

Evaluating Word Usage Overtime in Parliament

Thomas Griffiths – TJG4

Supervisor – Praboda Rajapaksha

G401

CS39440

V1.1 - Draft



February 16, 2025

Project Outline Content Discussion

Document ID: MMP\_S02\_ProjectOutline\_Examples

Project Description

This project will involve looking at the records of debates in the houses of Parliament, specifically the House of Commons, on Hansard. Debates from the past, as early as 1803, will be gathered, processed and compared to debates of the present, in order to examine how word usage has changed over time, as well as how word usage can vary based on the region or party of an MP. Debates being processed will be stored in a database and processing will be done using NLTK.

The debates from the Hansard website will be manually collected, or a crawler that will take specific debates from a time period may be developed. This will depend on whichever will be more efficient timewise. For the debates, two specific time periods will be focused on. The first half of Regency Era, although starting from 1802, as this the furthest extent that the records go back on the website. This period will be contrasted to the last 20 years. Using a 20 year period should give a wide enough range to account for irregularities.

The software development approach taken towards the project will be the Agile approach. Due to a personal lack of experience with machine learning, it is likely that changes will need to be made throughout the project’s development. Therefore, the Agile approach is most appropriate.

Proposed Tasks

Research will need to be done into the software available for processing words, as NLTK seems appropriate, but it may turn out that there is software that is better suited to the tasks that need to be performed.

A decision will need to be made on which aspects of word usage should be investigated. Potential ideas include; have MPs become more or less polite over time? Which UK region has the funniest MPs? Can we predict the party of an MP based on a paragraph?

(Also; How are disease outbreaks handled and discussed?, How are different vaccines approved? Do organisations disagree about making them available to the public? Why?, Discussions about war after, during and before World Wars?, The sentiments of particular MPs before and after they become PM?, Who was the funniest PM? – need to ask about these)

The debates will need to be gathered so that they can then be processed. As mentioned, it will either be done by manually collecting them, or creating a crawler to gather them automatically. If a trawler is used, it will likely be the TACIT Hansard Crawler, which will gather the desired debates into a corpus.

Before the data can be processed, the raw data collected will be pre-processed into a format that can be used by the language processing software.

The next task will be to processes the data using NLTK. This task should be broken down into different parts.

Documentation will need to be produced detailing all the tools used. This documentation will be in the form of a text document with a link to each tool used and a brief explanation of how it will be used and its relevance to the project.

All files for the project will need to be put on a GitHub page. When a file is edited, I will need to upload it to the page. This will allow version control.

On the issues page of the GitHub an issue will be created for each task that will need to be done. Larger tasks will be broken down into smaller tasks to make it easier to track project progress.

When the project has been completed, a project report will need to be created that will describe how well the project went.

Project Deliverables

TACIT Corpus of Debates – The TACIT Crawler will gather the desired debates into a corpus so that they can be processed.

List(?) of Results – The results of the language processing. This might be a list assigning a different value to each factor that we are looking at, for example, if word usage has become politer we could use a high value.

Documentation – This will list all the tools used in this project, which should include TACIT and NLTK.

GitHub Project – The Project will be on GitHub to facilitate version control and to make it easier for others to view.

GitHub Issues Page – All potential tasks will be included and tracked as GitHub Issues.

Project Report – This will describe how the project went.

Bibliography

*NLTK*. Available at: https://www.nltk.org/ (Accessed: 13 February 2025).

Usc-Cssl, *USC-CSSL/tacit: We introduce tacit: An open-source text analysis, crawling and Interpretation Tool. Tacit’s plugin architecture has three main components: 1. Crawling Plugins 2. corpus management 3. analysis plugins. Tacit’s open-source plugin platform allows the architecture to easily adapt with the rapid developments text analysis.*, *GitHub*. Available at: https://github.com/USC-CSSL/TACIT (Accessed: 13 February 2025).