Network analysis of 19th century British mathematicians and philosophers

University of
St Andrews

Megan Briers

Supervisors:
Dr Isobel Falconer
Dr Mark-Jan Nederhof

21st March 2022

Contents

1 Abstract

The portrayal of mathematicians in popular media often centres around the idea of a lone genius. Traditional co-authorship studies that look at the trends in named collaboration on publications often continue to perpetuate this myth, with a slower uptake in co-authorship seen across mathematics compared to more physical sciences such as physics and astronomy. This report takes a different approach at assessing collaborative behaviours of both 19th century British mathematicians and philosophers. By looking at the networks present on their Wikipedia articles instead of at the often flawed co-authorship statistics, new insights can be provided into the collaborative networks of these groups.

Code has been developed which assesses the performance of various Named Entity Recognition methods at extracting a list of mentioned people from given Wikipedia articles. Additionally the links to additional Wikipedia articles from a given page were identified. Once a level of confidence was achieved in these methods, the techniques were applied across the networks of mathematicians and philosophers. The NER method had a F1 score of 0.7697 and the link identification method had a F1 score of 0.8328. It was discovered that mathematicians mention significantly more people per 1000 characters, but the length of an average philosophy article is longer. Additionally, the network present on the Wikipedia article of Mary Somerville is compared with her known correspondence network as a test of what sort of information can be extracted from this comparison. This method has provided a promising alternative to traditional co-authorship evaluations of collaborative networks.

2 Declaration

I declare that the material submitted for assessment is my own work except where credit is explicitly given to others by citation or acknowledgement. This work was performed during the current academic year except where otherwise stated.

The main text of this project report is 20,253 words long, including project specification and plan.

In submitting this project report to the University of St Andrews, I give permission for it to be made available for use in accordance with the regulations of the University Library. I also give permission for the title and abstract to be published and for copies of the report to be made and supplied at cost to any bona fide library or research worker, and to be made available on the World Wide Web.

I retain the copyright in this work.

3 Introduction

3.1 Motivation

This project initially stemmed from an idea to investigate the portrayal of mathematicians through online biographical sources, such as Wikipedia and MacTutor, with the focus being on whether these sources presented a representative picture of members of mathematical communities. To people who have not spent a significant amount of time in a mathematics department, maths is commonly viewed as a fairly solitary activity, with people shut away, spending hours working alone on the same question until an answer is discovered. It is easy to see from these assumptions how the idea of a "lone genius" is so popular in depictions of mathematicians, most famously Einstein and more recently Andrew Wiles, who allegedly worked on Fermat's Last Theorem for years without even discussing the content of his work with his colleagues (1). As discussed by Montouri and Purser, the idea of a "lone genius" stems from the fact that "This modern view of creativity has venerated the artist or genius as a cultural hero, because he or she is someone who has forged something new and original by struggling against [and rising above the limiting, stultifying forces of] the conforming masses. To maintain such a stance, the creative person must in a sense disengage him-or herself from the environment" (2). Were these images of mathematicians similarly being portrayed on their Wikipedia articles?

Wikipedia was chosen as the tool of reference due to its popularity and ease of access to a wide set of internet users. Since being founded in 2001, it has continued to grow in popularity, with 113 billion views of English pages between February 2021 and January 2022 (3). English Wikipedia was chosen for analysis, due to the high level of articles in English (11.69% as of February 2022 - the highest percentage of any language (4)) and that English serves as a common language between people involved in the development and marking of the project. Wikipedia tends to appear high in search ranking algorithms and it does not have a paywall, which more professionally produced information may be behind, so it is not surprising that students commonly use it as a definitive source of information rather than weighing up multiple sources (5). The ethos of Wikipedia, as well as characteristics of articles and common contributors may lead to bias in the articles, as discussed in Section 8, but with the popularity of these reference articles continuing to grow, it is important we assess how well historical situations are being represented, so that incorrect or biased information is further amplified. Figure 3.1 illustrates the number of page views of select 19th century mathematicians and philosophers, and shows that these pages are popular sources to find out information about these figures.

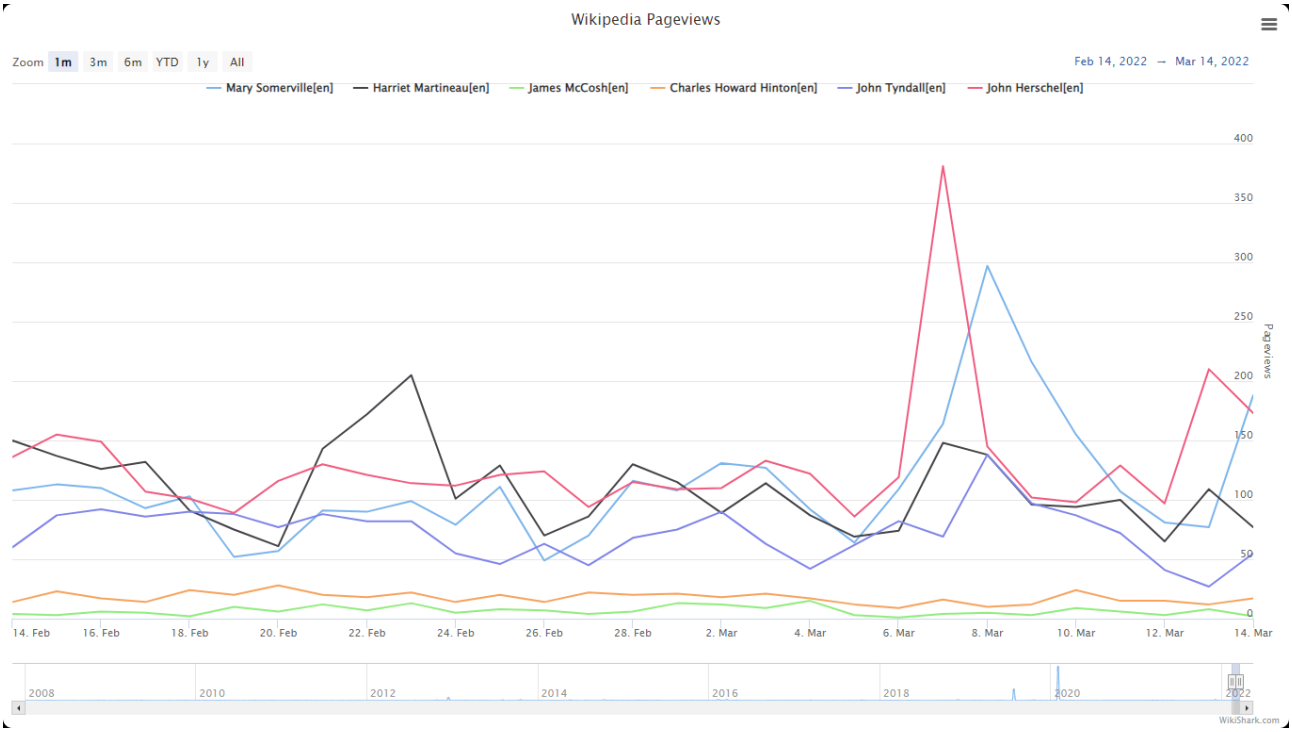

Figure 3.1: Number of page views between February 14th and March 14th 2022 of a selected group of analysed articles, created using the WikiShark tool (6)

Irrespective of the portrayal of mathematicians on Wikipedia, mathematicians have always striven to find ways to communicate with others and work together. Distributed, collaborative research has been increasing since World War I, as scientists have increasingly moved towards big science (projects such as building rockets and the Large Hadron Collider (7)), and collaborations were used to share expensive equipment and the expertise of increasingly specialized scientists (8). With the rise of the Internet and virtual communication, there has again been another shift in the type of collaboration performed, with fewer geographic and temporal restrictions in place for the modern researcher. A collaboratory, as named by Finholt and Olsen, is defined as the 'use of computing and communication technology to achieve the enhanced access to colleagues and instruments provided by a shared physical location, but in a domain where potential collaborations are not constrained by temporal or geographic barriers' (9). It is now much more common for research to take place in these collaboratories across universities, countries and continents, and the easing of communication means researchers all round the world are just an email away.

In the 19th century new advances, such as the learned societies and cheaper communication strategies, meant collaborative practice became more common within scientific communities. The first coauthored paper was published in 1665 but it was not until the professionalization of sciences in the 19th century that collaborative research began to take off. According to deB Beaver and Rosen, the move to professionalization of the sciences led to a significant change in social structures in science over this period, with a movement from a 'loose group of amateurs' into a scientific community by the end of the century. Professionalization defined who belonged to this class of scientists with a common set of attributes that made it clear what held the group together and what set them apart from wider society (10). Scientists had many motivations for collaboration including gaining experience, gaining visibility and recognition, becoming more efficient in use of time and gaining access to special facilities (10). The 19th century was chosen as the period of focus for this project due to this professionalization movement. Additionally, earlier periods risk a lack of collaboration to analyse, and later periods have the introduction of revolutionary tools for collaboration such as the internet to grapple with in analysis.

The aims of this project are as follows:

1. To perform analysis of the portrayal of 19th century British mathematicians and philosophers on Wikipedia by comparing the networks created within the two fields
2. To be able to construct the networks formed between other mathematicians/philosophers based on a Wikipedia article, including all mathematicians/philosophers linked on the article respectively that do not have their own article, by using processes involving web scraping and Natural Language Processing (NLP)
3. To evaluate the success of the NLP techniques applied by comparison with a manually extracted network

It is hoped that by achieving these aims, an overview of the information that could be extracted from analysis of Wikipedia networks will be identified, which will lead to further ideas to explore how well these popular and well used resources are portraying mathematicians of the 19th century.

3.2 Project Summary

Software has been implemented that scrapes information from the Wikipedia pages of 19th century British mathematicians and philosophers, using the category lists found on the Wikipedia page for each group. The code is able to produce a list of named entities for each article that Spacy has recognised with a test set F1 score of 0.7697. Additionally, it can produce a list of linked Wikipedia articles that refer to people, using Wikidata information with a test set F1 score of 0.8328. From these lists a variety of statistics were derived, detailing the quantity and breakdown of mentions of people in the articles, as well as commonly occurring nodes in the network. These statistics were used to analyse patterns of linkage between Wikipedia pages of the mathematicians and philosophers, providing the network analysis. Wikipedia articles can be edited, so all the statistics are correct as of a run of the code on the 13th of March 2022.

The report provides an analysis of collaborative work in the 19th century between these two groups, and assesses the benefits and disadvantages of using this method of analysis over the typically used co-authorship statistics. It details further expansions and pathways that this kind of work could take, as well as briefly discussing the ability to compare the software's analysis with additional historical information that could provide further context from projects such as the Epsilon correspondence of 19th century scientists initiative. Throughout development, a variety of different Named Entity Recognition (NER) tools have been tested on the data, and a comparison of the strengths and weaknesses as well as a discussion around the reason for choosing Spacy as the selected method is also found in the report.

3.3 Objectives

Below are the objectives for the project.

3.3.1 Primary Objectives

1. To perform analysis of networks of 19th century British mathematicians and philosophers
2. To derive a historical overview of the collaborative work of 19th century British mathematicians and philosophers
3. To evaluate the success of NLP processes in extracting mentioned names from given Wikipedia articles related to 19th century mathematicians and philosophers

3.3.2 Secondary Objectives

1. To choose NLP tools that maximise accuracy of named entity recognition within the Wikipedia articles
2. To be able to construct network diagrams based on mentions of people in the article
3. To assess the portrayal of mathematicians/scientists collaboration practices on Wikipedia against the records of their known correspondents

3.4 Report Structure

The remainder of the report is structured as follows.

4. Discussion of collaboration between subject groups over the 19th century, factors that led to the growth in collaboration
5. Discussion of the design of the code produced and justification for particular implementation choices
6. Discussion of the uses and capabilities of the code
7. Analysis of the NER methods used in various stages of development
8. Analysis of the results from code running on the network
9. Comparison of output from code with known correspondence data
10. Evaluation of the success of the project and future work
11. Conclusions

4 Scientific Collaboration

Scientific collaboration is a hard term to define. Many studies that have used computational analysis to assess general patterns have often used co-authorship as the standard statistic to assess collaboration trends. Co-authorship is often used as a standard as it gives a very objective assessment of how collaborative practices have changed over the previous century, as collaboration between scientists becomes more popular. A decrease in barriers to long distance travelling, as well as the massive growth of the internet and virtual connections has allowed researchers from all over the world to come together and work on common problems. Based on statistics compiled by Döbler from the Royal Society's (RS) catalogue of Scientific Papers and the Zentralblatt für Mathematik und ihre Grenzgebiete, you can see the clear increase in collaborative behaviour, illustrated in Figure 4.1. However, it is worth bearing in mind that the increase may not be just as straight forward as illustrated as there have been a number of different standards for authorship throughout the history of scientific publication.

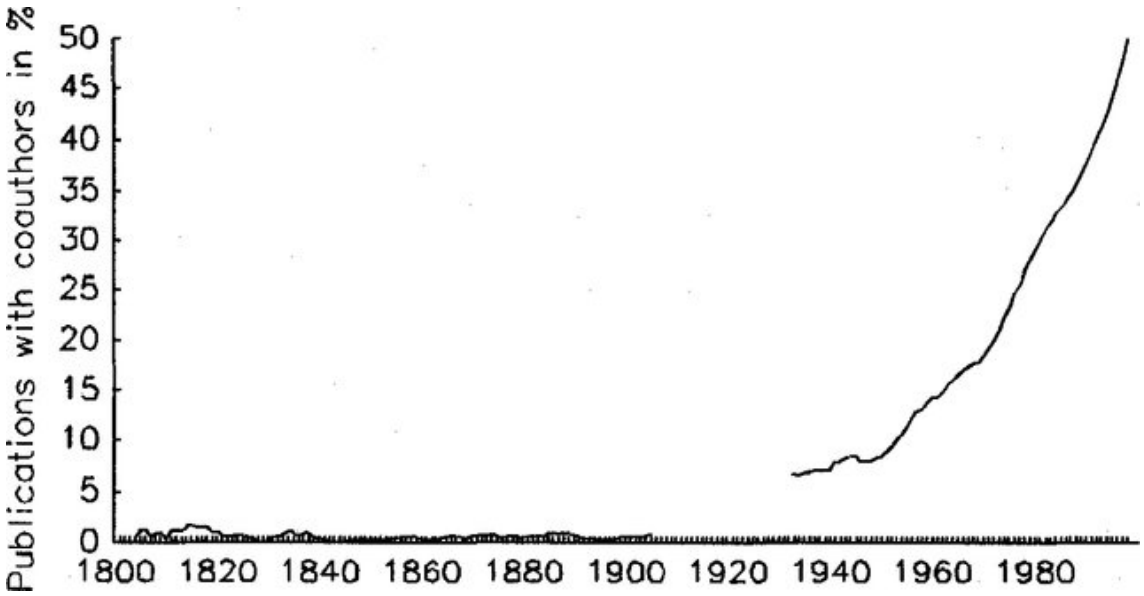

Figure 4.1: Figure from (11) illustrating the growth of publications with coauthors in the RS over the 19th century and Zentralblatt over the 20th century

4.1 Co-authorship and collaboration

Many papers have examined how collaboration and co-authorship have evolved ((11), (12), (13) to provide a few examples), but this project is not going to use co-authorship as a measure. Using co-authorship as the standard on to the 19th century relationships may be projecting our current standards of working relationships onto previous periods. Mathematicians and philosophers had many different working relationships that may not have resulted in publication, but were instrumental to their respective careers. By analysing mentions of relationships present on sites with biographical descriptions of these figures, it is hoped that different and new insights on networks of collaborations can be presented. Using mentions of relationships as opposed to co-authorship also allows for informal relationships, and relationships with lesser known figures to emerge, where they may have been neglected through co-authorship analysis. A relationship from the period of focus that would be downplayed by co-authorship analysis is that of William Henry Young and Grace Chrisholm Young. Grace was a pioneering female mathematician in the end of the end of the 19th century, becoming the first person to achieve a first from both Oxford and Cambridge and going on to gain her doctorate in mathematics from Göttingen Unversity in 1895 (14). Grace married one of her previous tutors, William Henry Young, also a prominent Cambridge mathematician, and they worked together on mathematical problems and sources. While there are instances of them having published together, William

decided that it was better for his name to solely appear on the papers. He wrote to Grace 'The fact is that our papers ought to be published under our joint names, but if this were done neither of us get the benefit of it. No. Mine the laurels now and the knowledge. Yours the knowledge only' (15). Both the named entity recognition code and the linked article code pick up Grace as a member of William's network, where as co-authorship analysis would fail to pick up Grace's contributions to a lot of William's work. While Grace does have publications to her name, this was not always possibly for women and other underrepresented groups during this period, and so these sort of relationships could be missed from co-authorship analysis.

Co-authorship continues to be a flawed statistic due to the nature of deciding who is a named author on a paper. The standards for who ends up on the byline of a given article conform to constantly changing standards, however exclusion from authorship tends to occur to groups whose members are commonly put at a disadvantage within the academic community. While the following groups are more likely to be awarded credit today compared with the 19th century; people of colour, females, non Europeans and indigenous people are still less likely to have their contributions deemed significant enough for authorship (16). Using co-authorship as a measurement means that these voices are yet again swept to the side and left out of analysis of collaborative networks, downplaying the importance of their contributions.

4.2 Factors involved in growth of collaboration

The following section will detail developments in collaborative style and practices over the 19th century, looking more broadly at the field of science as a whole.

At this time, a stronger community was developing around the people involved in the process of science. During the 19th century, an increase in both number and membership of societies and associations provided ample opportunities for discussion and networking (for those that were allowed to become members), as well as publication outlets that let news of foreign research reach a range of readers. The 19th century also saw increasing specialisation of the societies (17), with groups such as the Royal Astronomical Society (RAS, 1820), the British Science Association (previously British Association for the Advancement of Science (BAAS, 1831)) and the London Mathematical Society (LMS, 1865), forming in Britain, to name a few.

It is the BAAS that is said to have pushed the professionalization of science, ultimately leading to the structural change occurring across the 19th century (18). The 19th century saw a change in the constitution of science. At the beginning of the century, science was dominated by wealthy individual amateurs, who were able to self fund their own work (19). By the end of the century, science was performed completely differently, with scientists being employed for their work and working within stronger scientific communities. While potentially less relevant to sciences without experimentation such as mathematics, this shift was strongly felt across more practical subjects such as astronomy and physics. In deB Beaver and Rosen's study of scientific collaboration over the 19th and 20th century, they define professionalisation as the process which organized a group of individuals by a set of attributes which define what holds the group together and also what sets them apart from non members (10). The professionalisation movement led to more of a cohesive group, providing the environment necessary for increased collaboration between members.

The professionalisation movement took place at different speeds in different countries, and this is often reflected in the collaborative output of each country over the 19th century. At the beginning of the 19th century, France was further on in the process of professionalisation and more supportive of a collaborative environment compared to Britain and Germany (10). According to deB Beaver and Rosen, after a tumultuous 18th century, France was quick to move science into it's professional era. Institutions were long established and well supported, and there was a general encouragement of scientific teaching. Early career collaboration often played a significant role in the career of a French scientist, and it was an accepted part of the pathway into higher roles (20). France was making good early progression towards more scientific collaboration especially compared to early 19th century Britain, which lacked a well established education system, government support and professional autonomy (10).

It was not just the movement to professionalization that led to a change in collaborative habits of scientists. There was also a dependence on the subject and material of study. Subject areas that required experimentation, such as experimental physics and astronomy, often saw an increase in co-authorship and collaboration sooner than theoretical based sciences. This is illustrated in the growth of co-authorship in physics compared to mathematics across the 19th century in figures 4.2 and 4.3.

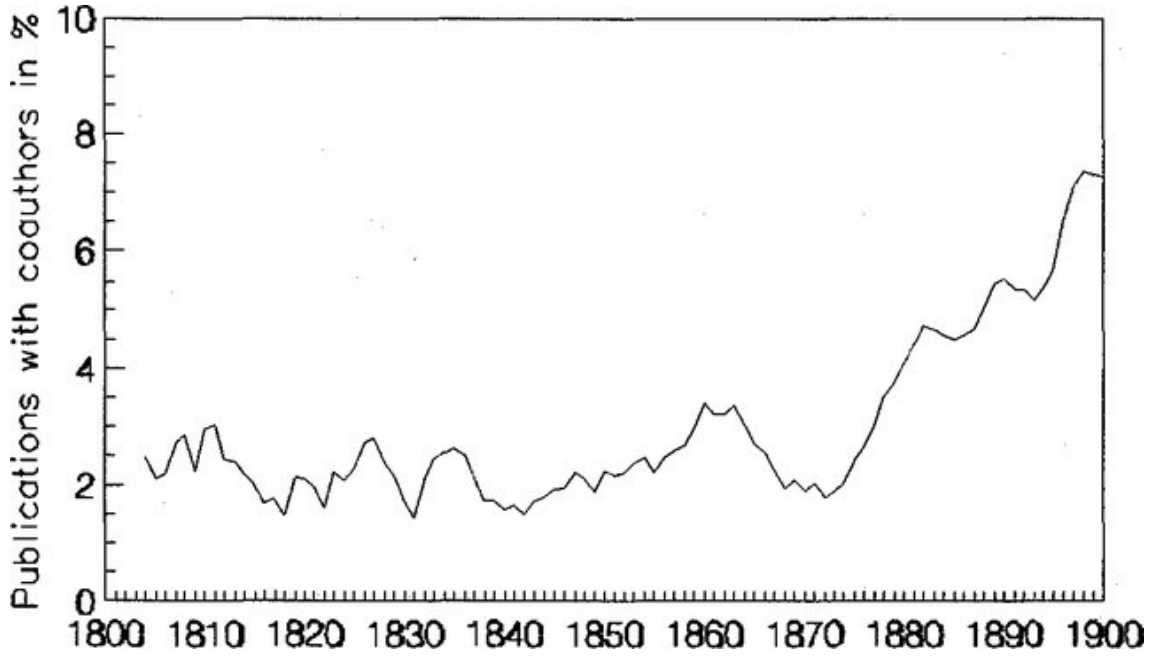

Figure 4.2: Figure from (11) illustrating the growth in proportion of coauthored physics papers

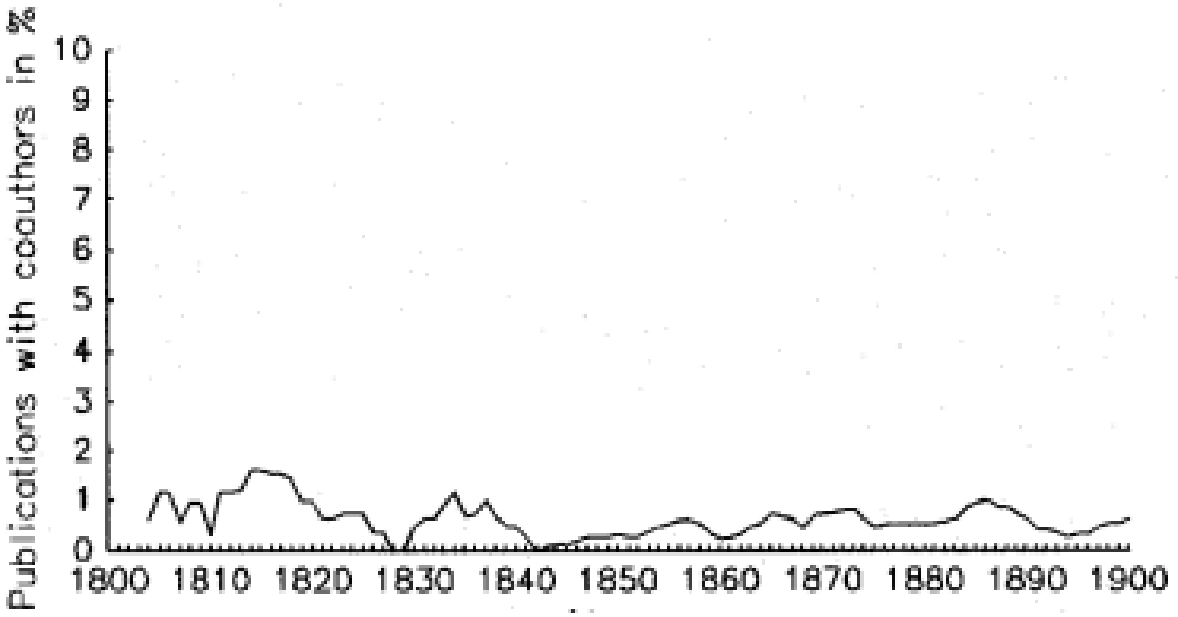

Figure 4.3: Figure from (11) illustrating the lack of growth in proportion of coauthored maths papers

A pattern that seems to emerge in various studies and discussions looking at collaborative networks across this period, is the influence of a 'common object'. As illustrated in Döbler's study, if scientists approach reality through experiments they tend to collaborate; if logicians come into touch with working devices such as computers, they also tend to collaborate (11). In this case, the common object extends to the devices used for experimentation. People interested in this equipment also had a need to share resources, due to the cost associated with gaining access to their own devices. We also see this pattern again emerging in astronomical communities. With a large common object such as the sky, the increase in teamwork needed for technical observations and the shared telescopes, it is not surprising to find strong collaborative networks and patterns within this discipline, both on a national and an international scale.

While you could potentially use the almost platonic mathematical objects as the ideal for a common object between mathematicians, this may be a factor that is not as obvious and hence helping to develop this idea of mathematicians as lone geniuses. An alternative perspective to this is using previous mathematical work as the common object of mathematicians. Performing mathematics is inescapably linked to the work of previous mathematicians, and the central topic of mathematics is often digging backwards into what is taken for granted in previous work (21). While more abstract than the previous examples given, this idea of famous work of previous generations being the link between these networks stands for both mathematics and philosophy. A question to ask about this is whether the abstractness of this common object makes the social structures of either group less obvious, potentially leading to misrepresentations of groups in public portrayals?

When comparing the mathematicians to philosophers, a similar style study into the collaborative patterns of 19th century philosophers would be useful to have. Unfortunately, research into collaboration within philosophical communities over any time period is a lot more scarce, and there is very little relevant literature for my period of interest. It is however, acknowledged that collaboration in philosophy does not tend to be represented using the co-authorship statistics. Philosophers still remain as one of the subjects with the lowest co-authorship rate.

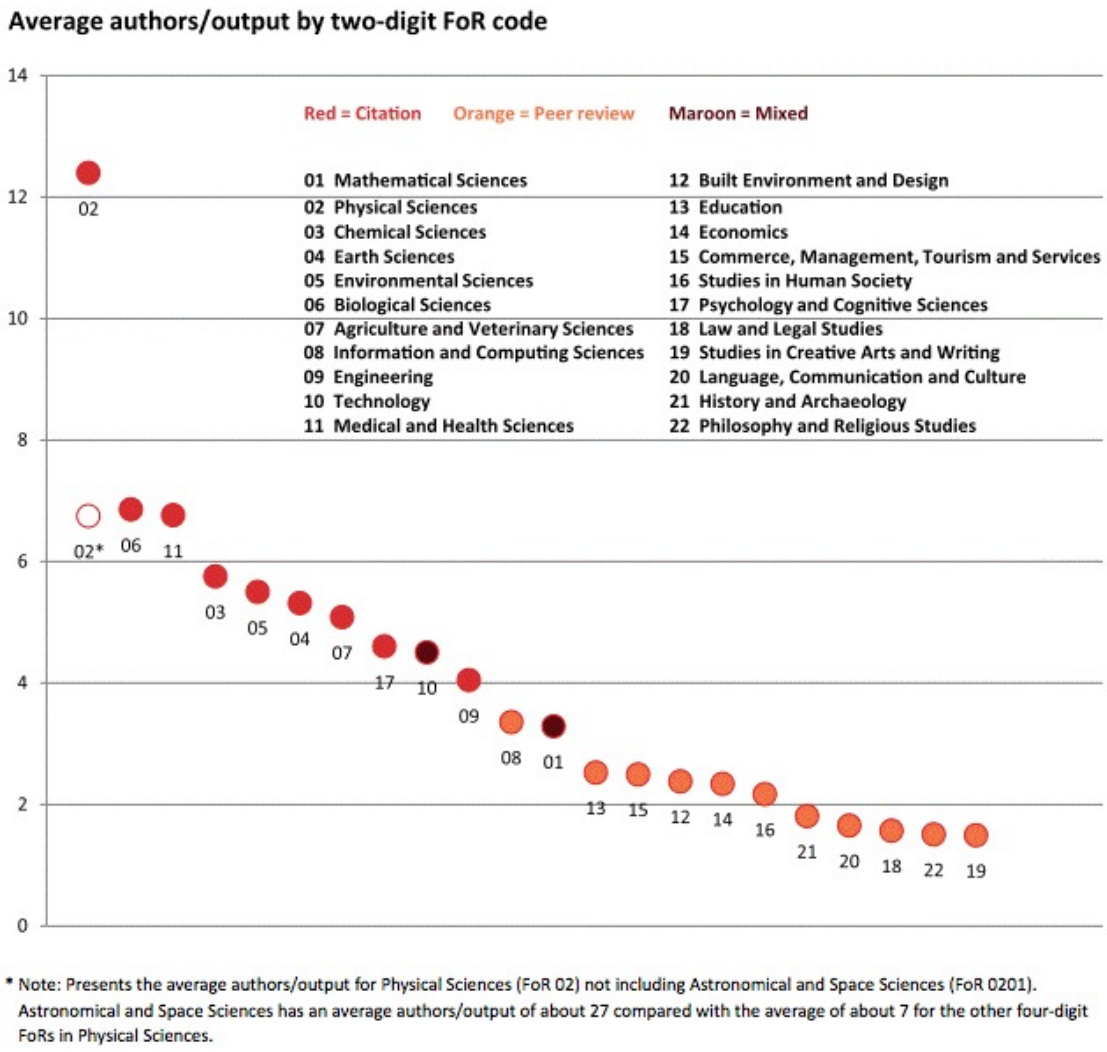

Figure 4.4: Diagram illustrating average co-authorship statistics for a range of subjects compiled from data in the Excellence in Research in Australia 2012 report (22)

With this lack of solid literature, any ideas proposed in the analysis section to explain patterns of collaboration needs to be combined with further research into the collaborative style of philosophers. It is hoped that the analysis can help begin to look at the patterns in comparison with a well studied group such as mathematicians.

4.3 Conclusion

In conclusion, it is hoped that this study will provide a new way of analysing collaborative behaviour of groups of mathematicians and philosophers in the 19th century, while avoiding the problems of co-authorship as a metric. It will hopefully provide a more representative sample of collaborations that didn't necessarily result in publication, and help assess whether the portrayal of these groups on Wikipedia is potentially being influenced by the public opinion of their subjects.

The 19th century, with its changing style of performing science and the growing professionalization move-

ment is an ideal period to study the collaborative style of mathematicians and philosophers. The lack of literature has made it harder to assess any changing patterns in philosophical collaboration over the 19th century, but it is hoped by comparison with mathematics figures from the results, that a discussion can occur regarding the collaborative habits of both groups.

The changes in the nature of science over this century suggests that mathematics articles should represent growing collaborative networks, despite what is seen in co-authorship data. The co-authorship data combined with the common object argument for groups such as physical scientists may suggest a higher level of collaboration within groups such as physicists or chemists, but collaboration still took place between mathematicians. This collaboration should be represented on the Wikipedia articles, but a lack of visible common object for mathematicians combined with a popular lone genius stigma could be resulting in a lack of acknowledgment of collaborative networks on the articles.

5 Design and Implementation

The following section will detail the design and implementation stage of the software side of the project; discussing the tools used in development, the functionality of the code, and the general processes that were followed to result in a final product.

5.1 Language and Libraries Used

All the code written as part of this project has been developed in Python, with the exception of one statistical analysis file in R. Alternative languages could have been used in development, but the support Python offers for NLP development, as well as my general proficiency in the language meant it was the most logical choice.

On top of well documented NLP libraries, Python also offered good support for integration with Wikipedia/Wikidata and web scraping in general. Below are a list of the most important external libraries used in development:

- Beautiful Soup - a library used to interact with the Wikipedia pages, and extract the contents of relevant pages, in preparation for NER on the text
- Wikidata - a library used to easily access the Wikidata information (structured data about given Wikipedia articles) for given figures
- Wikimapper - a library used to map the Wikipedia article addresses to their corresponding Wikidata IDs
- Spacy, Natural Language Tool Kit (NLTK) - NLP libraries used to identify the named people in given Wikipedia articles

5.2 Code flow

Figures 5.1 and 5.2 give a high level overview of the movements between files during execution. All code files stated can be found in the submission folder for this project, and should be able to be run locally if you follow instructions stated in Appendix B.

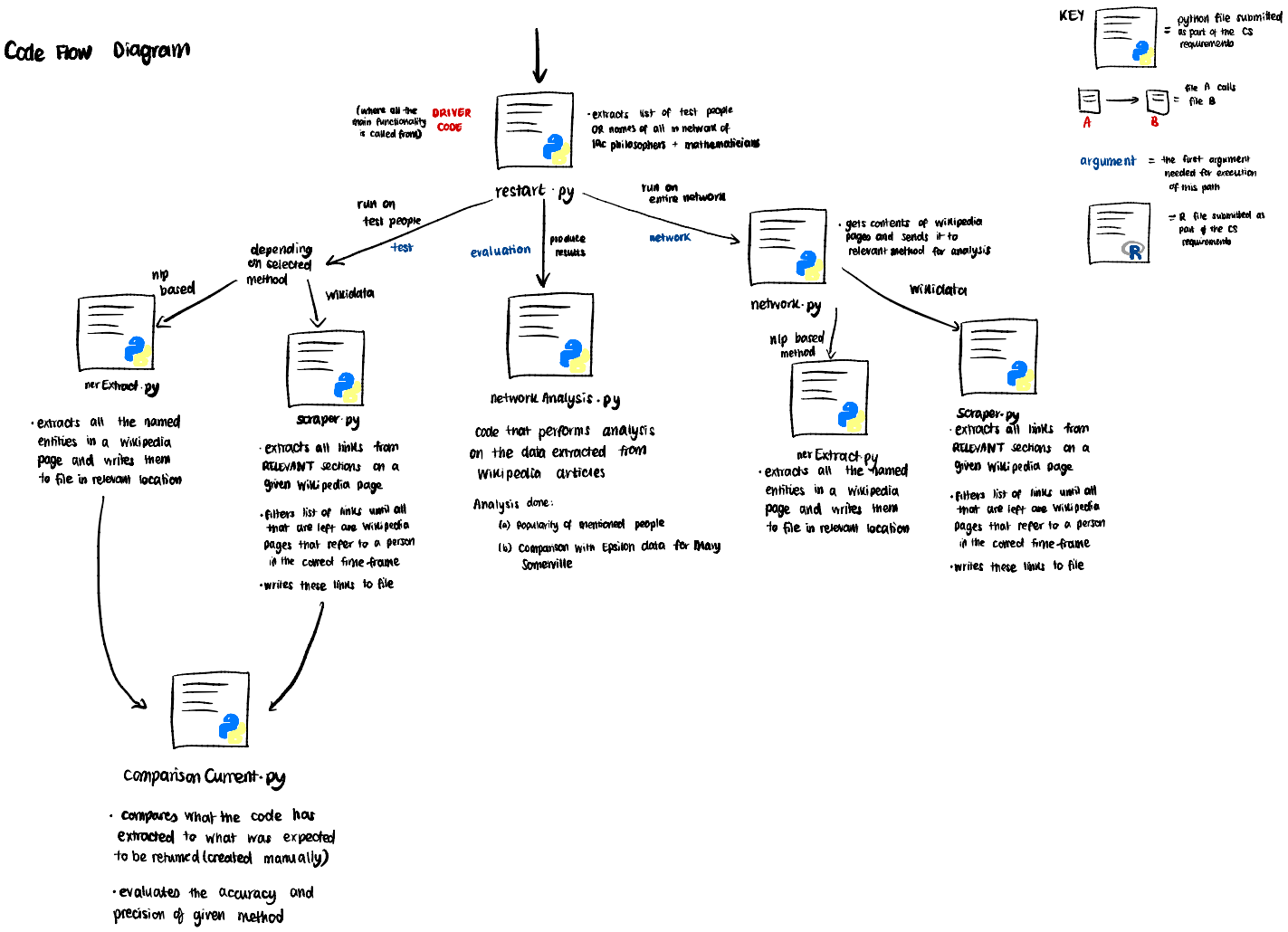

Figure 5.1: Diagram illustrating relationship between Python files submitted for the project

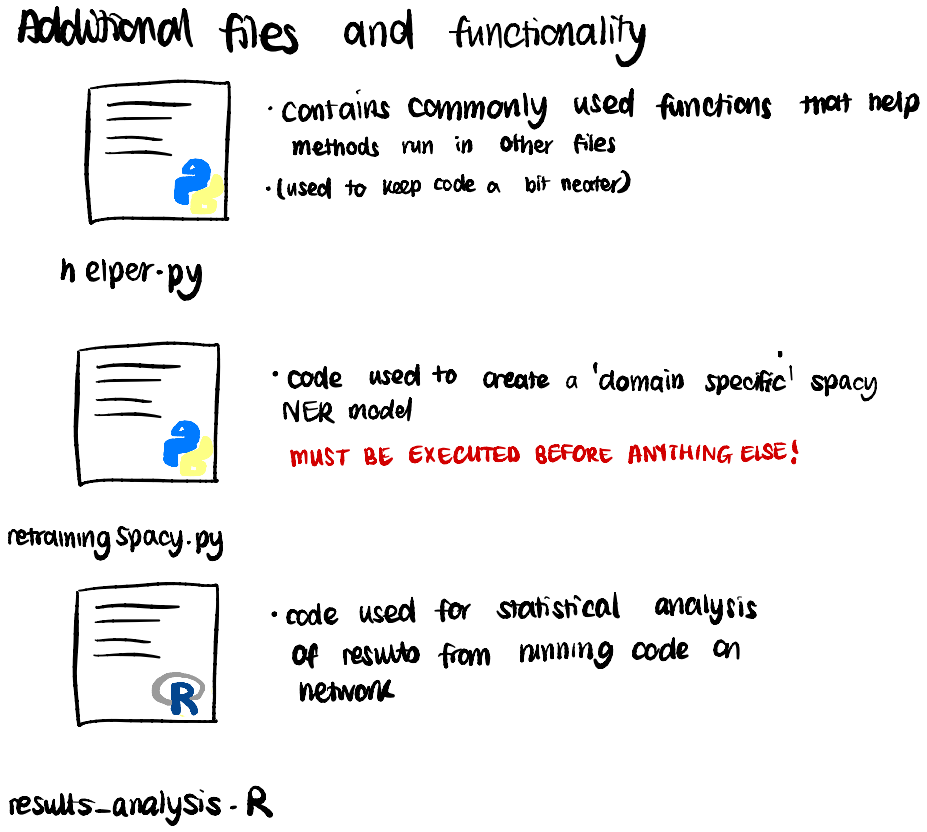

Figure 5.2: Diagram illustrating additional code files and their functionality

5.3 Methodology

The following section will detail the processes followed during development of code for the project.

5.3.1 Software engineering process

For a project of this scale, it was important to have a process to ensure regular work was being completed on the project. I used a process of iterative and incremental development (IID) to work on the software side. Incremental development is a style of development where sections of the code can be developed individually and integrated when the development has finished. This allowed me to have code early in the development that could be adapted to meet the goals of the next stages of the project. As part of this cycle, I held weekly meetings with supervisors. The weekly meetings consisted of a recollection of the work that had been completed in the previous week, asking any questions about problems I had encountered, and stating what I planned to complete in the next week. By having a consistent structure throughout the meetings, and a form that I prepared before the meeting, it allowed me to analyse how well my time had been spent on the project.

The following benefits came with using IID as my development methodology :

- Incremental development
  - This was one of the largest projects I have completed during my time at university. Paired with the fact that the scope was not set in stone at the start of development and adapted as I became more familiar with capabilities and literature, it often felt like the amount of work that needed

to be done was quite daunting. By using IID, the coding tasks began to feel more manageable as they were broken into smaller chunks. This meant I was able to quickly identify tasks that I could easily work on to continue to make progress on development.

- Working product
  - Using incremental development meant I often had a small version of the product running, allowing me to test features and abilities of the current system. This meant I could decide what were the next steps needed to get closer to the goals and identify errors earlier on in the development process
- Improved communication
  - Meeting a minimum of once a week meant that my supervisors were updated with the progress of the project and I kept myself accountable for ensuring I had progress to report back on a regular basis.

5.3.2 Version Control

Despite being the only developer on the code, I still used Git for version control and to ensure my work was regularly backed up. This allowed me the freedom of working on more experimental features while maintaining a working version in case of needing to revert any changes.

5.4 Ethics

There will be no ethical considerations during this project, as indicated in the ethics self assessment form.

6 Data and Methods

In order to have a sense of the project outcomes, this section will aim to provide a brief overview into the workings of the software. As described above, the main purpose of the code is to extract the named entities from Wikipedia pages of 19th century mathematicians and philosophers. There are two main ways that the code can be used; testing the methods on a strict set of test figures whose Wikipedia pages have been manually combed to identify named and linked people or applying the developed methods to the network of philosophers and mathematicians. Testing the network on a strict set of test figures with manual data allowed for modification of the methods used and identification of weaknesses, as well as ensuring that there was a high enough level of confidence in the methods before allowing the code to run across the whole network. The code can also be used to produce statistics about the results of either of these runs of the code.

6.1 Wikipedia and Wikidata

All the articles extracted for analysis in this project have come from Wikipedia. The lists of 19th century British philosophers and mathematicians used to run the methods on can be found here for philosophers, and here for mathematicians. The code uses the wikipedia library to extract the main text of the article, before removing any content that was below common names of headings that we were not interested in the data from. The data after the following headings was not collected; Works, Bibliography, Further Reading, References, Main Works, See Also and Select Bibliography. These headings were chosen after analysis of the additional names that were being picked up and where they had come from in the article.

With the Spacy NER, the program just works with the text given back by the previous steps. The built in functions for Spacy are applied to generate a list of names mentioned in the article, and this, along with the length of the article text in characters is written to a file in the relevant folder for the discipline of the figure being analysed (either mathematics or philosophy).

In the case of recognising the articles to other people linked from a given Wikipedia page, Wikidata is used as the main tool to recognise these types of articles. Wikidata stores structured information about the articles hosted on Wikipedia. For the purposes of this project, the 'instance of' statement about an article is the most needed feature of Wikidata, as if the instance of a given article is human (also known as Q5 in Wikidata tags), then we can identify that given article as relevant for our analysis. The code works its way through the links in the article, and unlike the text analysis, where we can cut off part of the articles easily by string manipulation, filtering the articles so only the relevant sections are checked is harder when looking for links. The chosen method for ensuring we are in a relevant section before returning links is to look at the previous header. The previous header is checked against a list of unaccepted headings; Works, Bibliography, Further Reading, References, Main Works, See Also and Select Bibliography. If the link has a previous heading that is not in the unaccepted headings then it is extracted and moved onto the next stage of the code.

All the articles at this stage are mapped to a Wikidata ID. This is the unique code used to extract the Wikidata information for any given article, and through using the WikiMapper database locally, which contains the Wikidata information for the majority of English articles, the ID can be extracted. If the given article is not in the WikiMapper database a request is sent to the Wikipedia API to return the Wikidata ID for the given article.

The information is extracted from each article with a corresponding Wikidata ID. The information that needs to be extracted is the instance of value (has to be human), the gender (used to divide the articles for analysis) and the date of birth and death. All the articles in the network are articles of 19th century mathematicians and philosophers, but there is no guarantee that the figures being linked to from the Wikipedia are also 19th century figures. We are looking to analyse the collaborators of the figure whose article we are looking through, and so steps are taken to only identify links to articles of figures who were alive within the same time period. If the date of birth/death of the figure whose article we are looking through cannot be identified, then they are set to default to 1800 as the DOB and 1899 as the DOD. This will allow us to

still pick up 19th century collaborators if there is a problem with the date field. These date of births are compared and only relevant people remain in the set of linked articles from a given figure.

After the irrelevant articles have been filtered out, as with the Spacy data, this list of linked people is written to a file. For wikidata, due to the ability to identify the gender of the article's subject, the files are placed in folders related to both their discipline and their gender.

Applying these methods to all the articles in the list of 19th century British philosophers and mathematicians is a computationally intensive task, with many links to be checked, and takes significant time to run in its entirety. However, once the data is generated (and it is submitted with the data from a previous run), the analysis steps can be run without the need to re-collect this data every time. Further details on the analysis will be found later in the report, but it uses the mentions, and occurrences of mentions found within each of these files to compare the types and quantity of mentions across 19th century British mathematicians and philosophers.

6.2 Manual test data

Before applying these methods to the networks, I needed to ensure that they were working at a high enough level of accuracy to produce reliable results. For all of the following figures, I have worked through the relevant Wikipedia pages and extracted a list of all linked and unlinked names. All pages were either chosen for the reason stated below or randomly generated using the category lists extracted from Wikipedia. If pages randomly selected had less than 10 links, another shuffle of the page titles was performed, so the results for accuracy were not inflated by the code performing well on a small number of links. The number of entities does not distinguish between the same entity written in different forms in the same article.

1. William Hamilton (WH)
    - Randomly generated 19th century British philosopher, contains 37 entities in article
2. Michael Faraday (MF)
    - Selected for potential cross reference with Epsilon data, contains 45 entities in article
3. Mary Somerville (MS)
    - Selected for initial test of methods in early development stages, and continued to be a member of the test group due to links with Epsilon, contains 85 entities in article
4. John Tyndall (JT)
    - Selected for potential cross reference with Epsilon data, contains 46 entities in article
5. John Herschel (JH)
    - Selected for potential cross reference with Epsilon data, contains 51 entities in article
6. James McCosh (JM)
    - Randomly generated 19th century British philosopher, contains 14 entities in article
7. Harriet Martineau (HM)
    - Randomly generated 19th century British philosopher, contains 80 entities in article
8. Charles Howard Hinton (CHH)
    - Randomly generated 19th century British mathematician, contains 40 entities in article

These names were used to evaluate the performance of the algorithms, and the performance of the methods on this given test set forms the basis of my analysis of the NLP tools.

6.3 Evaluation of performance

The code runs the specified method on the articles of a selected test group of figures and compares the outcome to what was manually extracted from the selected articles. To evaluate the performance of the methods, the program uses the F1 score. The F1 score evaluates the success by looking at the true positives, false negatives and false positives that have been identified by the method.

Figure 6.1 illustrates some examples of names that would fit in each category. True positives are the names that were manually identified and identified by the method. False negatives are entities that have been manually identified but not picked up by the methods. False positives are entities that have been picked up by the methods that do not correspond to a name manually identified.

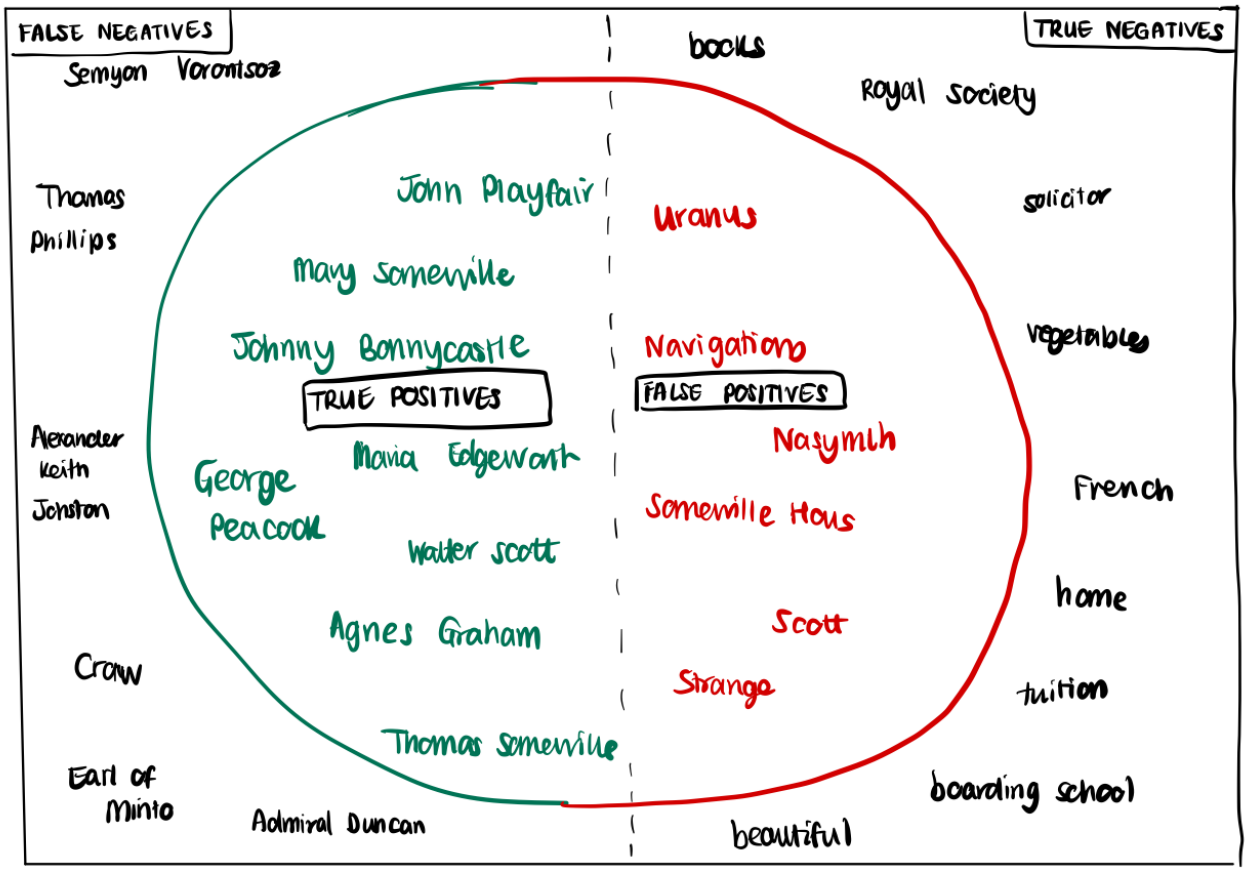

Figure 6.1: Diagram illustrating false positives and negatives from the Spacy run on Somerville's Wikipedia page, adapted from (23)

There are three measures used to evaluate.

- Precision $= \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$ (a measure of how well the positives identified match the test data)
- Recall $= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$ (a measure of how many people have been identified properly)

- F1 score $= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

The F1 statistic was chosen to evaluate the methods so the false negatives and false positives could be accounted. Both are equally going to make the output less representative of the true network present on the pages, and using an F1 statistic allowed me to factor this in. It is also commonly used across studies of NLP tools, allowing me to compare the performances of the methods on my corpus compared to the performances discussed in papers.

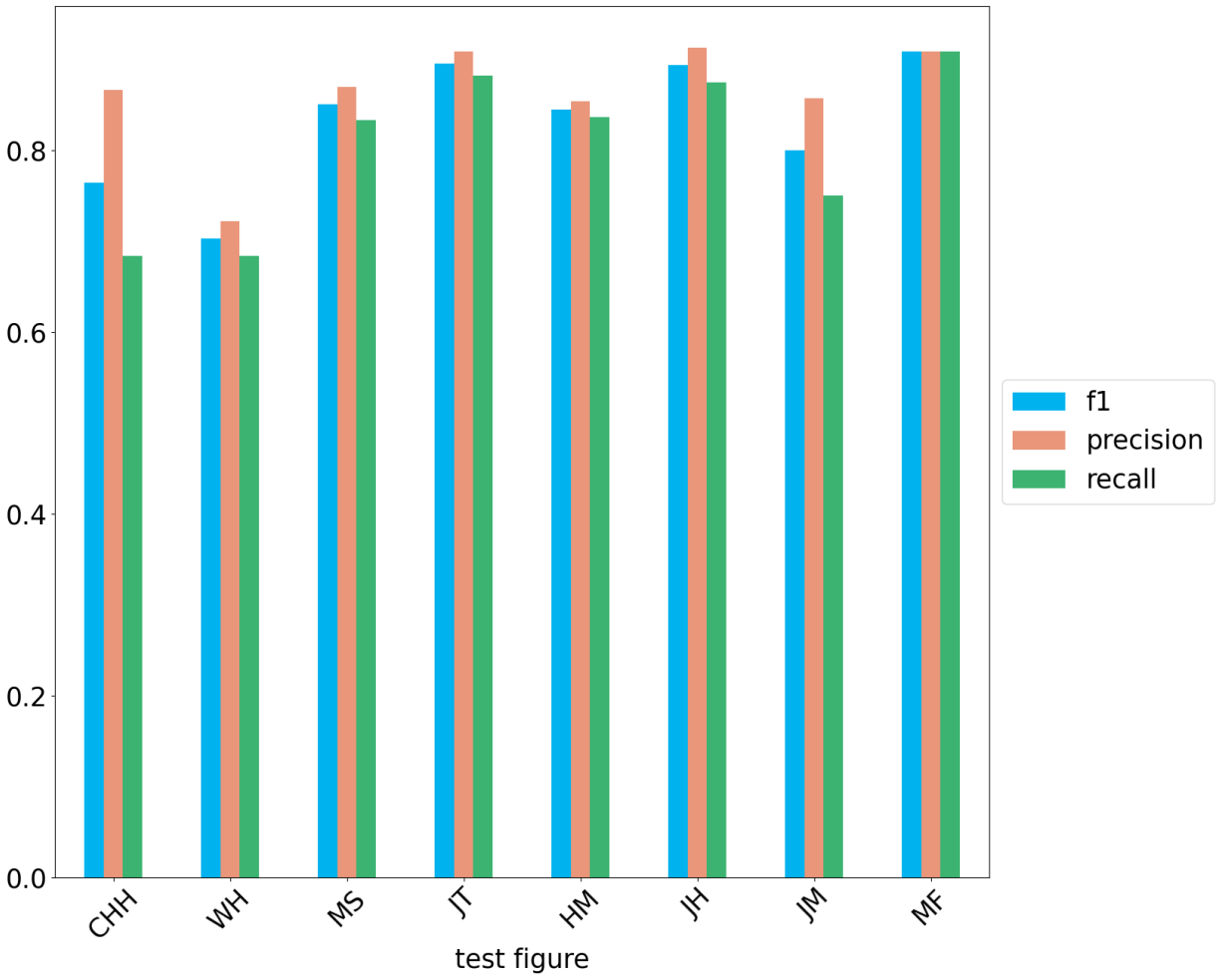

Figure 6.2: F1, Precision and Recall scores for the test articles using Wikidata to pick up linked figures

Figure 6.2 illustrates the accuracy of using Wikidata to pick up linked articles. Due to the size of the names and the size of graph needed to make the names readable along the x axis, the shortened versions of the test figure's names have been used, taking their initials. Figure 6.2 shows that the Wikidata way of identifying articles is performing well on the test network, with recall always slightly worse or the same as the precision score. As discussed in the results analysis section, occasionally irrelevant links are returned, but overall I believe that the Wikidata method is performing well enough that I can place confidence in it behaving correctly when applied to new data in the network.

6.4 Assumptions

The following section will detail how cases that may cause the code to perform incorrectly were dealt with, and any assumptions that have to be made.

6.4.1 Calendars

The library used to interact with the Wikidata information (wikidata) is unable to process Gregorian calendars. While this is not a problem due to the time period of interested, if a similar style of analysis was performed on figures from previous centuries where the Gregorian calendar was still in use, there would need to be some adaptation of the Wikipedia/Wikidata code.

6.4.2 Table of Contents

Originally in order to filter out the information on Wikipedia that was in irrelevant sections it was assumed that all of the articles had to have a table of contents to be processed. This allowed me to quickly identify the valid sections that a link could come from. When this was implemented and applied to the network, too many articles of mathematicians did not have table of contents to keep this in the software. While the process of finding the relevant headings was more complex than the table of contents method, it had to be implemented to ensure that the size of the mathematics data set was not reduced too drastically.

6.4.3 Internet Connection

The code will require internet connection to run due to the requests made to the Wikipedia pages at various points during the program. If connection is lost mid execution the code should be able to continue when the connection is restored, but it is recommended rerunning to ensure the data that is generated is actually representative of the pages of the two groups.

6.4.4 Multiple reference problem

When comparing what is identified by the methods and what has been manually identified, there needs to be some way to look at the similarity of the string. Strings that may be referring to the same entity but not exactly equal include when a method has extracted additional information around the name (such as Dr, Mrs, Mr etc). A selection of commonly used titles are stripped from any string returned from the method, but in case another title is used or additional irrelevant information is brought back with the name, the Levenhstein ratio is used to identify if the strings are referring to the same entity. Levenshtein distance is a measure that can be taken between two strings. It is the minimum cost of edits needed to go from the first string to the second string. Edits do have difference costs, with insertion and deletion having a weight of 1, but substitution (removing one letter for another letter) has a cost of 2. Levenshtein ratio will combine the information of the distance between the two strings, as well as the length of the strings, to provide a similarity measure between 0 and 1. It can be written as (lensum - ldist) / lensum, where lensum is the sum of the lengths of the strings and ldist is the distance of the two strings as described above.

There are two conditions that allow for strings to be identified as referring to the same entity:

- Retrieved/manual string is a substring of string to compare with
  - An example where this would occur is Queen Elizabeth II. When manually extracting names, I did not mark down the titles associated with the figures, so the manual entry to compare with would be Elizabeth II. As the manual string is a substring of the retrieved string, they must have a Levenshtein ratio of greater than 0.75.

- Retrieved string is not a substring of identified string
    - The Levenshtein ratio must be greater than 0.9.

The values for the Levenshtein ratio to allow strings to be identified as referring to the same object were chosen after experimentation across Wikipedia articles. The values were chosen to maximise the number of true positives that should be identified, but also ensures that false positives are not picked up as being a correct identification for the method.

6.4.5 Types of relationships

When picking up names on the Wikipedia article with Spacy or Wikidata methods, this is not a guarantee that the two people were in a working relationship. All names are assumed to be in some form of collaboration with the given subject of the article, due to the difficulty of being able to filter the names by relationship type. Additionally names picked up by the Spacy method are unable to be filtered by date, and so some names in the Spacy analysis may be of people who did not live in the same time period as the subject. The data related to date of birth would be unnecessarily computationally complex to retrieve, and there isn't even any guarantee of being able to find it (for example people who do not have many records on the internet) so there has been no further steps to filter the Spacy entities by date. When the filter for date was added for linked articles, it was found that the majority of linked people in the main text of the article were born within the time span of the subject of the article. It is imagined that the same pattern would apply for the Spacy names, but the data returned has not been analysed enough to know this for certain. There may be some inaccuracy in the lists representing named people, but I believe the inaccuracy is accounted for more due to the F1 score of Spacy discussed next, as opposed to leaving in irrelevant figures.

7 NLP Analysis

Like many modern day disciplines of Computer Science, Turing's *'Computing Machinery and Intelligence'* was crucial to the development of the field of NLP. Combined with emerging ideas around the structure of the brain, and an interest in creating a computer capable of imitating the brain, NLP began to emerge with the new field of AI in the 1950s (24). Now used for a variety of tasks, including spelling and grammar correction, language translation, handwriting recognition and text summarisation.

NLP has many aspects that can be used within digital humanities projects, including sentiment extraction, named entity recognition (NER) and machine translation (MT). NLP has helped humanities scholars and social scientists deal with their often large corpus of texts, allowing for new insights into their data without the need to spend a prolonged period of time manually extracting information. It also allows insights on a scale beyond the reasonable work of a single researcher, with this project demonstrating the potential of combining technological tools with a large corpus of information. Manually, it would have not been possible to extract these collaborative networks in a reasonable time, but combining tools from NLP with the Wikipedia articles has allowed for new insights to emerge. NLP has been combined with work in the humanities for a wide range of disciplines and uses, with examples including identifying historical place names (25), transcription of historical documents (26) and mining historical newspapers (27).

7.1 Named Entity Recognition

NER is a sub-area of NLP and Information Retrieval (IR), that involves identifying expressions in text that are members of various categories including people, organisations and locations. The aim of NER is to go from unstructured text into a semi structured output illustrating what entities are referred to within the text and where. For the purpose of this project, the main concern is the ability of a program to recognise people in a corpus of unstructured text, relating to 19th century scientists. NER is used in a wide variety of ways, due to the power of being able to pull structured data out of unstructured sources. Some alternative uses for NER include content recommendations and creating efficient searching algorithms (28).

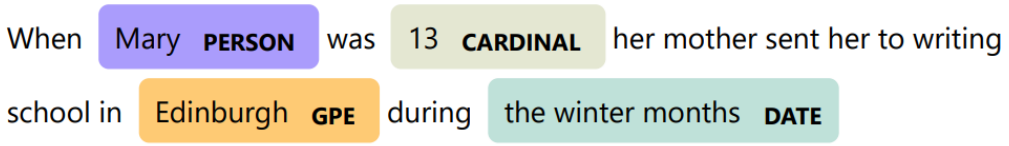

Figure 7.1: Illustration of the output that NER can produce when applied to a sentence from Mary Somerville's Wikipedia page, where GPE means Geopolitical Entity

This project applies the NER methods to the collection of 19th century mathematicians identified from Wikipedia, and to ensure it is performing well enough, test data was produced, that was annotated from eight 19th century figure's Wikipedia pages (Charles Howard Hinton, Harriet Martineau, James McCosh, John Herschel, John Tyndall, Mary Somerville, Michael Faraday, Sir William Hamilton). As well as identifying any named person in the article, it was recorded whether they were linked or not, and whether they were alive during the lifespan of the person whose Wikipedia page been analysed (however this feature is only used within the Wikidata analysis).

The following sections will detail the individual steps of the methods chosen to compare the performance of. Figure 7.2 below illustrates the three main stages that a standard NER tool will take to get from the unstructured data to the annotated data. As illustrated, the success of the NER is measured in this project

by looking at the F-measure, the precision and the recall.

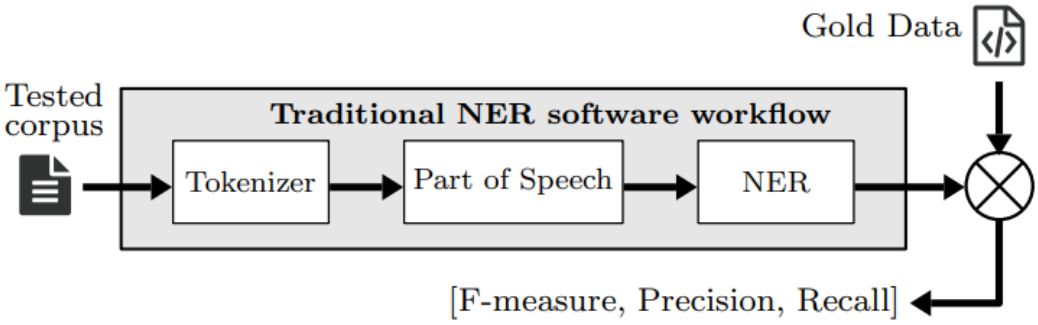

Figure 7.2: Illustration of the traditional workflow from unstructured text to recognised entities from (29)

Like many NLP tasks, NER is a challenging task to complete, for the following reasons:

- Syntactic Ambiguity
  - Ambiguity appears in all layers of NLP, from the syntactic to the semantic level. One example of ambiguity that may affect NER is attachment ambiguity. Attachment ambiguity is when a token can take up multiple positions in a parse tree (30). A classic example illustrating this problem is the phrase *The boy saw the girl with the telescope*. There are two interpretations of this, either the boy saw a girl who had a telescope with her, or the boy used the telescope to see the girl. Sentences such as the the example given can lead to the parse tree being made incorrectly, and named entities going unrecognised.
- Foreign words
  - The variety in structure of different languages means that one method performing on text that includes text in multiple languages an often encounter problems. At the beginning of the development process, an English model was used as it was presumed that as the program was working on 19th century British mathematicians, this would be able to cope with the text. However, in the 19th century, collaboration and relationships across countries was not uncommon. This resulted in the models failing to recognise names that contained foreign keys or non English language names, and French names especially were not being recognised. The model has now been switched to a multi language model, but this is a common problem for processing text that can contain a combination of languages.

The problem of multiple references has also occurred during development. While not a problem with the actual recognition of the entities, ideally the program should identify when the text is talking about the same figure, as if we manually parsed a Wikipedia page for named entities, we would not note the same person multiple times if they were just referred to in a different way (i.e. by their surname). This is a problem of NLP more broadly, not specifically NER.

For the purpose of this project, NER is performed with the goal of identifying only the entities that are people. The next sections will compare commonly used libraries for NER in Python, and discuss some of the workings of these libraries. It will also detail the relative advantages and disadvantages of the methods, with a discussion of where they were performing interestingly on the data set presented.

7.2 Theoretical background

This section will cover the steps taken by each of the methods to identify the entities in given text and why they were chosen to develop with.

7.2.1 Spacy

Spacy is a natural language processing library, released by Matthew Honnibal and Ines Montani in 2015, with the aim of creating a library that was 'industrially orientated' and allowed companies to perform NLP tasks in a more efficient way than the libraries on the scene in 2015 allowed for (31). My project had two uses of the Spacy library, both using the pretrained named entity recognition model ('xx_core_web_sm') and the ability to train a new model. The following section will lay out the various steps in the NER models used, as well as the relative advantages and disadvantages of this library.

There are many reasons for choosing to use Spacy as one of our models to compare performance of:

- Well documented
  - Spacy provides exceptional documentation. This not only allows for support while developing code, but also it allows for an in depth exploration into the process it is following to identify the entities. Spacy provides both tutorials and a free online course to help developers who are new to natural language processing, as I was at the beginning of this project.
- Accuracy performance for people
  - Spacy has recently been shown to have a 0.6827 F1 score for PERSON entities, which was the second best performance in the study (compared to Stanford NLP - 0.8153, NTLK - 0.4528, Gate - 0.4105 and OpenNLP - 0.5041) (29). Although the ability to recognise organisations decreased it's overall F1 score, for the purpose of the study, it is more important to look at Person tags accuracy.
- Speed of parsing
  - In comparison with other greedy parsers (parser will accept the token with the most characters as they can before a rule will force them to create another token), Spacy had the fastest running time, see Figure 7.3 (32).

	Sent/Sec	Tokens/Sec	Language
ClearNLP$_g$	555	10,271	Java
GN13	95	1,757	Python
LTDP$_g$	232	4,287	Python
SNN	465	8,602	Java
spaCy	**755**	**13,963**	Cython
Yara$_g$	532	9,838	Java
ClearNLP	72	1,324	Java
LTDP	26	488	Python
Mate	30	550	Java
RBG	57	1,056	Java
Redshift	**188**	**3,470**	Cython
Turbo	19	349	C++
Yara	18	340	Java

Figure 7.3: Speed of parsing in comparison to other NLP libraries (32)

Below is a diagram illustrating the various phases that selected text must go to in order for a user to extract entities from the text.

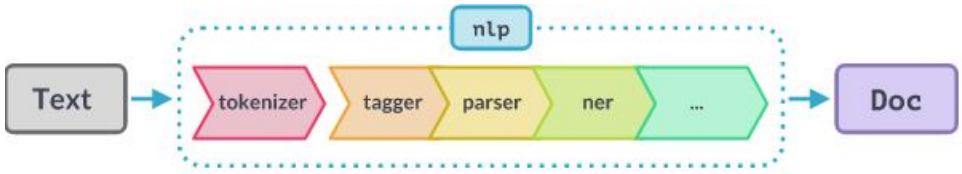

Figure 7.4: Pipeline from Spacy documentation (33)

When we use Spacy to perform NER, we create a nlp object. Within this nlp object, specific to the model you want to use, there are four main parts of the pipeline, with four main components that will be explored; tokenizer, tagger, parser, ner. All but the tokeniser are components of the processing pipeline.

The role of the Tokeniser is to take a string of text and turn it into a Doc, that will then be further processed by everything in the processing pipeline. The broad idea of the tokeniser part of processing is to break the text into segments, called tokens. These tokens can then be classified (34). There are many ways a set of text can be split into tokens, and Spacy uses a rule based model that is specific to different languages.

The tokeniser parses from left to right, and Spacy's parser has a particular emphasis on checking for suffixes or prefixes. A suffix is an affix which is placed after the stem of the word (-ed, -able), where a prefix is an affix that is placed before the stem of the word (un-, pre-). It looks for rules, for example there is one relating to the punctuation at the end of a sentence. A full stop at the end of a sentence needs to be split from previous tokens, where as acronyms with a full stop such as J.M.W. Turner should stay as one token (34). The rules are highly dependent on the language that the text is written in, and if this project was extended to Wikipedias of other languages, this part of the model would have to be modified. An example of what the tokeniser does for the text is illustrated in the following figure.

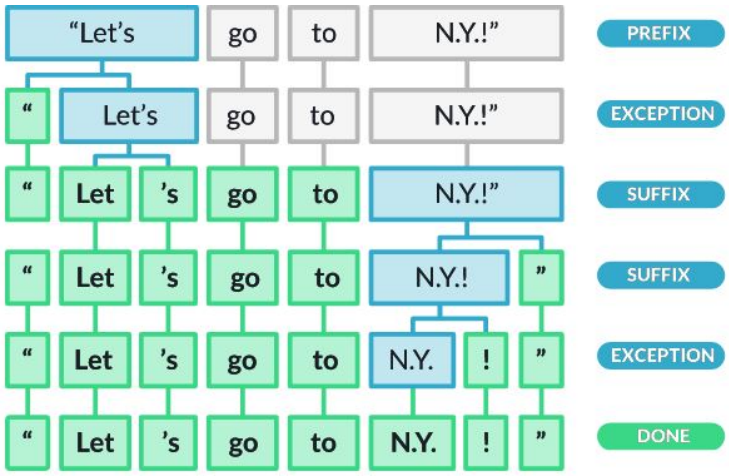

Figure 7.5: Example tokenisation progress from (**?**)

After the tokeniser has transformed the text into a Doc, the tagger is able to work it's way through the text and make predictions for which label or tag applies to each of the tokens. The tagger is not applying tags for what kind of word it may be (organisation, person etc), but instead is looking at features of the token. Example features that the tokeniser is looking to identify include retrieving the lemma (stripped form of the token without any capitalization and suffixes) of a token and whether the token is in the stop list. A

token present in a stop list would be a token that is commonly found in sentences but doesn't contribute to the meaning of the text (a, an, the etc). For a full list of features identified about each token at this point, please refer to the documentation (35).

The next step in the pipeline is to use the dependency parser. This part of the pipeline places emphasis on how the syntactic structure is representing the relationships between the tokens.

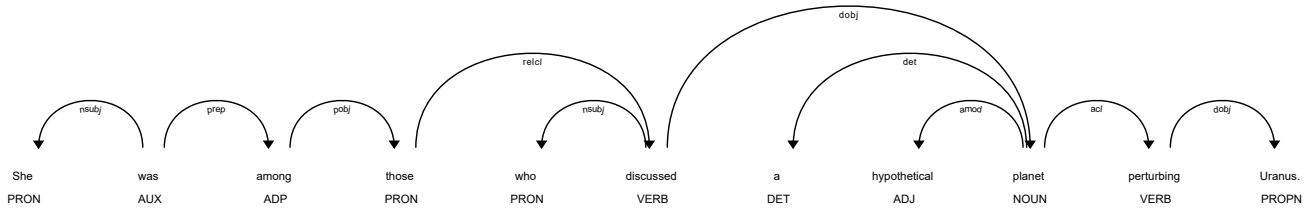

Figure 7.6: Sentence from Somerville's Wikipedia after parsing occurred

As you can see in this diagram of a sentence from Somerville's page that has been passed through the dependency parser, directed labels go between the heads and the dependents (36). For example, between Uranus and discussed there is a dobj label. This stands for direct object, representing that the noun (Uranus) is the object of the verb (to discuss). There are 42 relations that could occur between tokens; to read more of the different types, please see (37).

The final stage of the pipeline is the entity recognition stage. This is the stage where the model will predict the types of the entities in the text, and this project utilises this stage's ability to predict the PERSON entities in text. A variety of types of entities can be picked up from nationalities/political or religious groups (NORP) to organisations (ORG). As seen in Figure 7.7, when applied to selected text from Somerville's Wikipedia page, it has picked up the highlighted entities as people.

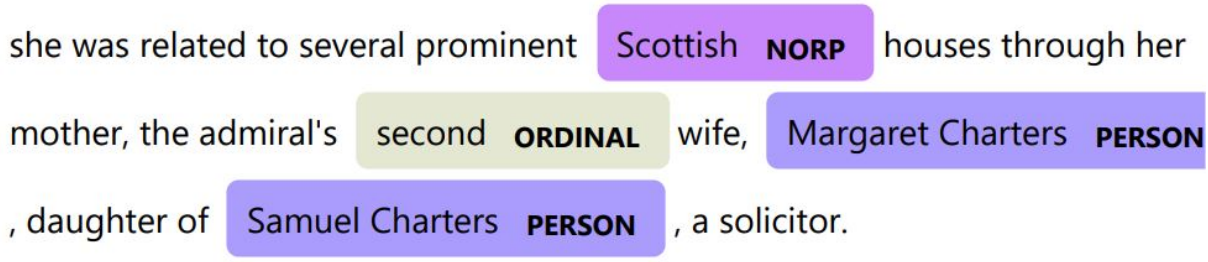

Figure 7.7: Sentence from Somerville's Wikipedia after NER has occurred

7.2.2 Retrained Spacy

During development, it was decided to try and retrain a Spacy model on domain data, to see if it would improve the performance of the NER. Domain data is data of the same form that we will eventually be running our NER methods on - Wikipedia articles in this case. Retraining Spacy on Wikipedia articles means that the model should be able to pick up features common across articles. It is important to choose models that are trained on data similar to the set being worked on. For example, NER trained on Wikipedia would not perform well on Twitter data, due to the lack of first person narrative in articles compared to tweets (38). It was decided to spend some development time on retraining a Spacy model, though as discussed later, the problem of lack of training and testing data was encountered.

A standard Spacy retraining was performed, and the steps in this process are fairly abstracted away to allow people without much machine learning experience to still access these tools. There are three main parts of the training process that will be explained; the creation of test data, the training of the model and evaluation of the performance.

The first task of training a new Spacy model is to get the available training data (Wikipedia articles of 19th century British mathematicians) into the correct form to allow the model to learn. For Spacy, the training data needs to be in the following form:

TRAIN_DATA = [(text, "entities":[(start,end,label)])]

Text will be a given sentence within the test data. The model needs to know where pre-identified entities are within each sentence, hence the (start,end,label) tuple. The standard Spacy multi-language model is initially used to identify any entities in the given text. This generates the training data in the desired form. Once the training data has been made, the model is ready to be trained.

The code is provided within the submission, so instead of stepping through all the parts of the code, a brief explanation of key ideas in the training will be provided. 30 iterations of training were performed (as is recommended as standard in the Spacy documentation) and before each iteration of training, the training set was shuffled to ensure that the model is not making predictions based on the order of the data that is given to it.

For each set of text and annotations, an Example class is created which will store both the gold standard reference data and the predictions of the pipeline in the currently developing model. The gold standard reference data is what the model should be picking up, identified from the training data generated by the Spacy multi language model. The currently developing model takes this Example data, assesses how well it is performing in comparison with the training data, and uses it to update its prediction for the next iteration.

A drop out value of 0.2 is given. This drop out rate means that 20% of the patterns learned at each level are dropped from the model. Having a drop out values allows overfitting to be prevented, which is where the model is able to perform very well on the training data as it learns the patterns too closely, and subsequently cannot perform well on unseen data.

Once this training of the model has been performed, a new Spacy model is generated. Whenever the code is re-run on a new computer, this model will have to be retrained as it is stored locally. The performance of the new model is evaluated in section 7.3.2. The model was applied to the articles in the test set that did not have mathematicians as their subjects. While this is slightly different domain data, it is assumed that the writing styles of Wikipedia articles is fairly consistently across the categories.

The problem of finding adequate training and testing data is familiar to many fields of NLP. In the case of this application, training on the mathematicians articles meant that the new Spacy model could not be reapplied to those articles, and so after testing the performance on the non mathematicians from the test set, there was no further development on retraining Spacy. Future developments in retraining models could train the model on articles of other 19th century groups and then apply this model to the articles of mathematicians and philosophers.

7.2.3 Natural Language Tool Kit

Natural Language Tool Kit (NLTK) is another open source Python library for natural language processing, developed to teach computational linguistics at the University of Pennsylvania, as the instructors often found that teaching the course required multiple programming languages and structures to effectively teach the various aspects (39). It was designed with four main goals in mind; simplicity, consistency, extensibility and modularity (40). Although like Spacy, you can train NLTK on your own corpus, NLTK is only used

in the current project in the standard 'off the shelf' form. The following section will detail the process and mathematics behind NLTK and illustrate how it is performing on the project.

```
def nltk_names(text, title):
    people = []
    for sent in nltk.sent_tokenize(text):
        for chunk in nltk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(sent))):
            if hasattr(chunk, 'label'):
                if chunk.label() == "PERSON":
                    text = ' '.join(c[0] for c in chunk)
                    people.append(text)

    write_to_file("ntlk", title, list(set(people)))
```

Figure 7.8: NLTK code for the project

As you can see in Figure 7.8, the NLTK cleary achieves its aim for simplicity. In just 9 lines, the code can break down text given to it (the text from the relevant sections of the Wikipedia article) into sentences, and tokenise these sentences, before checking if the chunks identified in the tokenisation are thought to be referring to people. I will briefly explain the steps taken by NLTK to deal with the text and how they differ from previously explained methods. The sentence and the words are both tokenised, and then assigned part of speech tags (POS) before Noun Phrase (np) chunking is used to identify the types of entities present in the text (41).

NLTK is a very popular Python library for NLP. it was chosen for testing compared to other methods for a number of reasons, including ease of use and flexibility for future expansions of the project.

However, NLTK has has been the worse performing of methods used on the articles. It has some major weaknesses:

- No support for word vectors
  - As seen in the discussion of Spacy's transformers, word vectors and embeddings can be incredibly useful for increasing the performance of your NLP application. NLTK was developed before the introduction of this approach and has not adapted to integrate word vectors. As discussed in the Spacy section, word vectors allow for much greater accuracy and this could help NLTK perform better on the given corpus.
- Low accuracy
  - As will be expanded in Section 7.3, NLTK performed at a significantly worse level of accuracy compared to the other methods tested in development.

Now that there has been a discussion of the theoretical background of these methods, Section 7.3 will analyse the performance of these selected methods on the Wikipedia articles, as well as identifying weaknesses and common problems spotted through the evaluation of results for the test data articles.

7.3 Performance of methods

This section will detail the performance statistics of the methods mentioned above on the Wikipedia articles of the test group of figures. As discussed above; F1, precision and recall are used to measure performance. A compiled table of the performance statistics for all methods can be found in Appendix C.

Due to problems with the size of some of the test figures names, the x axis values have been shortened to the initials of the individuals in the test set. The abbreviations are as follows; JM - James McCosh, HM - Harriet Martineau, MF - Michael Faraday, SWH - Sir William Hamilton, CHH - Charles Howard Hinton,

JT - John Tyndall, MS - Mary Somerville and JH - John Herschel.

Figures 7.9 and 7.10 illustrate the average F1, precision and recall values of the various NER methods across the test data set. These performances had to be split because of the retrained Spacy model. In order to evaluate the retrained Spacy model, only a selection of the test set could be used, as to test retrained Spacy (here after referred to as spacy_new) cannot have its performance evaluated on its training data. Figure 7.9 looks at the performance of Spacy and nltk over all 8 test figures, whereas Figure 7.10 looks at the performance of all three methods over 6 test figures (with Mary Somerville and Charles Howard Hinton excluded).

Figure 7.9 clearly indicates that Spacy is the best performing of the methods, and begins to highlight problems faced with the NLTK method, as the precision is too low to apply the method to new data.

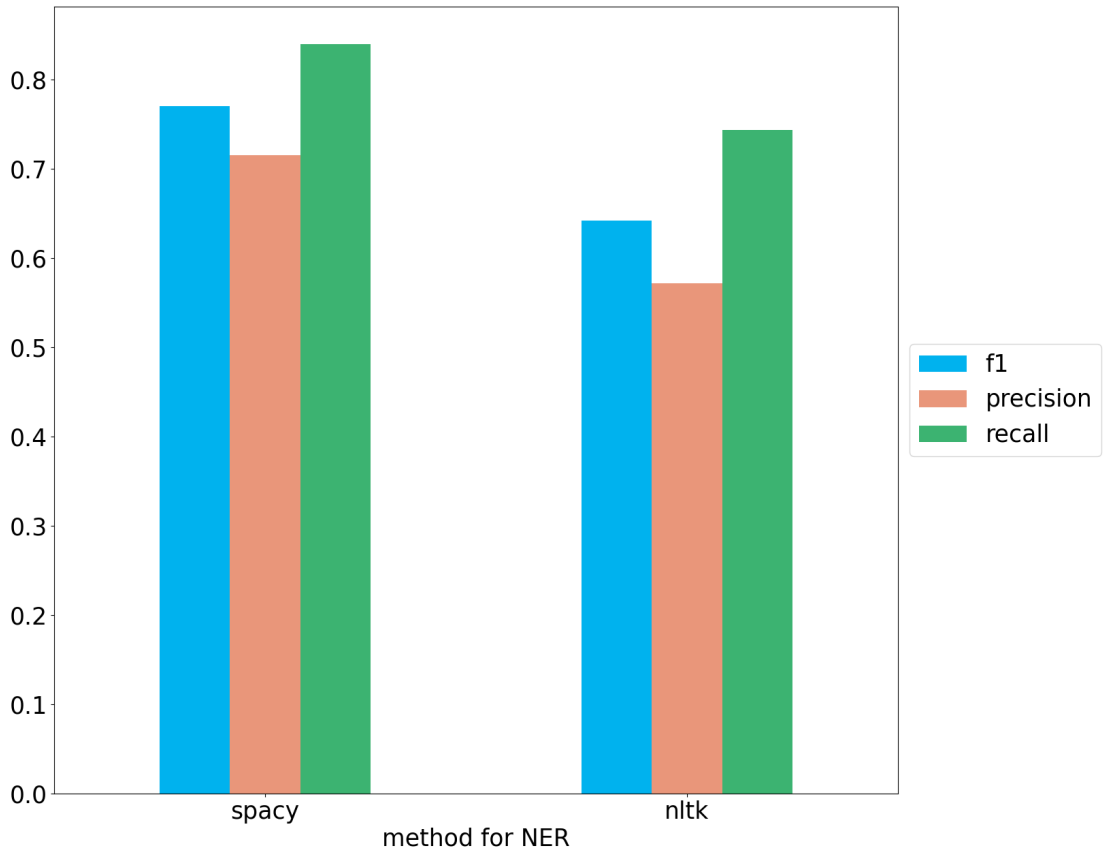

Figure 7.9: Comparison of NER methods through F1, precision and recall values across the entire test set of figures

Figure 7.10 illustrates the performances of Spacy and NLTK are slightly lower when not evaluating with the two maths figures, and that the retrained Spacy is performing well, but is less precise than the original model.

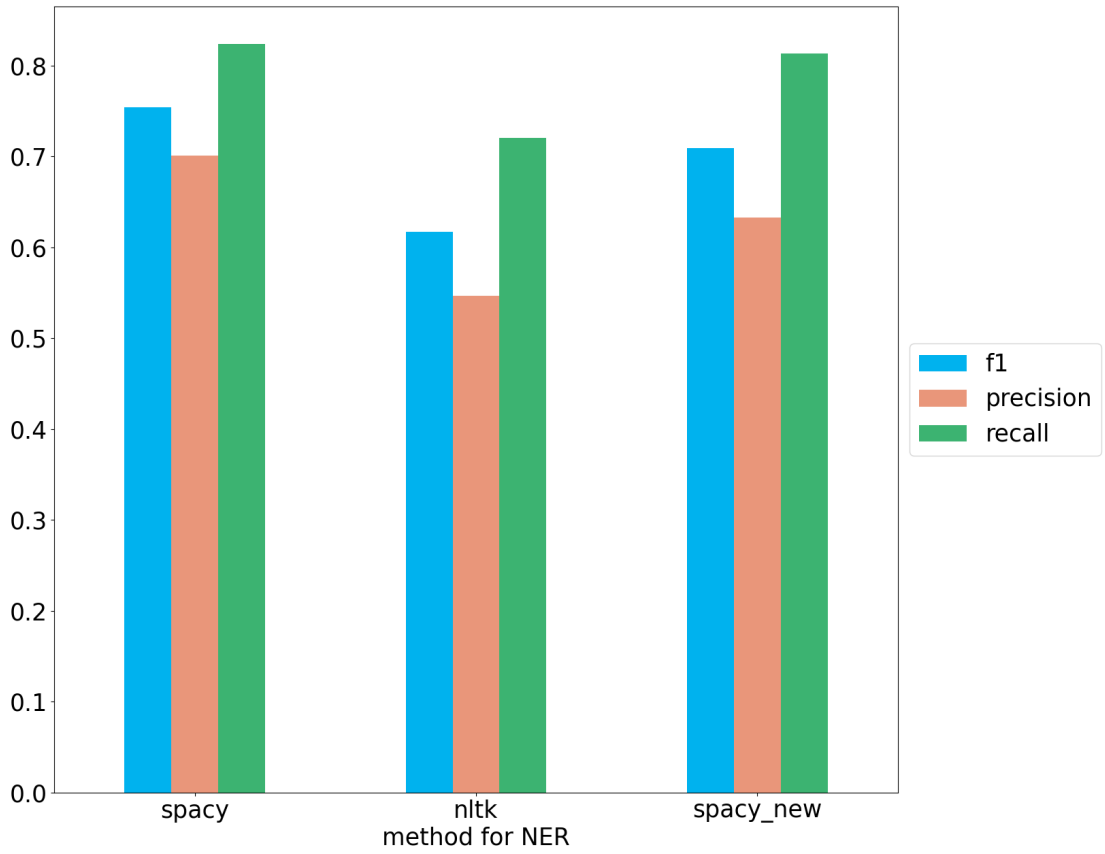

Figure 7.10: Comparison of NER methods through F1, precision and recall values across a subset of the test figures that did not include mathematicians

7.3.1 Spacy

Spacy on the whole is performing quite well at picking up named entities, and certainly performing the best of all three methods tried. Figure 7.11 illustrates all three performance statistics across the full test set of figures. Spacy's strength at recall is highlighted. With the exception of Spacy's performance on John Tyndall's article, precision values are also fairly high.

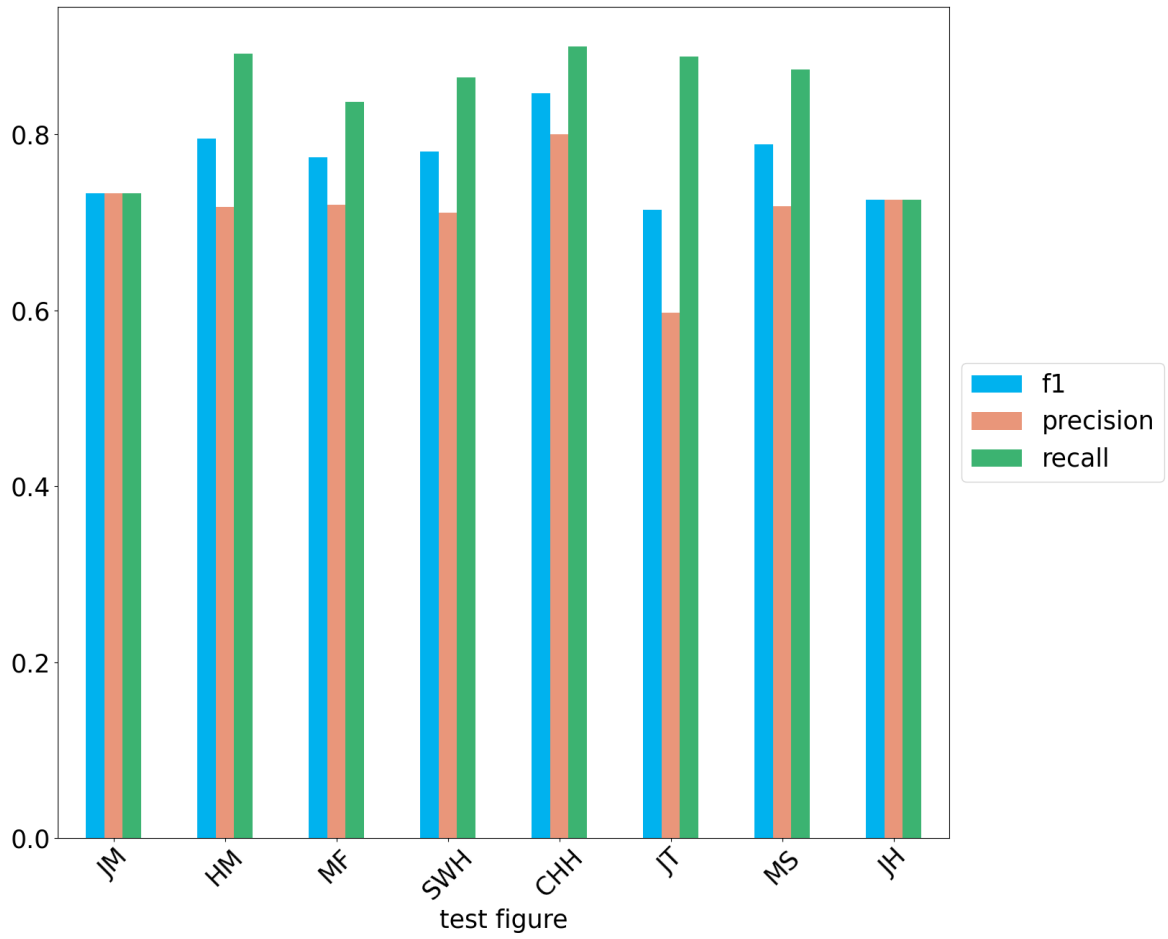

Figure 7.11: F1, Precision and Recall values of Spacy across the entire test set of figures

There do not seem to be many patterns between the additional names picked up by Spacy or the names that were failed to be recognised, but the following problems were raised when assessing the returns from Spacy:

- Picking up the name...and more
  - 'Madeleine L'Engle's A Wrinkle in Time', 'Jane Carlyle. Life in the Sickroom' and 'Henry W Field of London' provide examples of Spacy picking up the identified names (Madeleine L'Engle, Jane Caryle and Henry W Field) but also returning the name with additional surrounding text. The similarity between the two strings is not high enough to recognise the two strings referring to the same person.
- A large amount of the additionally returned names for Spacy tend to be single word entities

7.3.2 Retrained Spacy

While the retrained Spacy is performing better than NLTK, it does not surpass the multi language model originally used. I believe that this is due to the fact the model is built on an empty English model instead of a multi-language model (as far as I am aware, there is not an option to train from an empty multi-language model). This hypothesis is supported by looking at the specific performances of spacy_new on the test data set. Out of the six names not identified by spacy_new for John Tyndall's mentions, four of them are non-English names, and spacy_new's inability to cope with non-English names is demonstrated across the

test group.

Similar to Spacy it also tended to retrieve a lot of additional single word entities (such as molecular, fields, alps etc). The training data is provided by the Spacy multi-language model, so it is expected that the mistakes made in that model will also appear in this model.

When a name was mentioned next to another entity (such as Michael Faraday Memorial, Charles Lyell's Principles of Geology), spacy_new struggled to separate the name from the other parts of the text. The identification filter means that even though technically a name has been found in all the instances given, the Levenhstein ratio is too far away to recognise these as referring to the same person.

Figure 7.12 illustrates the performance statistics of spacy_new across the figures that it could be tested on. As seen again in the NLTK evaluation, the code is performing the worse on the articles of Michael Faraday and John Tyndall, with the precision values being lower than acceptable on these articles. There is not particularly any stand out negatives of the retrained Spacy, it just is not performing to the standards of the original multi-language model. Combined with the fact it can't be run across the mathematics network as that formed the training data, it was decided not to move on with any more development of this model.

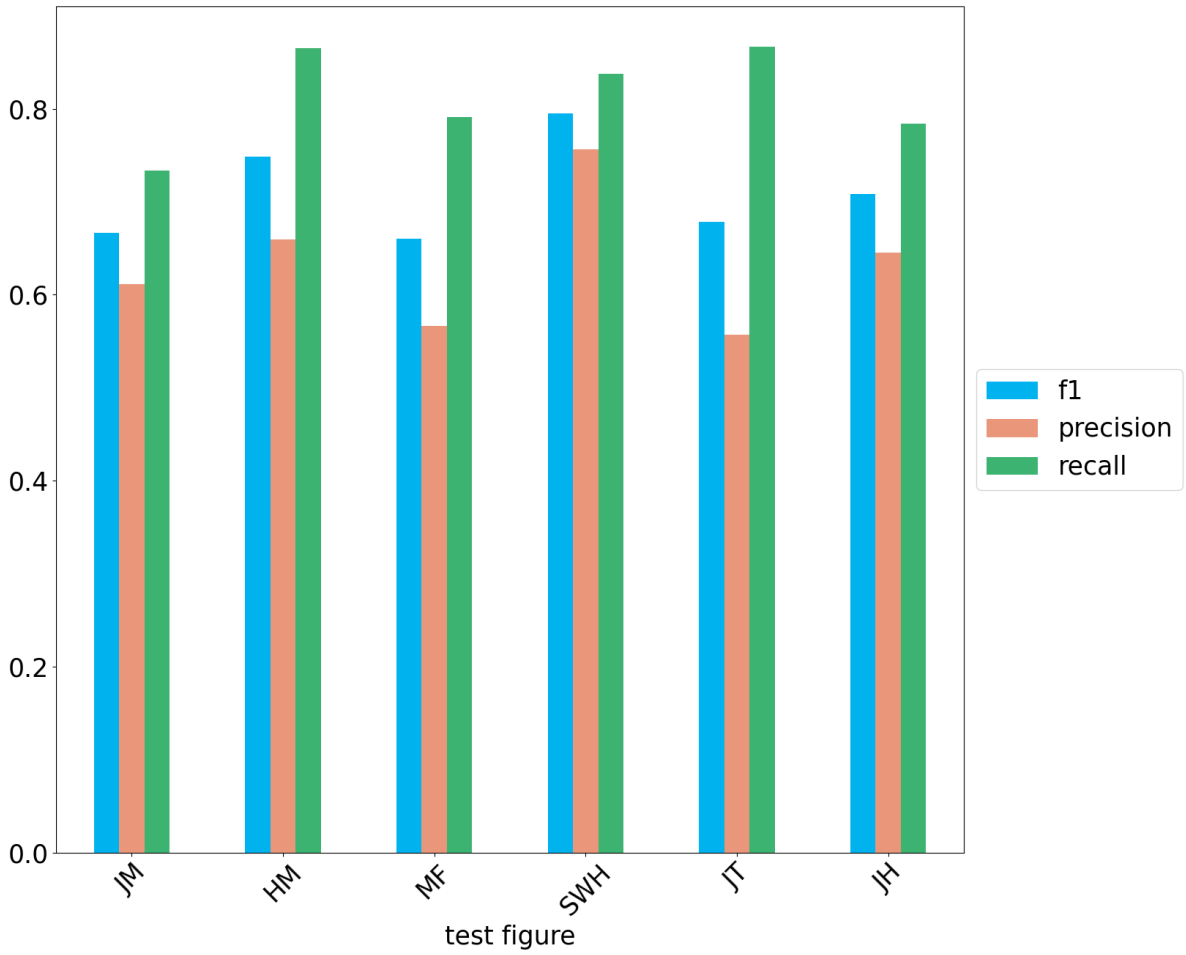

Figure 7.12: F1, Precision and Recall values of the retrained Spacy across the entire test set of figures

7.3.3 NLTK

The performance of NLTK was generally quite disappointing. Compared to the other methods, it was not very precise, with Figure 7.13 illustrating the precision levels of the library over the test data compared with Spacy. Spacy remained pretty consistent across the articles, but the precision value for NLTK fluctuated a lot more, and dropped to under 0.4, which is too unreliable to be used for a greater set of people without known test data to compare against.

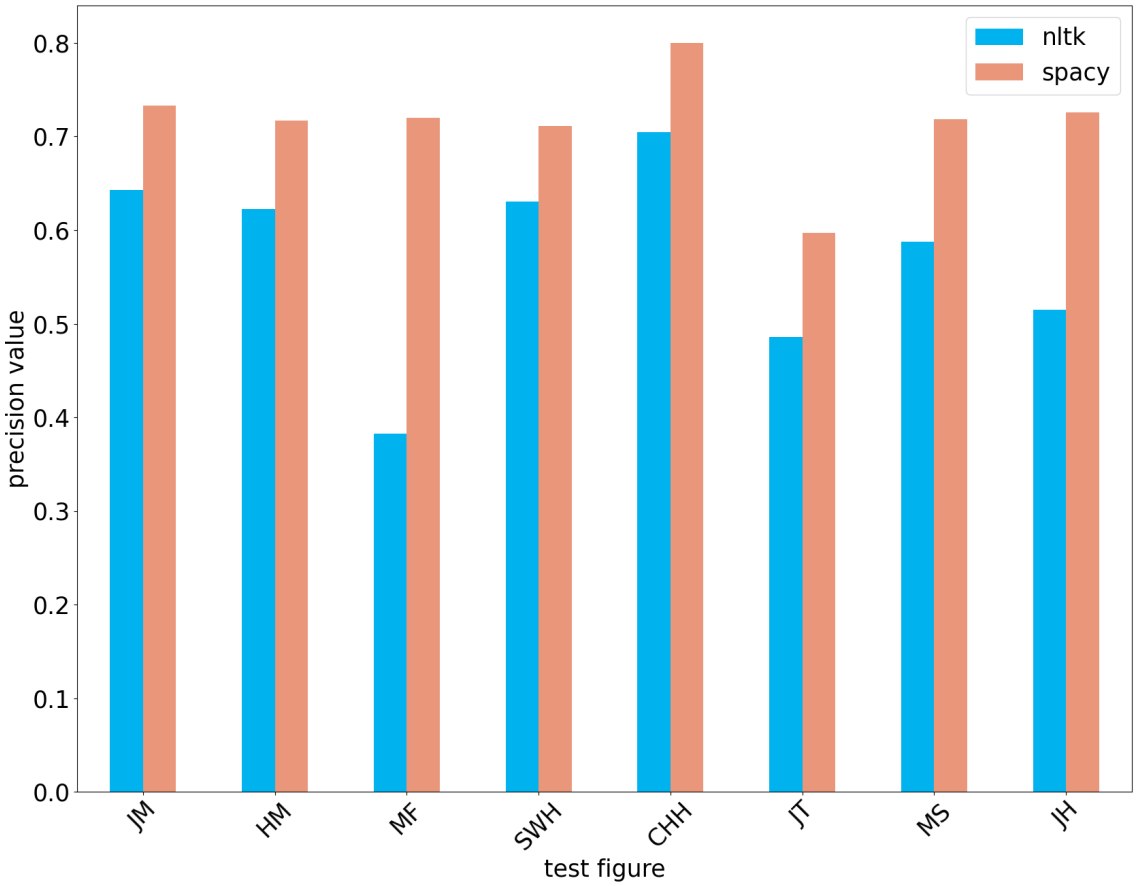

Figure 7.13: Precision values for Spacy and NLTK across the test set of figures

Figure 7.14 illustrates the recall performance of NLTK. Again, in all cases but one, the recall value for NLTK is below Spacy's. There is a greater difference between the values of recall in NLTK compared with Spacy, with the fluctuations again showing the misperformance and unreliability of NLTK.

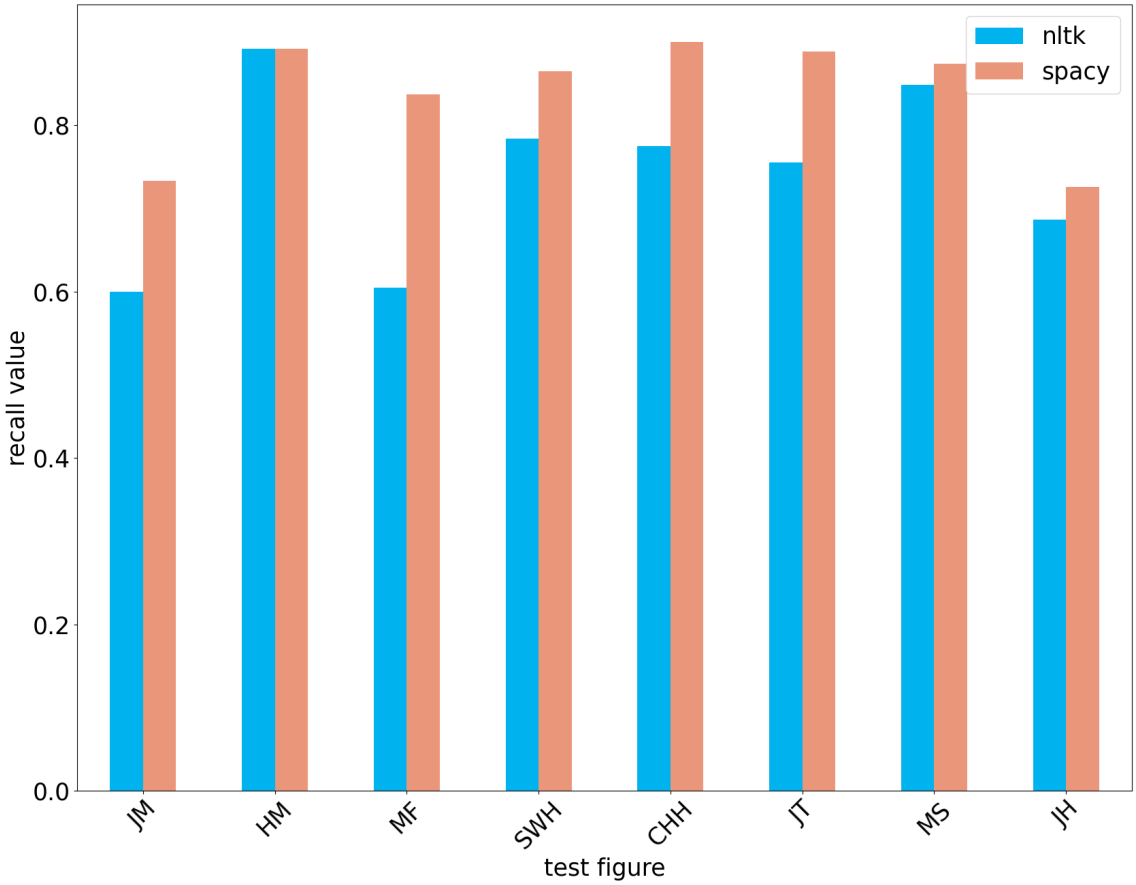

Figure 7.14: Recall values for Spacy and NLTK across the test set of figures

The code was able to output both names that were manually identified and not recognised by NLTK as well as additional names retrieved by NLTK from each Wikipedia file. This allowed common patterns between the errors to be analysed, and the following are particular problems that NLTK ran into:

- Struggling with names with punctuation in them
  - For many of the articles, NLTK failed to pick up names with punctuation in them, like H. P. Lovecraft. Being able to parse this is a more challenging task, as the full stop is not indicating a sentence end, but NLTK was unable to handle this.
- Struggling with three part names
  - Similar to the above point, if names consisted of more than two parts, NLTK sometimes failed to pick them up. This includes names such as Francis Landey Patton.
- Struggling with foreign language names
  - The parts of NLTK that are used in the code for this project are trained to deal with English data specifically, therefore it is no surprise that NLTK struggles with foreign names (especially French names) throughout the file. Unlike Spacy where there was an easy option to change to the multi language model, NLTK does not provide this option, although it does offer support for multiple languages in a select few features.

- Picking up irrelevant items
  - The precision of NLTK was worse than any of the other methods, and it often picked up many irrelevant entities, even things like chemistry compounds (including C2C16 for example in Michael Faraday's page).

7.4 Conclusion

In conclusion, the performance of all the NER methods on the Wikipedia articles is lower than I desired, but Spacy consistently outperforms the other two methods across the test set and so was the chosen method to move forward with in development. While retraining a Spacy model was done with the intention of surpassing the performance of the standard model, the lack of training data and the inability to create a multi-language empty model to train on meant that there was no benefit brought in any further retraining.

The code still provides options to extract names from the test data sets using NLTK and the retrained Spacy if further evaluation is desired.

8 Results Analysis

The following section will detail the results that can be generated using the lists of 19th century mathematicians and philosophers from the category articles on Wikipedia. It is worth briefly discussing the sample sizes. While it would have been ideal to have a thorough comparison by gender for a lot of these statistics, the comparative sample sizes of the male and female groups for both disciplines means that I will be unable to draw any strong conclusions about patterns emerging.

Figure 8.1 states the number of articles in each category that were used for analysis. There are 81 total pages on the 19th century British mathematicians category page, so 84% of the articles were involved in analysis. There are 55 pages on the 19th century British philosophers category page, so 91% of articles were involved in analysis. The remaining articles caused an error in extracting linked names.

	Male	Female
Mathematicians	62	6
Philosophers	42	8

Figure 8.1: Gender breakdown of the number of articles involved in the analysis from each category

The Wikipedia category articles do not claim to be complete, and at a glance, you can easily see both female and male contributors to either discipline missing from the list. While it would be unrealistic to suggest these sample sizes would ever balance out completely, it is suspected that there are many more articles about female mathematicians and philosophers that have not been included in this list that could be added to a future analysis to help further explore the relationship between gender and linked articles. As it was not a main aim of the project to do this breakdown however, this will be left for future work.

The results section will analyse three main statistics about the data collected; average number of mentions per 1000 characters, average number of characters in a Wikipedia article and number of mentions for an average length article. As will be discussed further in a later section, the lengths of the Wikipedia articles of mathematicians and philosophers greatly varied, so by comparing a normalised version (per 1000 characters), as well as comparing across the average article length, it is hoped the analysis can provide insight from a variety of perspectives on the data. The square root or log of these values are taken at various points during the analysis, so that the assumptions of Analysis of Variance (ANOVA) tests could be satisfied and comparisons drawn. The assumptions of ANOVA are that the residuals are normally distributed and the variance of the groups are approximately equal. Additionally observations should be independent of each other, which all of the articles are.

All code used to produce the following graphs can be found in results_analysis.R, and while there will be discrepancies between earlier figure styles, the ease of analysis in R meant that I was more comfortable performing the statistics using that language as opposed to trying to perform the same thing in Python.

Violin plots have been produced to help assess the distribution of the statistics and variation between groups. The mean in the plots is indicated with the diamond, the rectangular bar indicates the interquartile range and black dots indicate outliers. While in a lot of these graphs the median is obscured by the diamond, the median is represented using a horizontal line. Like probability distribution functions, the wider the graph is at any given point, the higher the probability that a given value will fall in this area.

8.1 Division by category

The following section will discuss the differences in statistics produced across the categories of articles (mathematicians and philosophers).

8.1.1 Average number of mentions per 1000 characters across categories

Figure 8.2 compares the number of mentions per 1000 characters across the two methods of named entity extraction and the two groups under consideration. 1000 was chosen as the factor to ensure that the numbers were a recognisable integer of mentions. The square root has been taken to ensure that the assumptions for ANOVA hold.

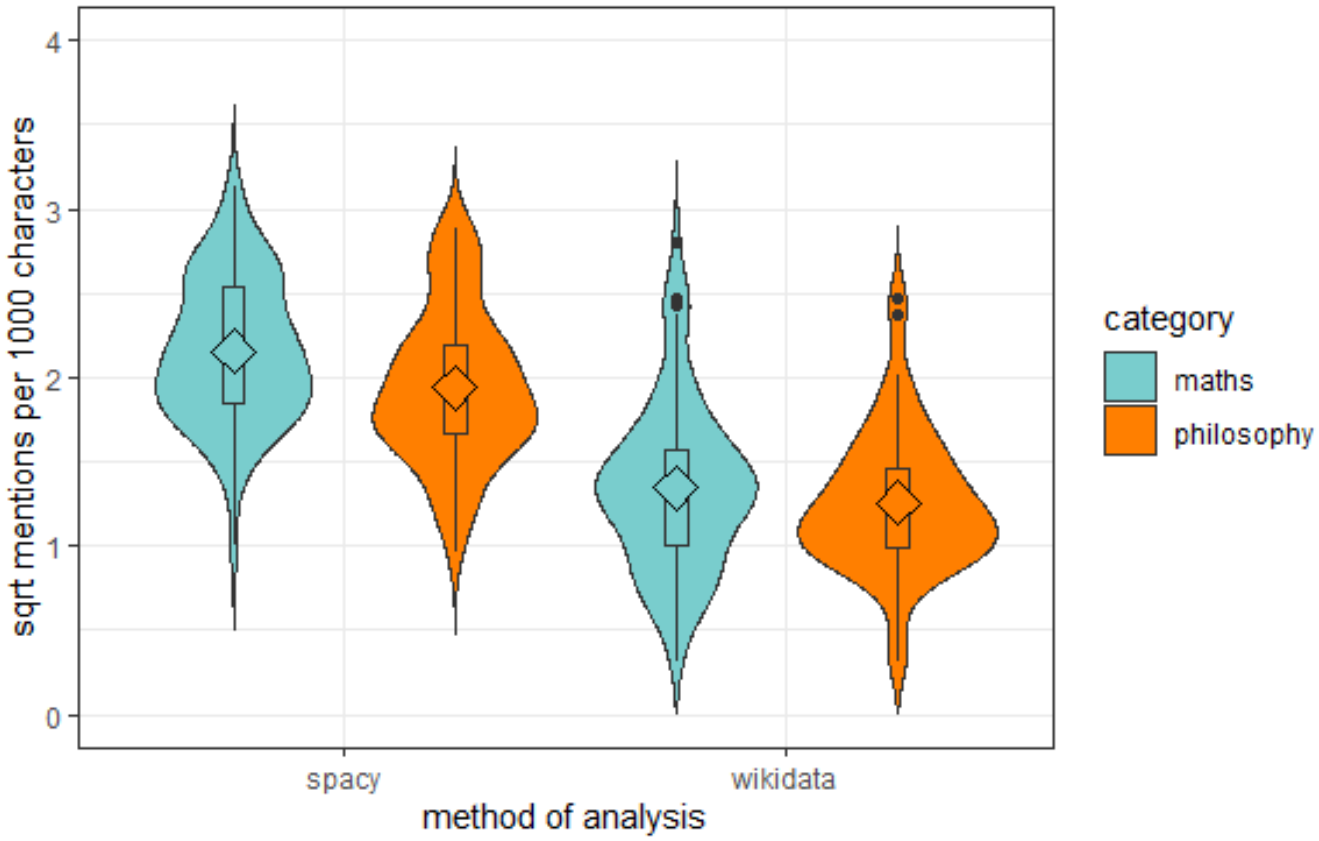

Figure 8.2: Square root of mentions per 1000 characters for Wikipedia articles on mathematicians and philosophers

There was a significant effect of category on the square root of mentions per 1000 characters (two-way ANOVA, $F = 6.28$, $df = 1, 232$, $p < 0.02$), with articles about mathematicians having a higher mean than for philosophers. There was also a significant effect of method (two-way ANOVA, $F = 106.88$, $df = 1, 232$, $p < 0.001$) with Spacy having a consistently higher mean than Wikidata. There was no significant interaction between method and category on the square root of mentions per 1000 characters.

This indicates, that unlike what was suspected, Wikipedia articles for 19th century British mathematicians are mentioning a greater number of people in a section of text of a given length compared with philosophers, regardless of the type of mention (linked or unlinked).

8.1.2 Length of articles

While given the same number of characters, articles about mathematicians tend to contain more mentions, assuming that these articles are of the same length is not valid. Figure 8.3 illustrates the differences in number of characters for Wikipedia articles of mathematicians and philosophers.

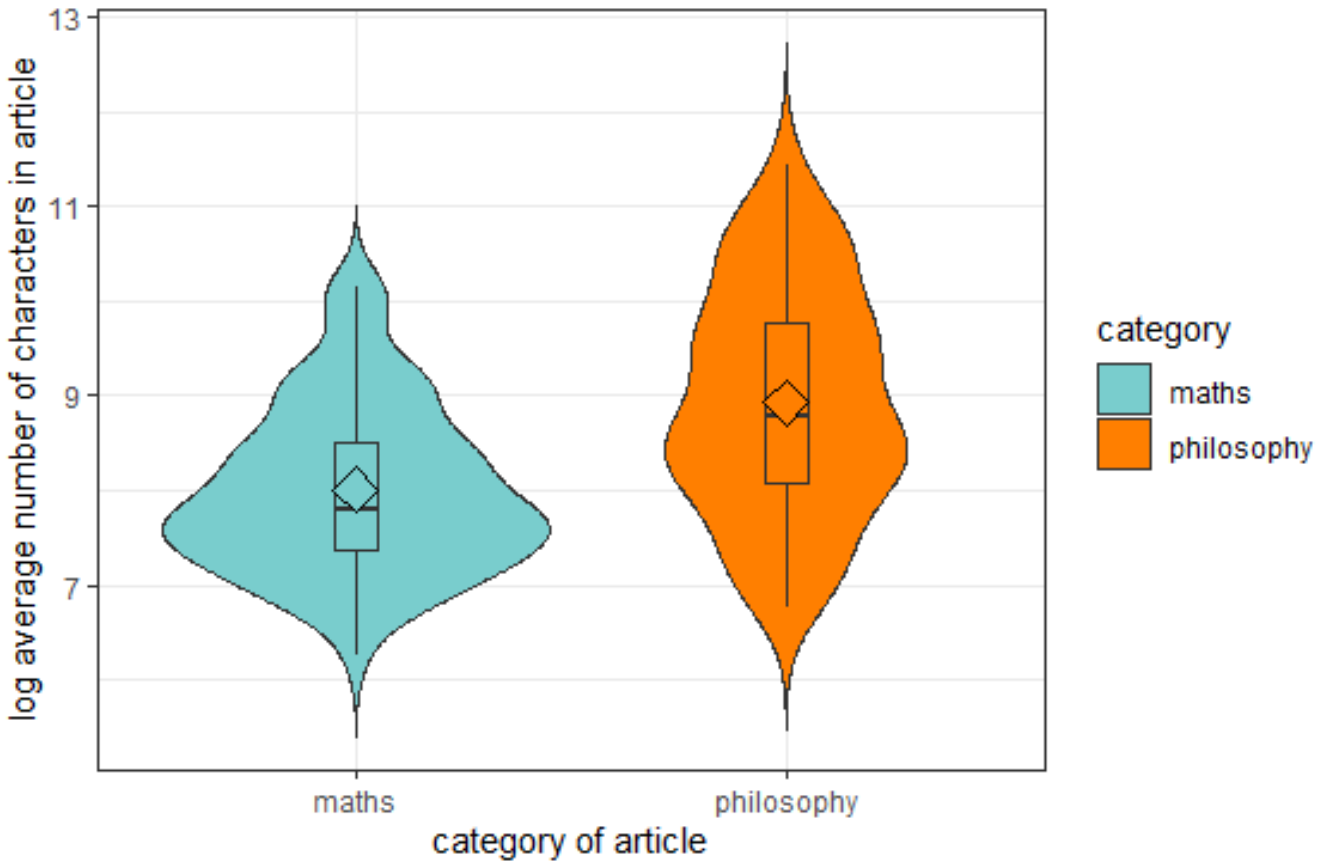

Figure 8.3: Log number of characters per article for Wikipedia pages of 19th century British Mathematicians and Philosophers

There are significantly more philosophy articles that are outliers when it comes to article size. Figures such as John Ruskin (longest length of article in data set) mean that the distribution of article lengths of philosophers is more spread out than article lengths of mathematicians.

There was a highly significant effect of category on the mean number of characters in an article (one-way ANOVA, $F = 45.47$, $df = 1, 234$, $p < 0.001$), with articles about philosophers having a greater log average number of characters. This indicates that even though given the same amount of text, articles about mathematicians will contain more mentions, the articles are not typically developed to the same extent.

8.1.3 Mentions in a typical length article

To account for this difference in article length, the mean length of an article (in number of characters) was calculated for each group and this was combined with the information about number of mentions and article lengths for each article to produce the mentions in a typical length article for each group, as illustrated in Figure 8.4.

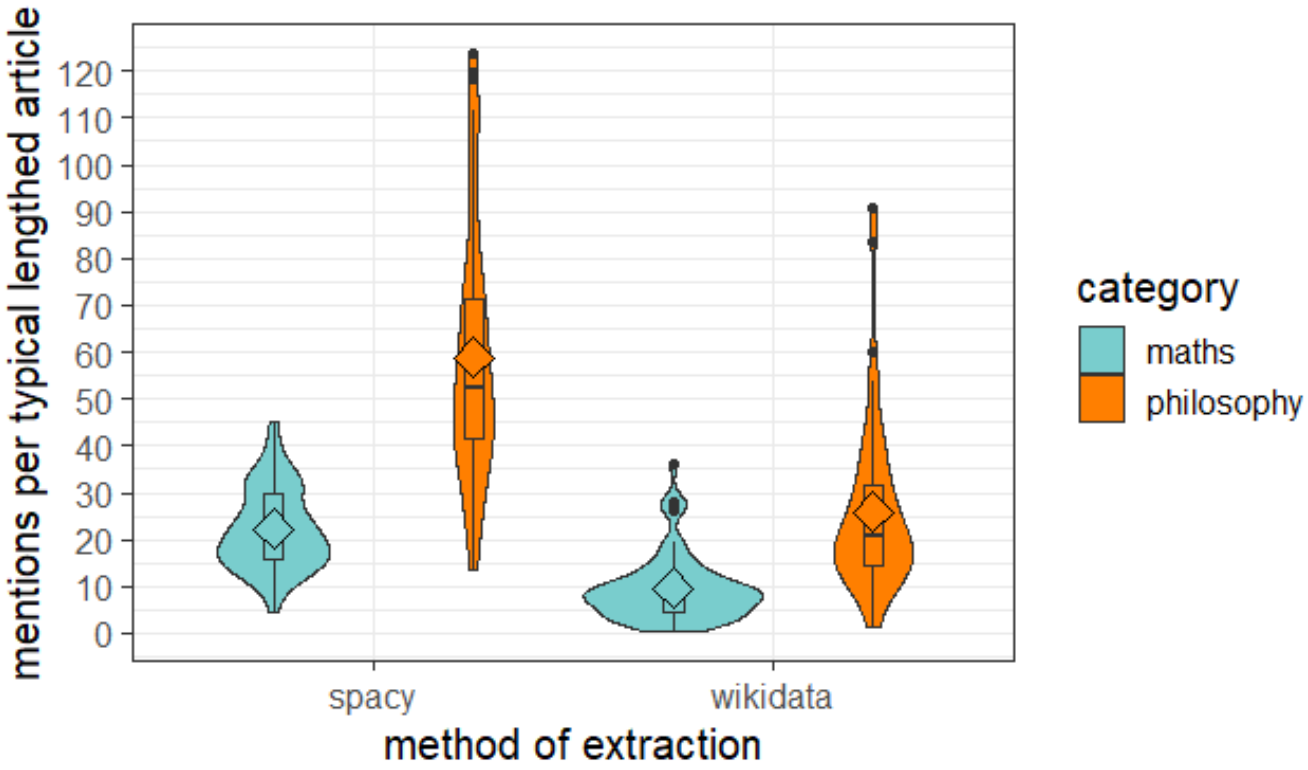

Figure 8.4: Number of mentions per typical article for 19th century British Mathematicians and Philosophers

Despite taking log and square root of the data, it was not possible to satisfy all assumptions of ANOVA. Due to this, a non parametric test had to be performed and the Wilcoxon rank sum test was used. This will allow for an insight into patterns of mentions between the categories, without the assumption required for ANOVA that the data is normally distributed and has equal variance. For a Wilcoxon test, the data had to be separated to test the hypothesis across Spacy mentions and Wikidata mentions. We already have shown above that Spacy consistently recognises more names than Wikidata so while we are taken out a layer of the analysis, it would not have presented any additional insight.

There was a significant effect of category on the typical mentions per article across both methods ($p < 0.01$ for both methods). As expected, as philosophers have longer length articles, the number of mentions in an article of typical length is higher for philosophers.

8.1.4 Conclusion

This analysis has highlighted that the expected difference in article mentions by category is completely down to the length of the articles. Although this goes against what was initially expected, it indicates that the networks of collaborations between mathematicians are being represented and based on these data, there is a higher number of mentions per 1000 characters for mathematicians. However, the difference in lengths of articles of mathematicians and philosophers indicate that the articles overall are less developed for mathematicians.

There may be many reasons why articles of philosophers tend to be longer. Academic philosophers, by training, are taught how to write prose and therefore may be more comfortable with contributing to writing Wikipedia articles. In comparison, mathematicians are often not equipped with the same writing practice during their training and may be less inclined to contribute to written representations of previous mathematicians. It is also suspected that philosophical ideologies are more often named after philosophers (think

Marxism and Platonism), leading to an increased contribution to articles of major figures as the articles about the movements are also developed.

While mathematicians have a greater number of mentions per 1000 characters than philosophers, there has been no indication how these values compare to other categories. It will be discussed further in the evaluation section, but the small amount of literature on collaboration statistics for papers of philosophers indicates a low level of collaboration. It would be incredibly interesting to compare these numbers with groups such as physicists and astronomers to reveal whether the mentions per 1000 for both mathematicians and philosophers are low compared to other groups.

8.2 Division by gender

The sample size of female mathematicians and philosophers in our network is incredibly small. It is worth discussing the gender bias that is present on Wikipedia. In the period of focus for this report, it was already going to be hard to equate the numbers in male and female groups due to the biases that existed in the society at the time, and still do, to a lesser extent. This will have been further amplified by the very prevalent bias in Wikipedia articles, impacting things from the number of articles, the content of articles and the connectivity of the articles. There is a higher ceiling of accomplishment that women need to achieve to ensure that their article is deemed significant enough not to get deleted ((42), (43)). If they do manage to have an article that is not deleted by Wikipedia editors (who are also overwhelmingly male (43)), there is often a difference in content of the articles. Articles about women are more likely to emphasis contributions unrelated to her field of work, such as family and personal connections or ways in which the women was a role model (44).

Despite the small sample sizes, there was still opportunity to perform some analysis on the differences caused to the statistics by the gender of the article subject.

8.2.1 Per 1000 characters

Figure 8.5 is a similar graph to Figure 8.1 and assesses how gender influences the square root of average number of mentions (of either Spacy or Wikidata) per 1000 characters.

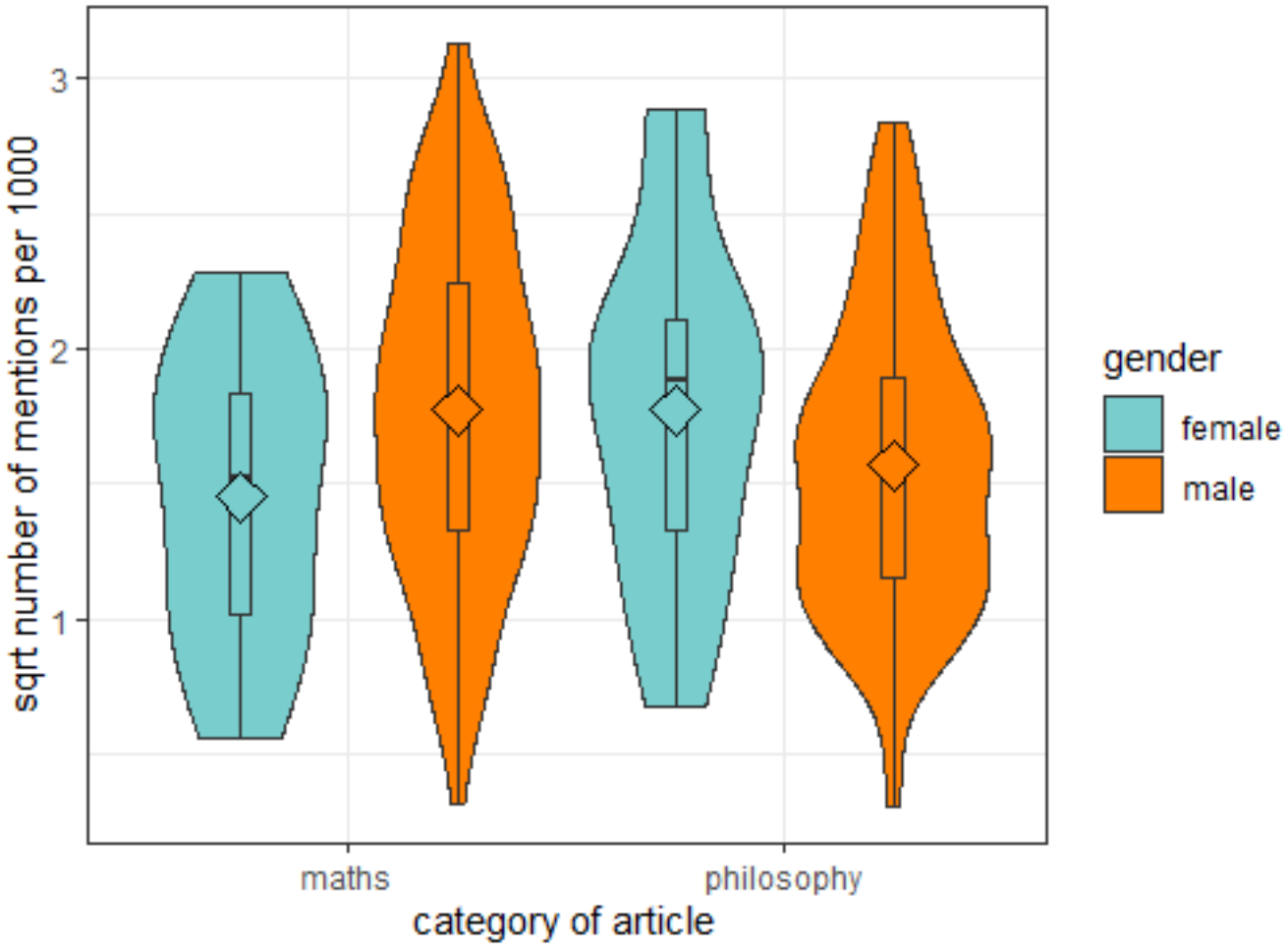

Figure 8.5: Square root of mentions per 1000 characters for Wikipedia articles on mathematicians and philosophers, divided by gender

A three-way ANOVA was performed on the square root of mentions per 1000 characters, to test the effects of method, category and gender.

There was a significant effect of method (three-way ANOVA, $F = 10.56$, $df = 1, 228$, $p = 0.0013$) with the means varying as has already been discussed with Spacy consistently recognising more names than Wikidata.

There was also a significant interaction between category and gender (three-way ANOVA, $F = 4.14$, $df = 1, 228$, $p < 0.05$). To assess how category with gender were influencing the response variable, a Tukey HSD test was undertaken to allow exploration into which combination of category and gender were causing a significant effect, by doing a pairwise comparison across the category and gender combinations. With a p value of 0.0087, only the difference between male philosophers and male mathematicians was significant. While this is not a particularly interesting comparison, the fact that the other combinations were not significant may indicate more. This suggests that within our set of Wikipedia articles, there was not a significant effect of gender on the number of mentions on pages. It is, however, worth remembering that the sample size does not provide much statistical power to make broad conclusions about the underlying differences between articles about male and female mathematicians/philosophers.

8.2.2 Length of article

Figure 8.6 begins the analysis of whether there is gender was affecting the article length measured in number of characters.

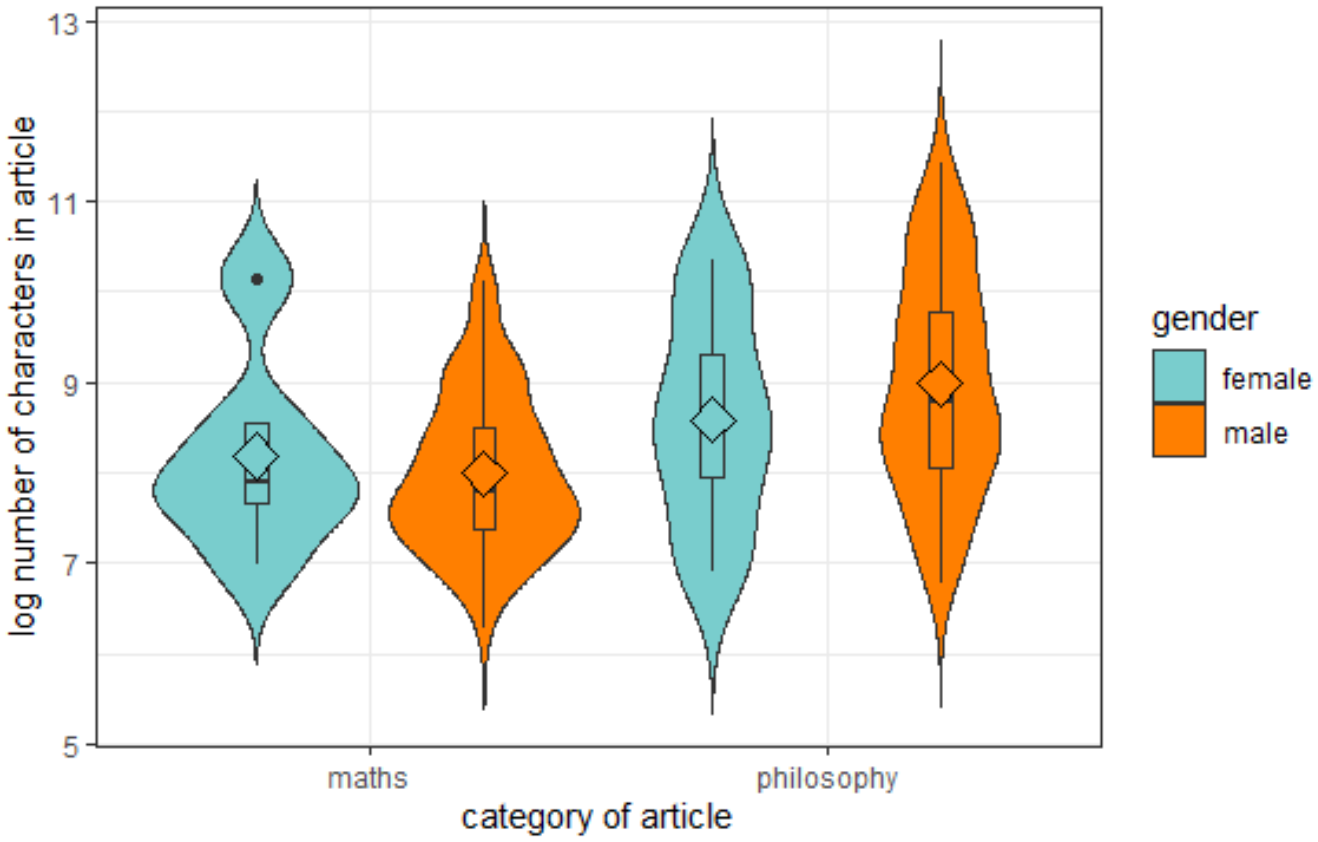

Figure 8.6: Log of number of characters per article for Wikipedia pages of 19th century British Mathematicians and Philosophers

Performing an ANOVA test revealed that none of the factors had a significant effect on the log number of characters per article of each category. This indicates, that despite the acknowledged bias in the creation and content of Wikipedia articles discussed at the start of this section, within our data set, the length of the articles are not effected by gender.

8.2.3 Conclusion

Looking at the breakdown of statistics by gender has revealed the opposite of what was expected when considering the bias that normally is found in Wikipedia articles. Not only is there no significant difference between the number of mentions in 1000 characters of the articles of members of both groups, but there is not a significant difference in the lengths of the articles. All of the results of this section must be considered with the size of the test group however, as there was considerably more articles about men in both of the category groups.

The lack of difference may be explained by the bias that is in place for which women are deemed impactful enough to have a Wikipedia article. The group selected that are female mathematicians have had to achieve more than their male counterparts to be recognised and therefore there is not a difference in length

because there is more to talk about for what the women has achieved with her career. The tails of the mathematicians articles indicate this may only be the case for mathematicians, with the distribution of the male log number of characters going lower than the females.

8.3 Commonly linked people

While not necessarily related to the representation of networks on Wikipedia for either subject group, an additional statistic that can be picked up through analysis can provide information on the most commonly linked and mentioned people.

8.3.1 Performance of methods

During this analysis, the imperfections of both methods were brought to light. As seen in Figure 8.7 and 8.9, the NER by Spacy is picking up additional entities that are not referring to people, such as Senior Wrangler and Natural Philosophy. While this is frustrating to see, it also allows us to spot the potential bias in the mathematicians who are on the Wikipedia category list. Even when choosing a sample group for testing, I had to be careful to ensure that not all the mathematicians were associated with Cambridge. Over the 19th century, Cambridge grew to be world leading in its training of mathematical physicists. Cambridge not only trained a large number of the brightest mathematical minds at the time, but produced extremely capable graduates, with the questions in the senior examinations being incredibly similar in difficulty to the research level mathematics at the time (45). The dominance is clearly illustrated with Senior Wrangler being the joint most mentioned entity recognised by Spacy.

Additionally, a link has been removed from the count after appearing as the second most linked to article from mathematicians. 'Recent changes in pages linked from this page [k]' was picked up as a link from Wikipedia pages 8 times from mathematicians. The flaws of the Wikidata link identification will be discussed in the following section. It does not add anything to our analysis at this point, aside from an alert that the code is not perfect, and has therefore been removed from the following graphs. This is the only such instance of removing a data point.

In the submitted folder, the file popular_figures.txt in the analysis folder contain the data used to produce the graphs, as well as a list of all the people who did link/mention each of the given names. It was believed that all of the raw data would be useful to provide further analysis on who was linking to certain commonly linked entities.

Names may have had to be shortened to initials due to problems with making the graph large enough to be readable. If this is the case, the expansion of the initials have been provided underneath the graph.

8.3.2 Mathematicians

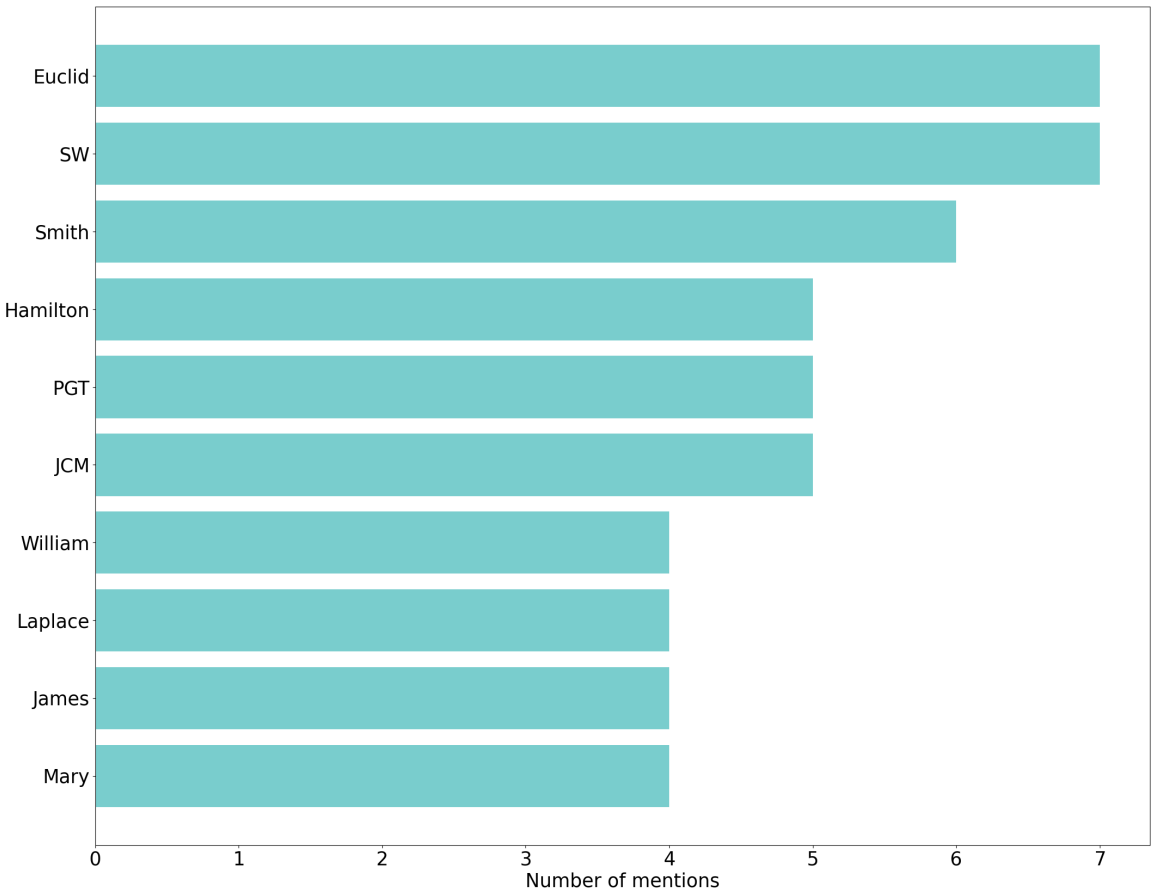

Figure 8.7: Most commonly mentioned entities for mathematicians

Where JCM is James Clark Maxwell, PGT is Peter Guthrie Tait, ADM is Augustus De Morgan, WRH is William Rowan Hamilton, HP is Henri Poincaré and ACB is Alexander Crum Brown.

Figure 8.7 illustrates the commonly picked up names in Wikipedia articles for mathematicians by Spacy. Without the additional context of surnames, single word first names cannot reveal much about the commonly linked nodes in the mathematics network. It is worth noting that Mary, the only female name to have made this top 10 list, is the first name of 3 of 6 female mathematicians in the network. Mary was linked from the articles of the following people; Mary Edwards, Robert Haldane, Arthur Stanley Ramsey and Mary Somerville.

Similarly names such as Smith, Hamilton, William and James are too ambiguous to state whose these may be referring to, and there is no guarantee it is a reference to the same person.

Euclid has been highlighted as a commonly mentioned name in the articles of 19th century mathematicians. This doesn't come as much of a surprise, considering the influence he had on modern mathematics, the

use of his work for mathematical teaching and the revival of interest in the fifth postulate from the Elements (46). While British mathematicians were not prominently involved in the early moves into non-Euclidean geometry, Euclid's *Elements* was the dominant textbook used to teach geometry to school children until the 1900s (47). It is likely that many of the mathematicians featured on the Wikipedia category were exposed to Euclid's work from a young age, explaining why he is such a prominent figure within the network of articles. Euclid, being a figure from the ancient period will be missing from the commonly linked articles with this filtering in place. While it will not be able to be performed as part of this project's analysis, the code could be slightly adapted so there wasn't a removal of linked articles that do not refer to figures of the 19th century. It would not be surprising if Euclid managed to place towards the top of this list when looking at commonly linked figures.

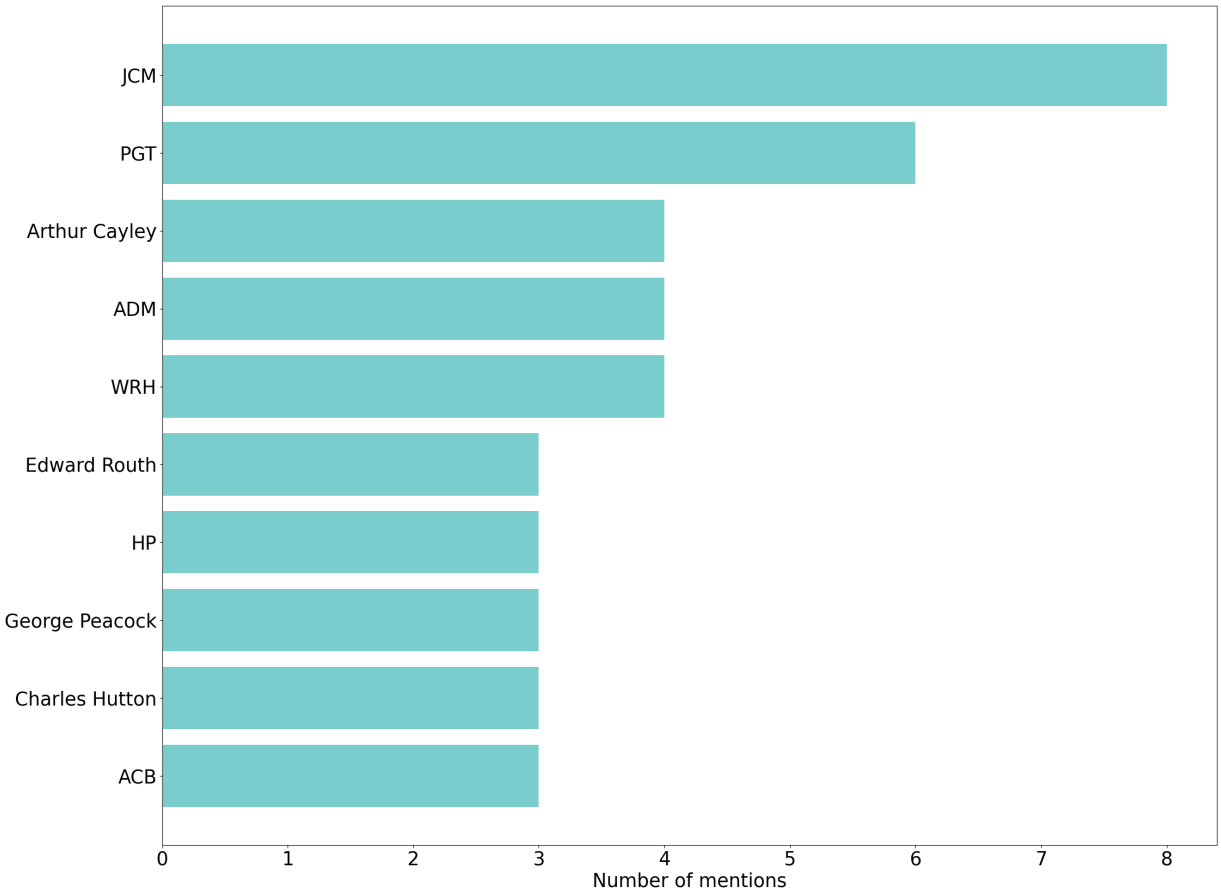

Figure 8.8: Most commonly linked articles for mathematicians

Where JCM is James Clark Maxwell and NP is Natural Philosophy.

Aside from the problem with the recent changes file, Wikidata analysis can provide more information due to the fact we know exactly who the article is referring to. In Figure 8.8 we see that Maxwell is the most linked to of the articles referring to 19th century figures. It is unsurprising that while there is no filter

to restrict the occupations of the subjects of these articles, mathematicians and natural philosophers clearly dominate the top 10 articles.

Unsurprisingly the majority of the figures listed were involved in teaching of some form. This is standard for the period, as this was the main occupation of anyone that was then classed as a mathematician. Additionally, it isn't a surprise that those associated with teaching or high up positions are then commonly mentioned throughout the articles of people in this category, due to their exposure to a large number of people and the potential influence of a mentor type relationship. Teaching roles held by these figures include the chair of natural philosophy at Edinburgh (Tait) and Sadleirian professor of mathematics at Cambridge (Cayley).

11 people are linked in 3 Wikipedia pages, and the mathematicians present on the graph are only chosen due to the ordering of the data structure. The people mentioned three times are Michael Faraday, Edward Routh, Charles Hutton, John Herschel, David Brewster, Nevil Maskelyne, Henri Poincaré, Alexander Crum Brown, Balfour Stewart, George Peacock and Philip Kelland. While it is not going to be possible to provide an analysis of common characteristics between all these figures, stand out people include Routh, whose role as as a coach for students preparing for the mathematical tripos at Cambridge would have meant there was a high number of interactions between Routh and prominent Cambridge mathematicians. Poincaré is the only non British mathematicians to make the list of most commonly linked.

8.3.3 Philosophers

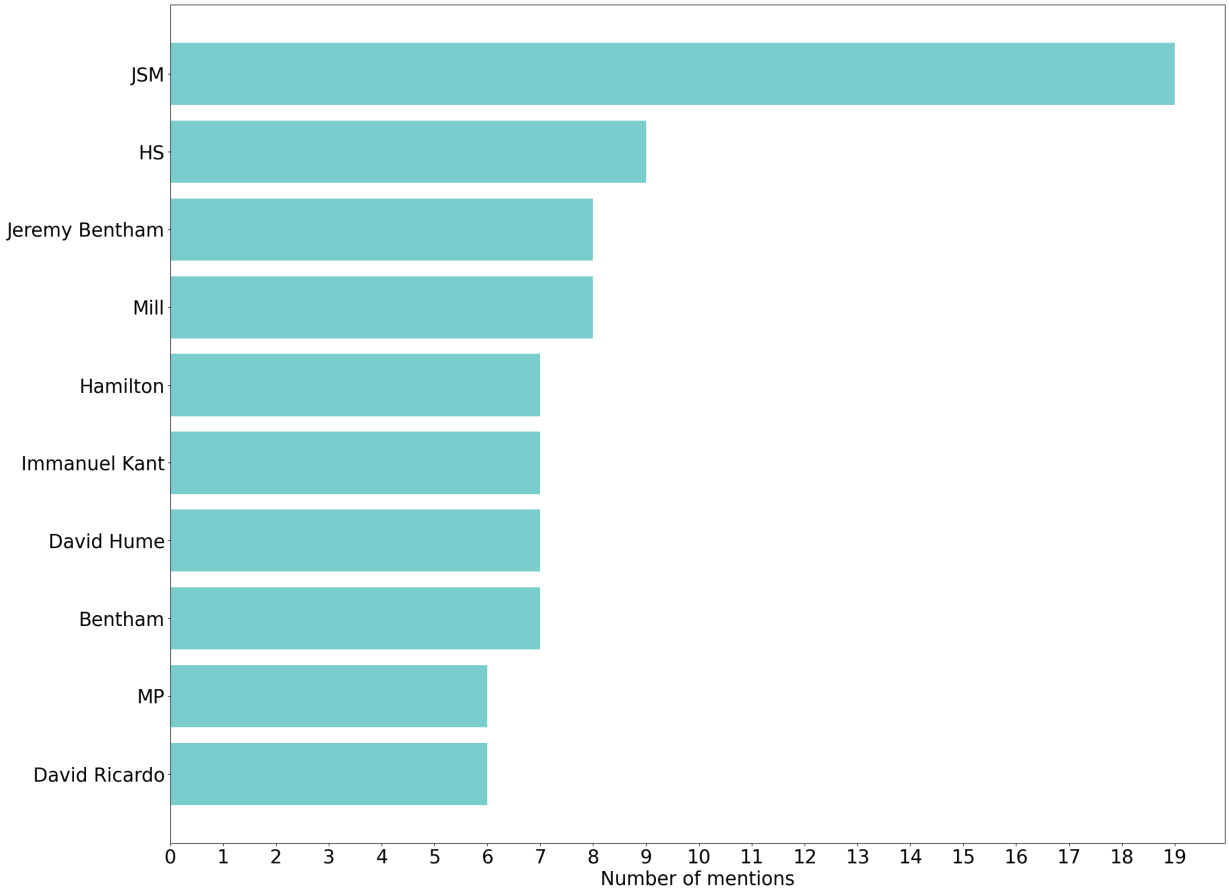

Figure 8.9: Most commonly mentioned entities for philosophers

Where JSM is John Stuart Mill, HS is Herbert Spencer, WEG is William Ewart Gladstone, SWH is Sir William Hamilton and TM is Theodore Mommsen.

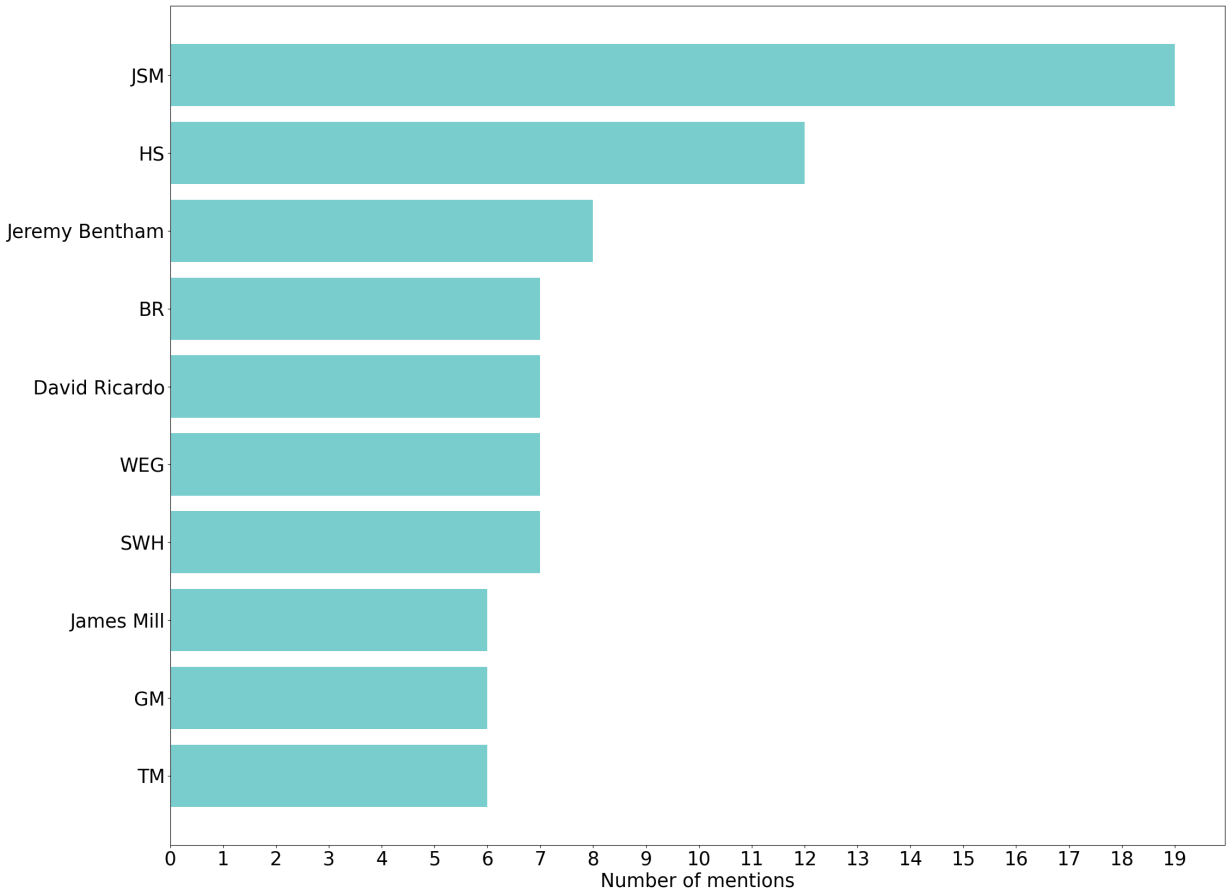

Figure 8.10: Most commonly linked articles for philosophers

Where JSM is John Stuart Mill.

Unlike the mathematics results, the philosophy results actually illustrated the methods working rather well on the data set. With the exception of Mill, Hamilton and Bentham, the majority of the top 10 names picked up using Spacy, illustrated in Figure 8.9, contain a first name and a surname. With the single word names picked up, it is at least easy to identify that Bentham is probably referring to Jeremy Bentham. Unlike the family name of Bentham, there is a chance that at least some of these mentions are discussing Harriet Taylor Mill. However, taking into account the patterns of the commonly linked people between both mathematicians and philosophers, as well as the fact John Stuart Mill has been shown in Figure 8.9 and 8.10 as a clearly influential person in the philosophy network, it is probably the case that Mill is another reference to John Stuart Mill. The name of Hamilton runs into the same problems as the single names in mathematics, but the additional presence of William Hamilton within this list may indicate who it is referring to.

Both John Stuart Mill and Herbert Spencer are linked from the Wikipedia pages of philosophers more than the most linked mathematician overall, and seven out of the top most linked to pages of from philosophers articles are linked at least the amount of times as the most linked pages from mathematicians. This suggests a higher level of interconnectivity between articles of 19th century Wikipedia philosophers. This

indicates a higher level of recognition in the writing of the articles of the connections between philosophers.

It is worth highlighting the indicated importance that these figures suggest about John Stuart Mill. He is both the most linked and mentioned in the articles of 19th century philosophers, and stands well above the second figure (Herbert Spencer), with mentions from 19 articles using Spacy, and 19 articles using Wikidata. With only 50 philosophy articles involved in the analysis, this illustrates a large number of the philosophy articles refer to Mill. Unlike the commonly linked mathematicians, Mill was never associated with a university, in either a teaching position or as a student, an the only official post he held was Rector of the University of St Andrews (48). While the influence of Mill's philosophy has already been widely accepted, I believe this further strengthens the point that his work was highly influential for 19th century philosophers. Nothing in the literature indicates a significantly more prolific collaboration style (aside from his work with his wife - Harriet Taylor Mill), and so I believe that the impact of his publications has led to this increased linkage between his article and the articles of other 19th century philosophers. While it is harder to find literature on the connections and collaborations of these philosophers, the statistics in Figure 8.9 and 8.10 indicate a few names (Mill, Spencer and Bentham) that may lend well to further study of the collaborative style, to help distinguish why these people are commonly linked within this network, looking into whether it is down to the personal connections of these philosophers that place them so high, or perhaps more likely, the influence of their work.

8.4 Conclusion

In conclusion, the results from analysis of the popular figures in the networks have provided a few insights. The methods for NER and link recognition are not infallible. This is illustrated in the commonly linked people section, with "names" such as Senior Wrangler and Natural Philosophy being picked up. The additional entities picked up by Spacy however still helped to illustrate the strong presence of Cambridge mathematicians within the network.

John Stuart Mill was the most commonly mentioned and linked person in the articles of 19th century philosophers, providing testament to the influence of his works. The philosopher articles tended to mention commonly mentioned people more often than the most mentioned figures for mathematicians, indicating a strong level of interconnectivity between pages for philosophers.

For mathematicians, Euclid was the most commonly named person and John Clerk Maxwell was the most commonly linked. The mentions of people in the articles of 19th century mathematicians seems to further indicate a heavy Cambridge bias in the institutions that the mathematicians were associated with.

9 Epsilon Analysis

The following section will cover what Epsilon is and how this project was integrated into my analysis.

9.1 Description of Epsilon

After discussions around collaborative networks of mathematicians, I was pointed towards Epsilon (Epistles of Science In the Long Nineteenth Century), a Darwin Correspondence Project and Cambridge University Digital Library collaboration. Epsilon is a platform used to store a record of correspondences between a variety of scientific figures in the 19th century, including Darwin, John Herschel and Michael Faraday. Although the content is unavailable for web scraping, there was a possibility that I was going to be able to get access to the equivalent data in XML form. This data could then have been used for a comparison against the networks that were formed in the Wikipedia articles, and the known correspondence networks, allowing an assessment of the contents of the given Wikipedia articles. Unfortunately, the timings did not work out for this project, but with the help of Brigitte Stenhouse, I have access to the data for the correspondence of Mary Somerville, which is due to be uploaded to Epsilon in the upcoming months. This has allowed me to perform a comparison with Mary Somerville's collaborations as represented through her letters and her Wikipedia article, and I'm deeply indebted to Brigitte for allowing me to use this data, and for discussions and resources on Somerville's networks.

9.2 Mary Somerville case study

Mary Somerville was the first person used to test the code by manually extracting all the named people on her Wikipedia to compare with the code output. Somerville, who was born in 1780 and died in 1872, was a British mathematician and 'scientist' (the term was actually first used to describe her in a review of her work). Somerville was born in Jedburgh, but spent most of her childhood in Burntisland. In her childhood, she learnt traditional skills for girls at the time, such as sewing and painting but also developed a keen interest in intellectual subjects such as Latin and Mathematics, through support from her uncle. Somerville's first marriage to Samuel Greig lasted only 3 years, before Greig passed away. Although this was a personal tragedy, it gave her more intellectual freedom and opportunity (49). Somerville eventually remarried, and the academic and personal support from her husband was critical in allowing her to rise to the position in scientific society that she did. She is primarily remembered for her 'translation' of the first two volumes of Laplace's *Traité de mécanique céleste*, where she was believed to be the only mathematician capable of understanding the material in Britain. Her translation was not just a standard translation from French into English, she managed to grasp the new techniques of the text well enough to provide her own commentary and thoughts, as well as adapting it to suit a British audience. Throughout her life, she continued to publish regularly, including her book *On the Connexion of the Physical Sciences*. She served as a model and inspiration to many women scientists (50) and is remembered today with an Oxford college named after her and her face on the Scottish £10 note.

Being connected to a network was important for female scientific figures like Somerville. Although there was a slight gradual opening across the 19th century, during Somerville's lifetime, many were closed to women. Somerville still managed to be part of the conversation around these institutions. She was also offered many honorary fellowships and memberships, including being one of the first women admitted to the Royal Astronomical Society (RAS) in 1835 with Caroline Herschel, but these did not bring much advantage outside prestige (17).

9.2.1 Results of analysis

Out of the 148 people (and organisations - places like RAS) that Somerville (or her husband - who acted as her publishing agent (17)), 17 of the correspondents feature on the Wikipedia article. Using both this analysis and a review of literature around her connections, the kind of relationships highlighted on Somerville's page and the others that are missing can be discussed.

The people identified both in the Wikipedia articles and the record of correspondents with the number of letters between the pair are as follows :

- Charles Babbage (33)
- Jean-Baptiste Biot (5)
- David Brewster (3)
- Henry Broughham (11)
- Charles Darwin (1)
- Maria Edgeworth (14)
- John Herschel (92)
- Charles Lyell (12)
- John Murray (18)
- Pierre-Simon Laplace (2)
- Sir James Mackintosh (6)
- Mary (1)
- George Peacock (1)
- Thomas Phillips (2)
- Siméon Denis Poisson (1)
- John Playfair (2)
- William Wallace (3)

This list of 17 common names between the Wikipedia and the record of correspondents contains some formidable figures of 19th century science and literature. With the exception of Mary (discussed below), all of the people found in both data have their own Wikipedia page and are *linked* from Somerville's article.

The presence of Mary in the data raises a question about what should be manually recorded as mentions of names in the article. In the Wikipedia article for Somerville, both she and her family are often referred to by their first names only. When looking for commonly named people between the data from the Wikipedia articles and the Epsilon data, my code is identifying a close enough match with the mention of Mary in the Wikipedia page and the correspondence to Maury. Although mentioned by only his last name, this correspondence can be identified through context as a correspondence between Somerville and Matthew Fontaine Maury, an American scientist famous for his work on oceanography. In this correspondence he was thanking her for a copy of the physical sciences (Bod MS Dep 381 MS 3/195 summarised by Stenhouse). For the purpose of comparing with the mentions of the Wikipedia data, standards such as referring to people by a single, full name throughout the file would have been easier for the analysing the data, but this data was never collected with the intention of this kind of analysis.

Once the names of unique (non family) correspondents of Somerville were identified from the Epsilon data, they were ordered them in terms of number of letters. While not a direct guarantee of stronger connection in terms of high number, or weaker in terms of lower (as verbal/in person relationships are not as represented), some interesting patterns begin to emerge. The top five most corresponded people and there number of correspondences are as follows:

- John Herschel (92)

- John Murray III (62)
- Ada Byron (39) - counted as two separate entities Ada Byron (King) - 23 and Ada Byron (16) - the difficulties of multiple entity recognition is discussed in the difficulties section
- Charles Babbage (33)
- John William Lubbock (22)

John Herschel (astronomer and mathematician, member of the test set), John Murray III (publisher of Somerville's books) and Charles Babbage (mathematician known for his difference/analytical engine) have all been identified as being in both the correspondence and the Wikipedia article. Ada Byron should have been identified as common between both the article and the correspondence network, if she wasn't referred to as her more famous alias, Ada Lovelace, in all of her references in Somerville's article.

John William Lubbock, despite having a Wikipedia article, is not mentioned or linked to from Somerville's page. Somerville and Lubbock had a productive working relationship, and Lubbock was especially complementary of her work, stating that Somerville puts the Cambridge mathematicians to shame (Bod MS Dep c.371 MSL 5/119 summarised by Stenhouse). They frequently used each other as proof readers, with examples including Somerville asking for his observations on her account on the Lunar Theory (RS, JWL, S.286 summarised by Stenhouse) and Lubbock asking Somerville to check his moon calculations and comment on his paper on comets before it went to press (17).

Two prominent people missing from the Wikipedia article are chemists, Jane Marcet (who had 8 correspondences) and Michael Faraday (who had 15). It may be interesting to note that while Marcet does mention Somerville on her Wikipedia page, there is no mention of Somerville on Faraday's page.

Jane Marcet, like Somerville, was the author of many 'popular' science books, including most famously *Conversations on Chemistry.* As well as chemistry, she was involved in the field of political economy, and her *Conversations on Political Economy* was the 'first explicit attempt to render the subject accessible to the general public' (51). Perhaps it would be interesting to explore the relationship between publishing books of these kind and women's access to science during Somerville's and Marcet's life, but this is outwith the scope of this project. Through her relationship with Marcet and her husband, a physician and fellow of the Royal Society, Somerville was introduced to many of the influential figures in the London scientific community (17). Through summaries of the correspondences involving Marcet and Somerville they seem to have a fairly informal relationship, with discussions often centered around party and dinner invitations (Bod MS Dep. c 370 MSE 1/177, 2/177-185).

Faraday, another missing figure in Somerville's Wikipedia page, was supposedly inspired into his life of science through reading Marcet's *Conversations on Chemistry.* Faraday, remembered for his work in electricity and electrochemistry, was a prominent figure in 19th century British science. He lacked any formal education past the age of 13, and made his break into the scientific world through correspondence with Humphry Davy at the Royal Institution (RI) (52). Faraday continued to have a strong relationship with the RI, founding the Christmas Lectures series in 1825 which still runs today. Faraday held a similar type of relationship to Somerville as that of Lubbock, sending and receiving papers, as well as helping with corrections on her work (17). It is worth noting, these were corrections on the non-mathematical works of Somerville (Connexion and Physical Geography), as despite his high level of competency in chemistry, he is often described as being mathematically illiterate (53).

Through the people missing from this analysis, it is hard to draw any pattern on the types of relationships being missed out on the Wikipedia representation. With the addition of more people into this analysis in the future it is hoped that more patterns will begin to emerge with a large subject group.

9.3 Limitations of Epsilon analysis

Like the fact that Wikipedia does not claim to be a complete representation of the information about the figures, the correspondence network can also fail to be truly representative of the collaborative networks that the figures were participating in. The problem is two fold; the archive of the material can be incomplete and all archives exhibit bias in the material that ends up being documented and collected. However, even with full(er) collections, there is the problem of a lack of representation for relationships conducted through other mediums, such as in person meetings. 'Archives are both documents of exclusion and monuments to particular configurations of power' (54) and it must be acknowledged that there is no current way to test how representative the letter collections used for comparison is of the true network of collaboration. While the limitations are still present in the correspondents network, I believe the benefit of using this data outwits the negatives. The correspondents network allows for a very objective idea of who a given figure was collaborating with, and is free of the bias in place when a given person is writing an article. Who they chose to include in Wikipedia articles may be the result of their background and knowledge, where as correspondent networks are not influenced by this, as they store information on the actual networks of the figure at the time.

9.4 Further uses for this data

When the Epsilon data is made available for public use for computational uses, a similar kind of analysis could be performed on other people in the Epsilon data set, including Michael Faraday, John Tyndall and John Herschel. It is hoped that a comparison with known correspondence networks will allow strong and solid conclusions to be made about the representation of the networks of scientists in their Wikipedia articles. If a similar data set of collected letters exists for 19th century philosophers, the comparison of how representative the articles were for these groups could be extended.

10 Evaluation

The following section will detail particular challenges that occurred in the development process and look at how well the project was able to achieve its original goals. Additionally, areas of interest for future study related to the findings of this report and work will be discussed.

10.1 Implementation Challenges

The development of this project was challenging at times. Trying to balance the two subjects meant I was often unsure on where best to focus my attention and this section will discuss the problems faced, in both code development and research.

10.1.1 Technical

Towards the middle of the project, the network analysis task seemed very overwhelming. To be able to correctly form judgements on the type of relationships mentioned in these articles, as well as which important figures were missing off the articles seemed an incredibly large task. Instead, pivoting the framing of the project to be able to think of it as an introductory explore of what sort of analysis could take place on the data collected, and realising that I would not have to have a detailed level of understanding of the collaborative style of each subject of the article to perform an analysis on portrayal meant that I felt more confident that the goals of this project would be achieved.

The hardest part of code implementation actually came out of a task that was believed would be easy to achieve within a short period of time. When looking at a Wikipedia page, the page may actually link to a significant number of people that are not direct collaborators of the subject, due to the category section at the bottom of the article, as illustrated in Figure 9.1. These people were not relevant to the analysis, but the way the link collection was set up did not lend itself well to being able to filter out specific sections of the article. Now, specific headings act as cut off points, where it is known after this point anyone who is linked is not relevant to the analysis performed. Example headings that act as cut off points include References and See Also. This seemingly simple problem turned into weeks of work to ensure the count of links for each article was not made unnecessarily large by references to irrelevant people. While if the articles were references to people from a different period of time, they would have been filtered out in the next stage of the code, there is no guarantee they are not also 19th century figures, and it is a lot of additional work to have to do all these checks for articles that we are aware aren't relevant, hence why there needed to be cut off headings, after which we did not look for any more links from the article.

Figure 10.1: Linked people present at the bottom of Mary Somerville's Wikipedia page

Currently, the code is not able to distinguish between multiple references to the same person. There is no guarantee that the count of the number of links does not include the same person being counted multiple times, due to this problem. While steps have been taken to try and circumvent this, including making the list of names into a set when the code is run to filter out duplicate mentions of the same person if they had been referred to in the same way each time, the multiple references problem was beyond the scope of work that could be completed in the given time. Ideally, to provide a more accurate representation of how linked these articles are, one person should only be counted once for each article, regardless of the number and type of mentions.

10.1.2 Research based

The major challenge this project faced in terms of research was the lack of literature on collaborative styles of philosophers, in the 19th century and beyond. Having this literature would have been both complementary to the survey of the collaborative styles of mathematicians, and helped to provide an insight into the reasons behind the statistics being the way they were. An awareness of the lack of literature earlier in the development process, maybe would have led to a changing of comparison groups. An alternative would have been to pivot the analysis to physicists or astronomers, but an immediate problem with this is the lack of distinctive line between people classed as mathematicians and people classed as scientists in adjacent areas. While there is still crossover between mathematicians and philosophers (George Boole in the 19th century, Bertrand Russell in the 20th), the line is clearer, and there is definitely less overlap than between mathematics and the physical sciences.

One problem that was not anticipated was the level of expertise required to identify the people that were missing from Wikipedia pages. While the statistics tell us about the people who have been included in the article, there is no way of knowing who hasn't without an in depth understanding of the life of the given figure. This is where tools such as Epsilon allow me to get a more complete idea of who has not been featured on the Wikipedia articles, but before this comparison, I often felt unsure of how I could present an analysis of the how representative the articles of these figures were. By instead focusing on the comparison I was able to pick up on statistics that allowed me to begin to assess how representative these articles actually

were, but I am also incredibly glad that Epsilon allowed for a more solid comparison.

10.2 Goals of the project

At the beginning of the project, the Description, Objectives, Ethics, Resources (DOER) report allowed me to create a set of goals that I wanted to achieve with my time spent on the project. Like every project, these goals evolved and adapted to the work that performed throughout the year. I came in with a developed interest in NLP, but not as much technical experience or awareness of the challenge that working with natural language is.

Below is a list of the original goals stated in the DOER.

- To perform analysis of the network of 19th century British mathematicians and philosophers on Wikipedia to compare the networks created between the two fields
- To be able to construct the networks formed between other mathematicians/philosophers based on a Wikipedia article, including all mathematicians/philosophers linked on the article respectively that do not have their own article, by using processes involving web scraping and NLP
- To evaluate the success of the NLP techniques applied by comparison with a manually created network

Below is a list of three secondary objectives of this project.

- To choose the correct NLP tools to allow for the networks to be created with the maximal amount of accuracy and looking at alternatives if a desired level of accuracy is not reached
- To perform a historical analysis of the collaborative style of mathematicians and philosophers in 19th century Britain to see how these correspond to what the networks suggest
- To provide a discussion about how collaboration in both subjects has evolved since the work period of the cohorts being compared

The main differences between these original set of goals and the objectives stated in section 3.3 is the decreased emphasis on construct networks, and looking at network graphs of mathematicians and philosophers. While the final product can still provide an analysis of the networks formed within these Wikipedia articles, I have not produced any network graphs since early in the development process. It was realised that the information that would have been represented in these diagrams could still be represented through a less visual, but also more comprehensible way, by using the statistics evaluated in the results analysis section.

10.2.1 Technical

All my primary technical goals have been achieved.The results section analyses the networks of the 19th century mathematicians and philosophers, and the NLP analysis section evaluates the differing success rates of NLP methods used to extract the names from the Wikipedia articles.

Out of the methods that I tested during development, Spacy was the most accurate, and therefore the secondary objective of maximising the accuracy of the NER models is partially achieved. However, relative to research standards, the accuracy of Spacy is still relatively low. Initially I was aiming for an F1 score of above .9, in order to ensure that the application was reliable enough to provide solid analysis, however throughout development, none of the methods ever performed to this standard. As a result, the application is a less accurate at NER than ideally expected.

I have not constructed network diagrams with the information retrieved from the program, although this is well within reach of the current scope. All the information to put together these network diagrams is currently all here, but for reasons stated above, I decided against continuing with these figures. I do not believe the analysis is missing anything due to the lack of diagrams, but this secondary goal was not achieved.

10.2.2 Research based

While the literature has restricted me from providing a thorough analysis of the networks of the 19th century philsophers, I have managed to give a background section that places the results of the analysis into context. I discussed how collaboration evolved in the sciences over the given location and time period, and have linked the knowledge gained from this literature review with the results of the analysis. This combination of the historical knowledge with the results has provided an insight and overview into the collaborative work of both mathematicians and philosophers, hence I have achieved all of my primary objectives that were research based.

I was only able to partially complete the secondary research goal of assessing the portrayal of mathematicians and scientific collaborative practices against records of their known correspondence. Despite the immense help that the team were at Epsilon, the licensing situation meant that I was unable to gain access to the records in time for submission of this project. Thankfully with the data that Brigitte Stenhouse supplied I was able to provide a preliminary case study of the kind of insights that combining the results of the Wikipedia scraping with known correspondence could provide.

10.3 Future Expansions

Having worked on the project for almost a year now, I am in an ideal situation to know where the most productive next steps that this work could take. Five potential areas of interest have been identified that could be combined with this work. It is believed these areas would provide a lot more insight into the portrayal of collaborative styles on Wikipedia.

10.3.1 Study of the collaborative style in Philosophy

The combination of this analysis with a collection of research on the collaborative styles of 19th century British philosophers would lend a lot more insight into why the patterns that did have emerged in the results analysis. To be either able to locate old or combine new research into this area with these results would allow for a much stronger case to be made in the analysis section, and for the much needed context to be given on collaboration within philosophy.

10.3.2 Expansion to consider other subjects

A fairly straightforward expansion of this analysis would be to consider alternative subjects. The comparison was kept across mathematicians and philosophers due to a wish to compare across the discipline boundaries of science and humanities subjects, but the existing literature around collaboration in the physical sciences could provide key insights. Subjects such as physics, chemistry and astronomy all have existing literature around collaborative style over this period, some of which has been discussed in previous sections. This research, combined with the fact these sciences are all possess common objects (the sky, the laboratory) that may indicate a higher level of collaboration, would be very interesting to combine with results from Wikipedia analysis.

10.3.3 Improvement of NER methods

Another immediate task to be completed if this work was returned to would be the improvement of the NER methods. As stated above, the Spacy library was the best performing of the methods tested, but it still was performing at a relatively low level of accuracy. Improving the accuracy of the NER methods should be a priority, and introducing features such as the identification of multiple references to the same person in an article should be a priority.

Up until Spacy version 3, which was released in at the start of 2021, Spacy followed a transition based approach for parsing the text, and a 'Embed. Encode. Attend. Predict' framework. A transition based dependency parser only creates one tree, starting from the left side of the data and working its way to the right. At every stage in the entity recognition, a prediction is given from the current state and this

process is repeated until the parsing finishes (and a termination state is reached) (55). When recognising names, the model creates multi-token names, so 'Mary Somerville' would be composed as a single item (56). While this transition based dependency parser is still used in this project, Spacy also has a new framework is now described as 'Embed. Encode. Reduce. Predict'. The first two stages of this framework are now the responsibility of transformers. Transformers are a type of deep learning model, that are responsible for extracting the 'context sensitive representations' of tokens (57), replacing the parser and tokeniser. They are a recent advance in the field of NER, and they have shown to significantly improve performance. Compared to Spacy's 'en_core_web_lg' (standard English) model with an accuracy of 85.5 for NER, the transformer based model, 'en_core_web_trf', which uses the roberta base, has an accuracy of 89.8 (58).

Using transformers does have some drawbacks, including a longer run time, but the following advantages offered make this a worthwhile trade off.

- Good at extracting vectors that represent meaningful representation
  - One problem that has faced NLP is the knowledge acquisition pipeline. This is where the information is all stored in the text, and the previous tools have been able to make objective statements about the content of the text but not necessarily the meaning. Transformers perform well at extracting meaningful representation of the text (59), which can also be useful in other applications of Transformers from sentiment analysis to translation.
- Increase accuracy
  - As discussed previously, moving from a standard Spacy training model to one using transformers provided an increase of over 4% in accuracy (58)
- Use GPU more efficiently
  - One benefit of using transformers is they work efficiently with GPUs, which although being less versatile than CPUs, are continuing to improve in performance (60)

One way to continue increasing the accuracy is to consider changing the Spacy NER to use the Transformer based model instead.

10.3.4 Continued Integration with Epsilon

Many of the figures for the hand analysed networks were chosen due to their appearance on Epsilon. It is expected that this data will be available for analysis soon, but it was disappointing to not be able to continue to integrate the Epsilon data and the results of this work, as it provides an effective means of determining which people are typically being missed out from the analysis. While it will be after this project, I am going to have the opportunity to work with the Darwin correspondence data this month, placing me in an ideal position to be aware of the ways this data can be manipulated, if I ever returned to this work and wanted to combine it with the Epsilon data.

10.3.5 Additional data collection

When requesting data from the Wikidata information associated with each article, there is also an opportunity to get information about their occupation. Adding this into the analysis could allow for an exploration into the type of relationships being portrayed on these Wikipedia articles, and how often the articles are referring to people with whom the subject has a working relationship with. This filtering of the types of relationships that we are interested in analysing on the Wikipedia page could provide a more accurate representation of the portrayal of the types of relationships contained within these articles.

Similarly, additional data could also be analysed on the quantity of mentions of particular people in articles. Currently the analysis extends to looking at the most commonly linked people between articles but this does not consider how many times these people are being mentioned in a given article. This extra information, which is stored implicitly in the files for each article could provide insight into the strength of

the relationships between the subject of the page and the mentioned name, with higher quantity of mentions perhaps indicating a stronger and more influential relationship.

11 Conclusion

This report has outlined a new way to assess the collaborative networks of mathematicians and philosophers in the 19th century. A review of existing literature around the changes in collaboration style over the century was presented.

The development of code used to extract names and links from the Wikipedia pages of 19th century British mathematicians and philosophers was discussed, and the methods used at various points in development to extract named entities were compared in order to choose the most accurate strategy. The code was tested on a set of test figures who had their Wikipedia pages manually scraped for names in order to evaluate the accuracy of the methods for recognising names and articles about people linked from a given Wikipedia page, and the F1, precision and recall statistics of these methods are presented. Once a desired level of accuracy was achieved with the methods, this code was applied to the networks of 19th century British mathematicians and philosophers, and the results from this are discussed.

The results from this analysis of Wikipedia pages found that mathematicians have significantly higher mentions per 1000 characters than philosophers, however philosophers have significantly longer character lengths. When comparing a typical length article for both groups, articles about philosophers will mention more people. This shows that despite the typical portrayal of mathematicians as lone geniuses in popular media, the Wikipedia articles are representing a collaborative network, even if the articles are shorter in length. It is positive that the number of mentions is significantly higher for mathematicians, as this helps ensure that the collaborative nature of the subject is being presented to the public.

Additional evaluations with data related to the gender of the article subjects were performed, as well as comparing the output from the code running on Mary Somerville's Wikipedia page with her known collaborative network. It was discovered that some key relationships identified in the correspondence data were missing, but out of the top five people corresponded with, four of them were present on the Wikipedia article.

Overall, a new way of assessing the collaboration that doesn't rely on the ever changing co-authorship standards has been presented. This allowed working relationships that didn't necessarily result in publication together to be added to the discussion around collaborative networks. This kind of analysis could provide additional insight when further applied to other groups with supporting background material on the collaborative styles.

A Bibliography

References

[1] S. Singh, *Fermat's Last Theorem*. Harper Perennial, 2007.

[2] A. Montuori and R. Purser, "Deconstructing the lone genius myth: Toward a contextual view of creativity," *Journal of Humanistic Psychology*, vol. 35, pp. 69–112, Jul 1995.

[3] Wikimedia Statistics, "Wikistats - statistics for wikimedia projects," Feb 2022. [Online]. Available: https://stats.wikimedia.org/#/en.wikipedia.org

[4] Wikipedia, "Size of wikipedia," Feb 2022. [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

[5] R. Rosenzweig, "Can history be open source? Wikipedia and the future of the past," *The Journal of American History*, vol. 93, no. 1, pp. 117–146, 2006.

[6] E. Vardi, L. Muchnik, A. Conway, and M. Breakstone, "Wikishark: An online tool for analyzing wikipedia traffic and trends," in *Companion Proceedings of the Web Conference 2021*, ser. WWW '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 558–571.

[7] A. M. Weinberg, "Impact of large-scale science on the united states," *Science*, vol. 134, no. 3473, pp. 161–164, 1961.

[8] J. Cummings and S. Kiesler, "Coordination costs and project outcomes in multi-university collaborations," *Research Policy*, vol. 36, pp. 1620–1634, Dec 2007.

[9] T. A. Finholt and G. M. Olson, "From laboratories to collaboratories: A new organizational form for scientific collaboration," *Psychological Science*, vol. 8, no. 1, pp. 28–36, 1997.

[10] D. deB Beaver and R. Rosen, "Studies in scientific collaboration," *Scientometrics*, vol. 1, no. 1, pp. 65–84, Sep 1978.

[11] R. Döbler, "Continuity and discontinuity of collaboration behaviour since 1800 — from a bibliometric point of view," *Scientometrics*, vol. 52, pp. 503–517, Nov 2001.

[12] M. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101 Suppl 1, pp. 5200–5, May 2004.

[13] J. Fagan, K. S. Eddens, J. Dolly, N. L. Vanderford, H. Weiss, and J. S. Levens, "Assessing research collaboration through co-authorship network analysis," *The journal of research administration*, vol. 49, no. 1, pp. 76–99, 2018.

[14] P. Rothman, "Grace chisholm young and the division of laurels," *Notes and Records of the Royal Society of London*, vol. 50, no. 1, pp. 89–100, 1996.

[15] I. Grattan-Guinness, "A mathematical union: William Henry and Grace Chisholm Young," *Annals of Science*, vol. 29, no. 2, pp. 105–185, 1972.

[16] J. Habgood-Coote, "What's the point of authors?" *British Journal for the Philosophy of Science*, forthcoming.

[17] B. Stenhouse, "Mary Somerville : Being and Becoming a Mathematician. PhD thesis. The Open University," 2021.

[18] H. Ellis, "Knowledge, character and professionalisation in nineteenth-century british science," *History of Education*, vol. 43, no. 6, pp. 777–792, 2014.

[19] A. Chapman, *The Victorian Amateur Astronomer : Independent Astronomical Research in Britain 1820-1920*. Gracewing, 2017.

[20] D. deB Beaver and R. Rosen, "Studies in scientific collaboration: Part II. scientific co-authorship, research productivity and visibility in the french scientific elite, 1799–1830," *Scientometrics*, vol. 1, no. 2, pp. 133 – 149, 1979.

[21] R. Collins, *The sociology of philosophies: A global theory of intellectual change.* Belknap Press of Harvard University Press, 1998, p. 858–883.

[22] J. A. Miller and E. Schleisser, "A case for co-authorship in philosophy (guest post by Joshua A. Miller and Eric Schliesser)," Dec 2017. [Online]. Available: https://dailynous.com/2017/07/20/a-case-for-co-authorship-in-philosophy-miller-schliesser

[23] Walber, *Precision and Recall.* Wikipedia, Nov 2014. [Online]. Available: https://en.wikipedia.org/wiki/F-score#/media/File:Precisionrecall.svg

[24] K. D. Foote, "A brief history of natural language processing (NLP)," Jun 2019. [Online]. Available: https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/

[25] M. Won, P. Murrieta-Flores, and B. Martins, "Ensemble named entity recognition (NER): Evaluating NER tools in the identification of place names in historical corpora," *Frontiers in Digital Humanities*, vol. 5, 2018.

[26] Read Coop, "Transkribus," Dec 2021. [Online]. Available: https://readcoop.eu/transkribus/

[27] Impresso, "Media monitoring of the past." [Online]. Available: https://impresso-project.ch/

[28] S. Gupta, "Named entity recognition: Applications and use cases," Feb 2018. [Online]. Available: https://towardsdatascience.com/named-entity-recognition-applications-and-use-cases-acdbf57d595e

[29] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, and Y. LeTraon, "A replicable comparison study of ner software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2019, pp. 338–343.

[30] eXplore Computer Science Engineering Courses, "Attachment ambiguity in natural language processing," Mar 2020. [Online]. Available: https://www.exploredatabase.com/2020/03/attachment-ambiguity-in-natural-language-processing.html

[31] M. Honnibal, *Spacy's Entity Recognition Model: incremental parsing with Bloom embeddings & residual CNNs.* YouTube, Nov 2017. [Online]. Available: https://www.youtube.com/watch?v=sqDHBH9IjRU&t=1081s

[32] J. D. Choi, J. Tetreault, and A. Stent, "It depends: Dependency parser comparison using a web-based evaluation tool," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 387–396.

[33] Spacy, "Language processing pipelines : Spacy usage documentation." [Online]. Available: https://spacy.io/usage/processing-pipelines

[34] ——, "Linguistic features, Spacy usage documentation." [Online]. Available: https://spacy.io/usage/linguistic-features#how-tokenizer-works

[35] ——, "Token, Spacy api documentation." [Online]. Available: https://spacy.io/api/token#attributes

[36] D. Jurafsky and J. Martin, "Dependency parsing," in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Feb 2008, vol. 2.

[37] M.-C. Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and C. Manning, "Universal stanford dependencies: A cross-linguistic typology," *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pp. 4585–4592, Jan 2014.

[38] Spacy, "Training pipelines &; models · spacy usage documentation." [Online]. Available: https://spacy.io/usage/training

[39] E. Loper and S. Bird, "NLTK: The natural language toolkit," *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, May 2002.

[40] S. Bird, E. Klein, and E. Loper, "Preface," in *Natural language processing with Python: analyzing text with the natural language toolkit*. 0'Reilly Media, Inc., 2009.

[41] ——, "Extracting information from text," in *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc, 2009.

[42] C. Wagner, E. Graells-Garrido, D. Garcia, and F. Menczer, "Women through the glass ceiling: Gender asymmetries in wikipedia," *EPJ Data Science*, vol. 5, no. 1, 2016.

[43] F. Tripodi, "Ms. categorized: Gender, notability, and inequality on wikipedia," *New Media & Society*, 2021.

[44] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier, "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia," *arXiv e-prints*, p. arXiv:1501.06307, Jan. 2015.

[45] A. Warwick, "Writing a pedagogical history of mathematical physics," in *Masters of theory Cambridge and the rise of Mathematical Physics*. University of Chicago Press, 2003.

[46] Scott Taylor, "Euclid in the 19th century." [Online]. Available: http://personal.colby.edu/~sataylor/teaching/S09/MA111/Euclid.pdf

[47] A. Moktefi, "Geometry: The Euclid debate," in *Mathematics in Victorian Britain*, Jan 2011, pp. 320–336 445.

[48] C. Macleod, "John Stuart Mill," Aug 2016. [Online]. Available: https://plato.stanford.edu/entries/mill/

[49] B. Kendall, "Mary Somerville: The queen of 19th-century science," Nov. 2021. [Online]. Available: https://www.bbc.co.uk/programmes/w3ct1rm3

[50] B. Renee, *Maria Mitchell and the Sexing of Science: An astronomer among the American Romantics*. Beacon Press, 2008.

[51] S. Bahar, "Jane Marcet and the limits to public science," *The British Journal for the History of Science*, vol. 34, no. 1, p. 29–49, 2001.

[52] BBC, "Michael Faraday (1791-1867)." [Online]. Available: https://www.bbc.co.uk/history/historic_figures/faraday_michael.shtml

[53] J. J. O'Conner and E. Robertson, "Michael Faraday - biography," May 2001. [Online]. Available: https://mathshistory.st-andrews.ac.uk/Biographies/Faraday/

[54] A. Procter, *The whole picture: The colonial story of the art in our museums and why we need to talk about it*. Cassell, 2019.

[55] M. Honnibal, "Transformers and 'embed, encode, attend, predict' in ner," Feb 2021. [Online]. Available: https://github.com/explosion/spaCy/discussions/6910

[56] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 260–270.

[57] Spacy, "Embeddings, transformers and transfer learning, Spacy usage documentation." [Online]. Available: https://spacy.io/usage/embeddings-transformers

[58] ——, “What’s new in v3.0, Spacy usage documentation.” [Online]. Available: https://spacy.io/usage/v3#features-transformers

[59] M. Honnibal, “Using spacy with hugging face transformers | Matthew Honnibal,” Jan 2020. [Online]. Available: https://www.youtube.com/watch?v=RB9uDpJPZdc&t=569s

[60] Explosion, “Spacy v3: State-of-the-art nlp from prototype to production,” Feb 2021. [Online]. Available: https://www.youtube.com/watch?v=9k_EfV7Cns0&t=235s

B Code execution

This document details the breakdown of the code, and which files are responsible for which sections. For an explanation of how to run the code, please refer to the guides folder. Within that folder, there are four documents detailing usage, required data and libraries as well as the structure of the file system storing the code.

B.1 retrainingSpacy.py

A file used to generate the retrained Spacy model. This must be run before anything else in the code as the user has an option to run the code on the. If the user does not want to evaluate the performance of the retrained model, the relevant lines that import the retrained model can be commented out in nerMethods.py, and the option for usage with the model can be removed in scraper.py.

B.2 restart.py

This is the main driver file. In order to run the code on the network or the test set of files this is the file that needs to be run. The usage of this file is detailed in usage.txt.

B.3 network.py

This is the code used if the name extraction wants to be run across the whole network of figures. It will extract a list of names from the category pages of both groups, and run the name extraction methods on all figures from both lists.

B.4 nerExtract.py

This file is responsible for extracting the names out of the article using NLP techniques. The names can be extracted using Spacy, a retrained Spacy model (which must be generated using retrainingSpacy.py before the code is run) or NLTK. Once the names have been identified they are written to a file in the corresponding subfolder within the output folder.

B.5 scraper.py

This file is responsible for extracting the linked articles. When this is run, internet connection will be required, although if connection is temporarily lost, it will continue executing and write that the articles currently trying to be accessed have no links. Once the links have been identified they are written to a file in the corresponding subfolder within the output folder.

B.6 comparisonCurrent.py

Code used to evaluate how well the NER and links extraction methods are working on the test set of figures. Generates a csv that stores the performance statistics for each method on each article of the test figure and then uses this file to generate required graphs.

The graphs generated by this method are as follows:

- Graph 1 - overall comparison of NER methods across whole set of test figures (nltk and Spacy), Figure 7.9
- Graph 2 - overall comparison of NER methods across subset of test figures (ntlk, Spacy, retrained Spacy), Figure 7.10
- Graph 3 - comparison of precision values across articles for nltk and Spacy, Figure 7.13
- Graph 4 - comparison of recall values across articles for nltk and Spacy, Figure 7.14

- Graph 5 - Spacy f1, precision and recall values across articles, Figure 7.11
- Graph 6 - nltk f1, precision and recall values across articles
- Graph 7 - retrained Spacy f1, precision and recall values across articles, Figure 7.12
- Graph 8 - wikidata extraction f1, precision and recall values across articles, Figure 6.2

If a new run has taken place on the network, the evaluation.csv file should be deleted so the statistics can be updated.

B.7 networkAnalysis.py

Code that performs the popular figures analysis and the Epsilon data analysis. Generates a csv that stores the number of mentions of each person, along with additional data such as what method was used to extract the mention and who they have been mentioned by. If the code is rerun on the network, this file (mentions.csv) must be deleted and the code will generate a new version.

Generates graphs for the popular figures of each method and category (Spacy maths, wikidata philosophy etc) as well as generating a .txt file that details output from epsilon analysis in the analysis folder.

B.8 results_analysis.R

Code that performs the statistical analysis on the data. Written in R instead of Python due to experience. networkAnalysis.py MUST HAVE GENERATED mentions.csv before this code is run. The following graphs are created:

- Graph 1 - sqrt mentions per 1000 characters divided by category (maths/philosophy) and method (Spacy/wikidata), Figure 8.2
- Graph 2 - log length of articles divided by category, Figure 8.3
- Graph 3 - mentions per typical lengthed article, Figure 8.4
- Graph 4 - sqrt mentions per 1000 characters divided by gender and category, Figure 8.5
- Graph 5 - log length of articles divided by gender and category, Figure 8.6

The following statistical tests are performed:

- Test 1 - ANOVA for the influence of category and method on mentions per 1000 characters
- Test 2 - ANOVA for the influence of category on length of article
- Test(s) 3 - Wilcoxon tests for the influence of category (and method) on mentions per 1000 characters
- Test 4 - ANOVA for the influence of category, method and gender on mentions per 1000 characters
- Test 5 - Tukey HSD to discover which combination of category and gender is significantly effecting the mentions per 1000 characters
- Test 6 - ANOVA for the influence on method, category and gender on the length of article

B.9 helper.py

This file contains functions that are used over multiple files, or have the potential to be reused. Never run independently, but imported over multiple files in the code to try and decrease cluttering and code repetition.

C Detailed performance of NER methods

	F1 overall	Recall overall	Precision overall	F1 subset	Recall subset	Precision subset
Spacy	0.7697	0.8394	0.7148	0.7537	0.8236	0.6999
NLTK	0.6433	0.7432	0.5714	0.6175	.7204	0.5474
Retrained Spacy	N/A	N/A	N/A	0.7101	0.8130	0.6334
Wikidata	0.8328	0.8069	0.8626	N/A	N/A	N/A