Summarizing Political Discussion Forums

Shay Merchant merchants18@students.ecu.edu
Cason White whitecas18@students.ecu.edu
Dillon Roberts robertsd13@students.ecu.edu
Jacob Craiglow craiglowj16@students.ecu.edu

1 Introduction

The general audience for most government documents are politicians. This often leads to proposals, speeches, Senate and Congress bills being wrought with confusion for the everyday person. Besides the technical nature of these documents, the texts can be extremely lengthy, and deceptively titled. A couple examples of this deception, whether intended or not, is the Patriot Act and Citizen's United. The Patriot Act expanded government surveillance privileges, and Citizen's United lifted the cap on what corporations could donate to politicians. Our goal is to remove the confusion of these documents and make a program which can summarize these documents in a more digestible format.

2 Related work

Milad Moradi's group worked on a similar program. A link to his work can be found here: https://www.sciencedirect.com/science/article/pii/S1532046418302156?via%3Dihub Rather than being focused on political documents, Moradi chose to focus on biomedical texts. Like us, Moradi outlines how the challenge they had to overcome was the different diseases and subcategories of biomedical texts which can affect the way a text is formatted. Moradi's approach was to tackle summarization in 4 steps. Preprocessing, topic extraction, sentence clustering and summary generation. The pros of this approach is that it is concise and relatively simple to implement. The cons could be that a simple approach may take out some important language and context out of the original text to summarize it.

Anna Kazantseva, and Stan Szpakowicz also had similar work with summarizing short stories. Their work can be found here: https://www.aclweb.org/anthology/J10-1003.pdf. Their team took a unique approach, whereas they made sure their reader knew certain details about the literature they were reading. The pro to this choice is that you are ensuring that your reader has some level of understanding at the end of the summary, even if the summary is less than ideal. The con to this approach is that a reader may have to have an understanding of the document before it's summarizes, which could potentially contradict the point of a summary.

3 Your approach

The approach that we shall attempt is to apply an LDA algorithm to the corpus of daily United States Congressional records to produce a topic model. Previous implementation and research into the strengths of an LDA topic model have shown that it functions well for data where there are some known labels or topics to exist. We expect to be able to apply this algorithm and experiment with fine tuning of its parameters as well as the location and size of what we consider to be documents.

3.1 Milestones & Schedule

1. Acquire, analyze, and preprocess data (2 weeks)
2. Build LDA modeler (3 weeks)
3. Fine-tune , do an error analysis (3 weeks)
4. Develop Presentation (2 weeks)

4 Data

The data that this will run on will be transcripts from (the United States Congress). The LDA algorithm is an unsupervised learning algorithm and thus does not require a training set persay. However, there will be testing done on small increments of a day's transcript throughout the length of the project. This data is freely available for download from congress.gov and is the main focus of our work. Its contents are records of politicians discussing issues stored in PDF format. This can be parsed using pdfminer, a Python library.

5 Tools

We intend to apply and fine tune gensim's LDA topic modeler. For preprocessing we shall experiment with utilizing NLTK's stemming, lemmatization, stopword removal, and extracting the text itself with pdfminer.

References