Background

**Hansard Dataset**

The Hansard Dataset is a set of documents produced by the British Parliament, which began in the 18th and 19th century. These documents contain reports and details of debates in the House of Commons, going back to the year 1803. Eventually, in 1907, these reports were made official and started being produced by Parliament itself, becoming “The Official Report”, though still unofficially known as Hansard. Along with becoming official, a report was officially defined as being one:

*“which, though not strictly verbatim, is substantially the*

*verbatim report, with repetitions and redundancies omitted and with obvious mistakes corrected, but which on the other hand leaves out nothing that adds to the meaning of the speech or illustrates the argument" (PUT REF HERE)*

Hansard is available in a variety of versions. The most commonly used and best known version is the Daily Hansard, which appears each morning and reports of the previous day’s proceedings.

One aspect of the Hansard dataset that needed to be planned for when designing the project was its lack of proper XML formatting. Though the data is presented in an Xml format, there were also HTML style tags within the bodies of text. In addition, each of the six series of data appeared to have slightly different formatting, and so anything used to read the data would have to be able to handle these differences.

**Natural Language Processing**

Natural Language Processing (NLP) is the process of getting a computer to read and understand written text. Computers are very good at dealing with numbers and performing complex calculations at high speed but are not as good at understand spoken or written language. Because of this, a large part of NLP is the act of processing the data, or text, to make it easier for the computer to understand and work with.

Sentiment Analysis, also known as Opinion Mining, is the process of learning and telling if a person is writing in a positive or negative manner, in relation to a topic, based just off the text they write. This can be difficult even for humans to manage due to the nuances of speech and presence of sarcasm, but even harder for computers.

Sentiment Analysis is often trained on data extracted from reviews of movies or products. The nature of reviews meaning they are more likely to express strong opinions, and thus make the training of the computer easier to do. From such reviews, a corpus of words can be compiled, with labels telling the computer if the presence of said word in a review usually means a positive sentiment, or a negative one.

In researching Natural Language Processing, the Natural Language Toolkit was discovered. This python module is a commonly used for Natural Language Processing, and provides methods and classes for many things, such as Named Entity Recognition, sentence splitting, and even some basic machine learning tactics for things like text classification, which could be used for the project.