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 employ will rely on an LDA algorithm we have researched. We will apply this algorithm to the corpus of daily United States Congressional records of House and Senate discussions. Using this algorithm, important topic models will be produced. We ultimately chose this algorithm because of it's a fairly straightforward approach to the issue of topic modeling, and its well researched strengths. Specifically, previous implementations of LDA topic models have shown to function well for data where there are some known labels or topics to exist within the corpus being used.
This is integral for our program, as the topic models play an important role in summarizing the data. 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 to produce a list of topics related to them for the average person to search. This will factor into our pre-processing stage, and will take up the bulk of our program. Once we have our LDA model down, and we are producing correct topic models, we will spend our time honing our interface to give the user the best summarization of the topic models.

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

When choosing our data set, we made sure that our data would be readily accessible and plentiful. A personal requirement for the data set, was our data had to be something that is relatable, something interesting, and something that could matter to a lot of people. The data that this will run on will be transcripts from congressional-record 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 will be using various tools to execute our program. Below is an outline of the tools we intend to utilize and what they will be responsible for.

5.1 Corpus Retrieval

We intend to utilize Python libraries to retrieve pdfs on demand from the live website: Congress.Gov. We specifically beleive that either the library "wget" or "requests" shall be appropriate for our needs. We believe, given additional opporunity for development, that we could utilize a database where the processed plain text portion of the pdfs could be stored.

5.2 Preprocessing

There are many libraries for assisting with preprocessing in the Python language. To strip the text out of pdf format we intend to utilize the "pdfminer" library, which also offers additional features which we believe will prove useful for understanding the structure of the corpus. Specifically, we intend to utilize the ability of pdfminer to recognize sections of pdf documents and extract headers in a way that we might leverage in our model.
After pdfminer has extracted our corpus, we intend to perform remainder of our preprocessing with the Python library "nltk". Nltk supports tokenization, stemming, lemmatization, among other useful preprocessing features.

5.3 LDA Application

The Python library "Gensim" offers many features for language modeling. We are interested in using its LDA function to retrieve a list of topics from our corpora. Since Gensim is an open source library we believe that, if given time and need, we could integrate techniques not currently part of its implementation such that it might be better fitted for our needs in modeling a brief summarization of our corpora.

5.4 Data Visualization

We believe that data visualization is an integral piece to understanding the output of something as complex as topic modeling. Thus, we intend to utilize "matplotlib", a Python library to tabulate our results. We also hope that we may be able to integrate the library "pyLDAvis" to create more visually interesting results.
If given enough time, we also hope to present our project as a usable application where a user may keep track of select topics that they may receive alerts on when they come up in the Congressional conversation. We believe that this would be a very useful tool for a reports on these conversations which are not often discussed or reported with minimal bias in modern American news media.

References