed query. However, scopus limits exporting the results to 2,000 at a time. This meant that the 1980 – 90 timeline had to be split into three intervals in order to comply with the 2,000 records per search instance (a simple CSV import allows upto 20,000 records but I also wanted to include PMID which reduced the efficiency to 2,000 records per search instance).

c) Matching PubMed data to open_citations in ernieplus

The following psql query was used to match doi from the PubMed data with the citing and cited data in open_citations

```
SELECT t1.doi, t2.citing
FROM pubmed_data_table AS t1
LEFT JOIN open_citations AS t2   ON   t1.doi = t2.citing
```

```
SELECT t1.doi, t2. cited
FROM pubmed_data_table AS t1
LEFT JOIN open_citations AS t2   ON   t1.doi = t2.cited
```

Once the cited and citing matches were made, I used pandas to combine the two search results, such that for every DOI from the PubMed data, corresponding citing and cited counts were displayed in the CSV.

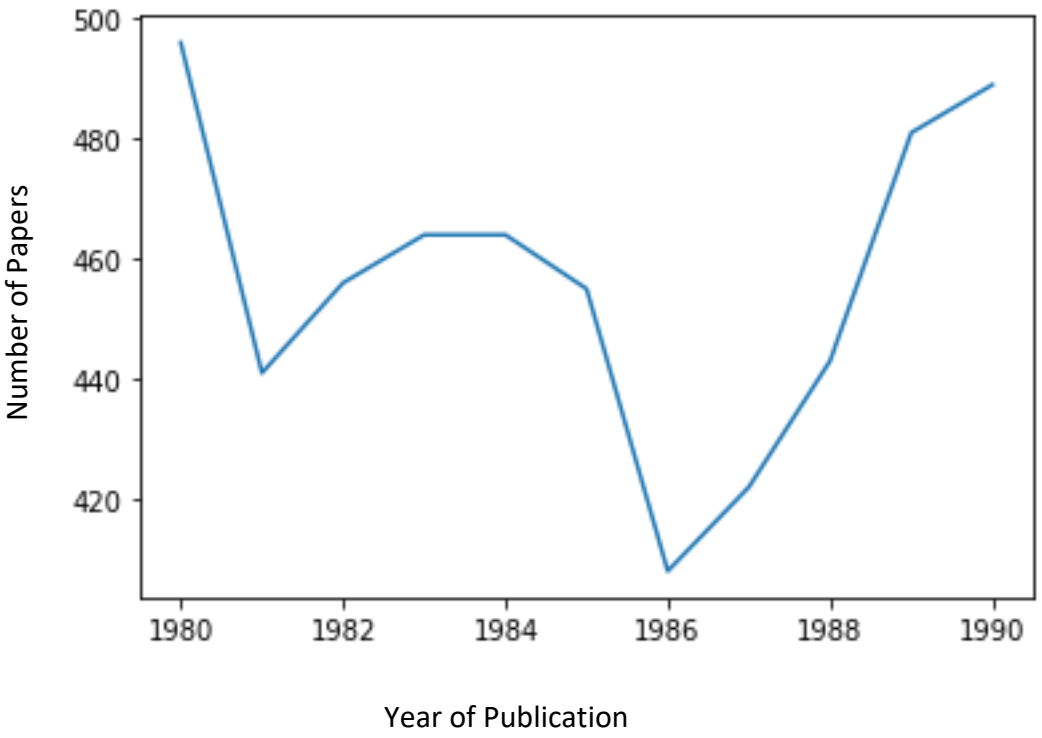
d) Degree Distribution of Nodes

I obtained 5,009 results from this exercise. That is, out of the 6,176 data records, there were 5,009 matches that were made with the open_citations. In total, there were 88,500 out edges (citing) and 166,712 in edges (cited) from 5,009 nodes.

e) Number of citations per year

The distribution of number of papers with the year of publication is shown on the following page. Although I don't feel confident in this graph output – I would expect that there would be an increasing or decreasing relationship as the years go by. However, the time span is relatively short (10 years) hence the variation is not as 'defined'.

Variation Of 'Histone' Related Papers By Publication Year



Conclusions

Three different sources were considered for histone related data. PubMed, which returned 6,176 matches for histone related journal articles between 1980-1990, was chosen given ease of use, and relatively easier extraction of results. The data obtained was then matched with open_citations database that is hosted on Valhalla. The results of this matching exercise yielded 5,006 nodes with 88,500 citing edges, and 166,712 cited edges. The yearly distribution graph matched a 'W' shaped curve, meaning histone related research went through periods of decline and fast growth, in quick succession between 1980 and 1990.