Project Report

Chronology of the project and member contributions:

Our project had multiple deliverables and a lot of moving components. It was a great team experience as everyone contributed as much as they could and genuinely cared about the project. The project idea and proposal were far from conventional; it was truly refreshing to work on a project that was so different from other projects that we had to work on as a part of our studies. The main purpose of this project was to create an NLP model that would extract the agenda, sentiment, and ideas from a political speech after summarizing it. This model and approach can have widespread adoption as a journalistic tool to objectively analyze the speech of a politician that is free from most human-centered biases. The biases presented by a news-media channel have been a topic of extreme polarization in the current times. While our approach is not completely devoid of any biases and is open to exploitation by data manipulation techniques, it does roll the ball a bit further towards accountability from our elected leaders and hopefully, serves as a useful tool to communicate the main agendas from seasoned-yet-garbled speeches from politicians; our project has the potential to cut through the ambiguity and help save the labor time of journalists as well as help them objectively analyze speeches by political candidates.

In the initial phases of the project, most of the work done was in terms of project ideation and dataset generation. A lot of time was spent gathering resources and consulting reference materials to come up with a successful approach for implementing the project idea. In the stages of the project proposal, the prototype implementation was split into 3; with each team member building upon the work done by other team members. It helped us create and maintain a clear division of labor and ensured that everybody's time was respected. Shahbaz handled the acquisition of the materials to be used for dataset generation and parsing. Gaurav handled the preprocessing and dataset attribute documentation while Jay tackled the implementation and generation of the word-cloud generation component of the prototype. Everybody worked in a highly cohesive manner which ensured the completion of the prototype model within the time constraints that we were given.

Gaurav Shinde took the lead after the project prototype stage and played a huge role in ensuring that we met regularly to work on the project and spent the time required to work on this project. He was very hands-on with the management of the project and took the initiative to implement a lot of the components whenever he had the time to do so. Gaurav went above and beyond in the implementation phase and worked with Shahbaz for consulting reference materials and understanding the core concepts that we used to implement our model.

The dataset generation using web links was handled equally by everyone on the team. Most of the dataset link acquisition and parsing was handled by Gaurav, followed by Jay. Shahbaz contributed the least amount of links for this part of the project but diversified the dataset acquisition from a web-domain perspective. It is due to this diversification that the overhead time associated with dataset generation was inflated and took longer than expected. There was a lot of pre-processing involved due to each web domain requiring a hand-tailored edit for extracting relevant information; most of this stemmed from the different ways text data is structured on different web pages and web-domains. Shahbaz worked on generating the data visualizations for our project presentation using the data provided by Gaurav and Jay.

Gaurav implemented virtually all of the project core components while Jay and Shahbaz helped with debugging. Some additional components such as the streamlined data-pipeline functions and the implementation of the subset approach were implemented by Shahbaz and Jay. The initial model was run and evaluated by Shahbaz and Gaurav while Jay worked with Shahbaz for the evaluation and execution of the model once the speaker-subset approach and summarized approach were implemented.

Everybody was actively involved in the development of the presentation slides. The work was divided based on the presentation documentation provided on the course canvas page. Gaurav chose to work on the first few slide objectives, followed by Shahbaz and Jay. The team met regularly and worked on polishing the project code and presentation slides almost every other day in the week leading up to the final presentation. The model implementation and evaluation had been set back by a couple of days due to some of the team members falling sick so everybody came together to put in the extra hours and effort that was required for ensuring the successful implementation of the project.

PROBLEM STATEMENT:

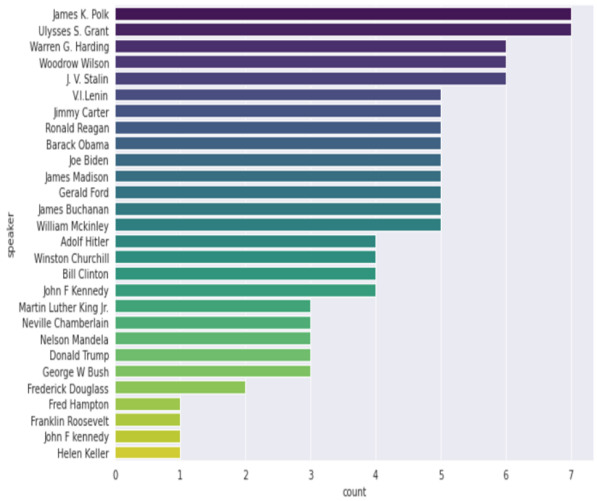
People lie all the time. While most of us lie about small or embarrassing things, or to avoid uncomfortable situations, some people lie in their professions to either save face or to evade accountability. The lies of the latter kind have far-reaching consequences and impact society negatively. Despite the establishment of republics throughout the world, we have one problem plaguing virtually all political systems, i.e., lies and a lack of confidence in their representatives. This is a problem across the board and is fodder for most news publications. We are going to attempt to summarize speeches by political figures to ascertain their agenda. The scope of this project, due to time constraints, is limited to this. The larger vision of this idea is to be able to visualize and quantify the agenda of a political collective/platform/party. Through this, the voters can be better informed about who and exactly what they are voting for. The root idea is to have metrics to analyze and ensure better governance for and by the people. The extended (and perhaps, utopian) application of this approach could/would be the association of confidence metrics to political figures running for elections which quantifies the gap between “what they say” and “what they do”.

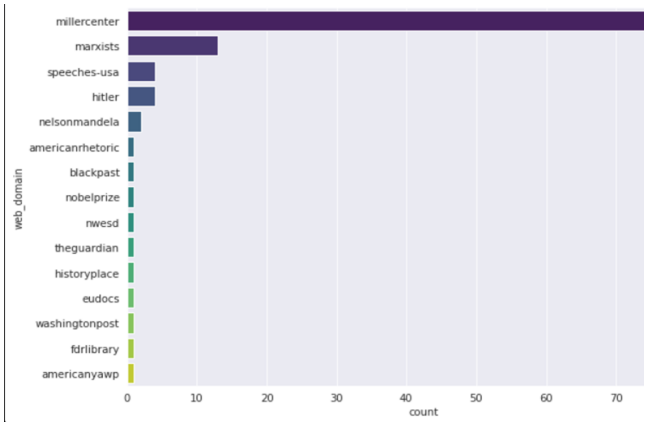
DATASET DESCRIPTION:

Number of words in corpus: 17177

Length of sentences:15880

Shape : (113,9)





NLP MODELS & TECHNIQUES UTILIZED:

Our project had 3 major components. Namely they are:

1. Text Scraper + Parser

   This component is technically 3 sub components as all members of the team searched for their own links to generate the dataset and wrote their own code to scrape and parse data from their listed links. A list of the URLs that we used has been provided in the "data_fix.txt" file. This component is directly connected with the text summarizer component for getting processed. The text scraper and parser utilized beautiful soup and selenium chrome drivers to scrape and parse data from "http" format requests.

2. Text Summarizer

   This was one of the core components of our project. It utilized spacy libraries to summarize the text. The code would receive text from the dataframe containing the parsed text, remove stopwords and punctuation and tag the words as parts of speech. After these preprocessing steps, it would pass the data through a frequency counter which would try to determine and generate the importance of words. The frequency counter would assign a "strength" score to each sentence. It would then filter out sentences with weak strength scores and keep the stronger strength scores as the summary. This data was then passed on to the Opinion miner component of our project.

3. Opinion Miner

   The opinion miner would re-process the data received from the summary since our project was supposed to deal with summarized speech data. The Opinion miner pipeline would first tokenize, and then lemmatize the words in each summarized speech field that it received. The component would then vectorize the words using a count vectorizer before passing it to the latent dirichlet allocation (LDA) model. The LDA model would extract details about dominant topics from the vectorized data and assign a relevancy score to them. Finally, the subcomponent would extract the sentiment and topic name from the vectors and relevancy scores given by the LDA model.

EXPERIMENTS CONDUCTED:

We implemented 3 different approaches for our project. With each one serving as a project checkpoint for our model. The first experiment, i.e. the baseline approach, was to run the entire speech dataset without a generated summary to efficiently debug the code as well as to gain insights into our model's functioning. The second experiment was one of the core objectives of our model, i.e. running the model on non-summarized speeches from each speaker. This was done by selecting a subset of speech based on the speaker label and passing it through our model pipeline. This served as a good benchmark to establish the working of our opinion miner component and for testing how good/whether our text summarizer performed. The last experiment that we conducted was running the speaker subset data after it is processed through the text-summarizer. This was the final benchmark and objective for our project. We could clearly see the changes that the text-summarizer caused to the speeches of some speakers and also noticed a lower processing time for the opinion miner component as it had to process fewer tokens.

INSIGHTS:

The project was very complex and utilized the implementation of multiple NLP models in a data pipeline. We could clearly see that summarizing the data had an impact on the dominant topics that were selected in the end and their sentiment scores. Since the Opinion mining technique was an unsupervised learning technique, we were expecting some errors such as the labeling of the token(s) "let us" from the speeches of Frederick Douglass. While this is clearly not a topic in itself and an error in our model, we were surprised at the number of errors that our model generated in the form of topics. We were expecting a lot more issues because it is an unsupervised learning technique but were glad to see that it was functioning as intended in most cases. The speech dataset contained a disproportionate amount of speeches by U.S. presidents so there was a prevalence of some terms throughout our entire corpus for each speaker. This could be seen when we analyzed the entire corpus post-model-processing as well as in the speaker-subset approach. One example of this was the classification of "United States" as a topic. While it was a general theme in most of their speeches, it is unclear when and where the "United States" was talking about the people of the United States, the geographic land mass of the United States and the nation of the United States. The last observation was that we needed more data for most of the speakers in order to analyze their speeches properly. Some speakers such as Nelson Mandela and Frederick Douglass had very few speeches in our dataset and therefore, our model did not produce desirable results for their speeches. The model only generated a couple of topics for them as compared to the 5 it generated for speakers with more speeches.

RESULTS:

1. Experiment on Entire Corpus - Opinion Based Mining

   We have iterated the topic modeling unsupervised learning for opinion-based mining. The latent Dirichlet Allocation (LDA) model was incorporated to create the topics based on the text speeches we got from the different politicians. The model parameters for LDA are detailed in the image below.

```
# initisalise LDA Model

lda_model = LatentDirichletAllocation(n_components = 5, # number of topics
                                      random_state = 10,          # random state
                                      evaluate_every = -1,      # compute perplexity every n iters, default: Don't
                                      n_jobs = -1,              # Use all available CPU'S
                                     )
```

We adapted this concept to the text corpus's representation of a bag of words. We used a CountVectorizer from the scikit-learn module for the bag-of-words representation. The depiction of the bag of words is a bigram.

The outcome of opinion-based mining on the speeches' whole text corpus is shown below.

	Dominant_topic	topic_name
0	3	[mr ford]
1	6	[senator kennedy]
2	4	[united states]
3	2	[great masses]
4	5	[health care]

sentiment	Negative	Positive
topic_name		
great masses	7	17
health care	5	19
mr ford	10	14
united states	5	15

The way politicians narrate their speeches causes several inaccuracies that have an impact on our text corpus results. "United States" is one of the terms in the dominant subject for two reasons. The first explanation is the manner that politicians describe their speeches, and the second is that a larger share of the data in our dataset relates to US presidents. We need additional information from the many politicians from various locations in order to enhance these outcomes.

The dominating themes were used to extract the right side (image) of the findings. It is the proportion of both the positive and negative talks out of 113 that have a dominant topic with a high relevance score.

2. Experiment on Speaker-by-Speaker Aspect - Opinion Based Mining
We developed a pipeline to conduct the complete speaker-by-speaker opinion mining method after obtaining the necessary findings across the entire corpus.

To obtain the chunks for each politician's speeches, we have divided the complete corpus according to the "speaker" column. We have now used our opinion miner to extract the dominant topics and speeches that these temporarily produced data frames are linked with.

There are 28 speakers in the dataset; we have attached a few of the speaker's results for opinion-based mining.

Speaker: James K. Polk

	Dominant_topic	topic_name
0	6	[missouri compromise]
1	3	[united states]
2	4	[laws nations]
3	5	[new mexico]

sentiment	Negative	Positive
topic_name		
laws nations	1.0	2.0
new mexico	NaN	1.0
united states	1.0	1.0

Speaker: Ulysses S. Grant

	Dominant_topic	topic_name
0	5	[within territory]
1	6	[united states]
2	3	[dead bodies]
3	4	[combinations conspiracies]

sentiment	Negative	Positive
topic_name		
combinations conspiracies	1.0	1.0
dead bodies	NaN	2.0
within territory	NaN	2.0

CONCLUSION:

In a nutshell, we collected all the speeches by carefully reading them all (particularly to the war genre) and then produced a link archive for each speaker. To collect the raw text speeches, each team member scraped the web in their own way. The dataset was then integrated based on three characteristics (links, speakers, and text speeches). Removed all punctuation, stop words, and unused hyperlinks from the raw text before converting the entire speech corpus to lowercase. Utilized the Latent Dirichlet Allocation (LDA) model for topic modeling to extract the dominant topic and relevance score.Conducted the experiments on the entire corpus and iterated the pipeline for the speaker-by-speaker aspect-based opinion mining.

FUTURE PLANS:

We can get additional information from many websites to use for our future actions. Additionally, we may add extra information to the data frame, such as the timeframe during which the speech was delivered and the political affiliation party with which the leader was actively associated and we can ensure that the dataset has fair distribution. We can achieve much more accurate findings with the data we now have if we add more of it and have much clearer insight into it. Having said that, the performance of the model is entirely contingent on the dataset used in our study.