What made Welsh scientists notable between 1804 and 1919?

TAB23

2019-02-02

1 Project Description

The National Library of Wales is home to a collection of newspapers dated between 1804 and 1919, all of which were printed and published in Wales in those years. The aim of this project is to automate the scanning of these newspapers for information regarding scientists of those years, to store relevent information in a database that can be searched from a website. The purpose of this project is to explore what made scientists from a variety of fields notable in their day and, potentially, to present this information in a way that is palatable to young people. Notable, in this case, means why were they in the news, and will look at both fame and infamy, where the latter exists.

According to the Oxford English Dictionary, the word scientist can be defined as "A person who is studying or has expert knowledge of one or more of the natural or physical sciences"[1]. The nature of this definition means that the scope of this project could be considered much broader than Professors and Doctors, and could in fact include anyone who made a contribution to an area of scientific study. I feel it best to narrow this scope, at least to begin with, to specifically the fields of Chemistry, Mathematics, Physics, Biology and Astronomy. Should I find myself with a surplus of time, or a shortage of material, then I could widen my search to include fields like geology[2], engineering and psychology[3], all developing areas of study at the time.

2 Proposed Task

As I am coming into this project with no prior experience in natural language processing, a significant task on my part will be research into both understanding what NLP is conceptually, and understanding how to put these concepts into practice.

The largest task in this project is the processing of the newspapers themselves. That is to say, searching through hundreds of news articles in search of a list of specific keywords, and then scanning over those articles and stripping each one for information on Welsh scientists only, is no small matter. Language changes over time; words that once had one meaning may have an entirely different meaning nowadays. There is also the matter of the multiple meanings one word can have in different contexts - the manner of differentiating between those potential meanings is called lexical disambiguation[4], and will play a vital role in cases such as the word 'chemist' is searched, since this project is not interested in looking at the kind of chemist that dispensed medicine. Over the course of their life, people can move, change jobs and titles, get married and change their name, etc. I can either attempt to cover all of these changes that one person may experience, and gather a much more accurate and complete collection of what they accomplished, or I can simplify my search to match a name to the reason they were in the paper. Given what I would ultimitely like this system to be able to achieve, the former is ideal, though the latter is my main goal.

What I anticipate to be at least as large a problem is in the mistakes made by the National Library scanning system. Where the newspapers are old and faded, the ink was not properly scanned, and so when that text was written up (by what I can only assume was automated software), there remained huge errors that can be recognised as errors only through reading the newspapers themselves[5]]. I am currently exploring different options for handling these mistakes, including a very specific list of keywords that I have deemed 'relevent', alongside their various common misspellings as I find them in the articles. An alternative approach would be in accounting for these errors when parsing the information, and telling the software how to handle unexpected words, perhaps with a dictionary to search through and find the closest approximation of a word that makes sense in context. This wouldn't always be accurate, but it would allow for a greater degree of autonomy, which is the nature of this project.

Alongside the Natural Language processing element of this project, there is also the issue of what to do with the data once I have obtained it. Storing the information in a database seems like the most obvious choice, as it allows me the most options of how to present my findings. Before I can build this database, however, I will have to decide on exactly the kind of data I am looking for, to best design the structure. An alternative to a database would be a single xml file that stores all the information the web interface needs to access, and can be searched by the user.

The web interface is the only part of my system that the user interacts with - it will be a simple website that the user can search for information based on name, date, location, or any number of keywords they deem useful, and receive back the information that the other parts of the system has found and stored. This part of the project will be built for ease of use, and designed with young people in mind. While it is intended at this point to be nothing more than a basic search function, there is potential for it to grow into something much larger than the scope of this project requires.

3 Deliverables

Key deliverables in this project include the requirements, design and testing specifications, the coded elements of the project (this is the text processing program, database and web interface), a list of keywords that the processing program uses to search the National Library archive, and two short reports. One of these reports will explore the different natural language processing toolkits available for Python, and discuss my options and final decision. The second pertains to use of that toolkit, and evaluating the potential need for training the model on data I have gathered myself, versus adapting the current model to suit the needs of this project.

Given the nature of the kanban process for software development, which is part of how I will be structuring the workload, I have no set timeline for when these deliverables will be finished, only a sense of the order in which they will need to be produced. For example, the requirements specification will need to

be completed fairly early in the project, perhaps as early as week 2, in order to have a strong foundation upon which the rest of the project can be built. Something like the testing specification, however, will not need to be designed until I am ready to begin work on the software.

4 Bibliography

References

[1] *https://en.oxforddictionaries.com/definition/scientist* The source for the definition of scientist.

[2] Dawson-Adams, Frank *The Birth and Development of the Geological Sciences chapter VII* 1938 This text illustrates the history of geology up to the year of publication in 1938, and discusses the changes that underwent the field. Chapter seven is the only one pertinent to this outline, as it discusses how geology became a real science during the 18th and 19th centuries. I have included it here as a reference should I wish to expand my search to include geologists.

[3] *http://www.newworldencyclopedia.org/entry/History_of_psychology* There is a brief mention in the second paragraph about the nature of psychology in the 19th century, and its emergence as a serious area of scientific study. This text has been included as a point of reference should I wish to expand my search to include psychologists.

[4] Ayetiran, Eniafe Festus *Enhancing word sense disambiguation using a hybrid knowledge-based technique*, Natural Language Processing and Cognitive Sciece: Proceedings, 2014 This research paper was found in a collection published in 2014, and the introduction contains a lot of useful information on the theory behind lexical disambiguation.

[5] *https://newspapers.library.wales/view/3561154/3561162/37* This news article, an example of those found on the National Library of Wales archive, contains examples of common misscannings of words, such as 'lie' where 'He' was meant. The errorrs happen consistantly down the article. *Further Reading:*

[6] Jurafsky, Dan *https://web.stanford.edu/~jurafsky/slp3/* 2018 This is a starting point upon which more research will be built as I look into specific areas of natural language processing more thoroughly. Chapters that I believe will be most relevent are 2, 11, 13 and 17.

[7] *http://www.nltk.org/book/ch01.html* This book belongs to the Natural Language Toolkit, a NLP toolkit for Python. This will be my starting point for the practical aspects of natural language processing, regardless of whether I choose to use NLTK or another toolkit.