Natural Language Processing to Analyze Mayor Speeches And Identify Emerging Local Policy Trends

Cristina Giraldo[1], Kanchan Ghimire[2], Renzo Castagnino[3], Taisha Ferguson[4]
Columbian College of Arts & Sciences, The George Washington University, Washington D.C.

[1]cristinagiraldo@gwu.edu

[2]kghimire92@gwu.edu

[3]rcastagnino@gwu.edu

[4]tferguson14@gwu.edu

Abstract

The purpose of this project is to use Natural Language Processing (NLP) techniques to analyze mayors' speeches from cities around the United States to identify local policy trends. The goal would be to use topic modeling to identify policy topics in speeches. The first objective would be to create a topic model to represent the distribution of topics across the full dataset. Once the overall distribution of policy topics is obtained, it will be further segmented by year, region, and population to analyze the similarities and differences in the distribution of the topics. This project idea emerged from an annual report called "States of the Cities" from the National League of Cities (NLC). Latent Dirichlet Allocation (LDA) topic modeling will be used in order to identify topics and major themes in mayors' speeches across the United States.

Introduction

The purpose of this project is to use Natural Language Processing (NLP) techniques to analyze mayors' speeches from cities around the United States to identify local policy trends. The goal would be to use Latent Dirichlet Allocation (LDA) topic modeling to identify policy topics in mayor's speeches. The dataset consists of about 150 speeches per year and is between the years 2016-2019. The first objective of this project is to use LDA to create a topic model to represent the distribution of topics across the full dataset. Once the overall distribution of policy topics is obtained, it will be further segmented by year and region and the similarities and differences in the distribution of the topics will be analyzed.

Background

This project idea emerged from an annual report called "States of the Cities" from the National League of Cities (NLC). Two of our group members, Cristina and Taisha, are employees at the NLC. For the State of the Cities report, NLC staff members have been manually tagging policy topics in mayor's speeches to analyze emerging policy trends. We intend to help improve the efficiency of the NLC staff's tagging process with the use of NLP topic models.

Data

It is worth mentioning that the speeches comprising the State of the City report are speeches given once a year by the local mayors of many cities in the United States. In these documents, mayors detail problems that are shaped by the local realities, as well as their vision and willingness to solve these problems. For the purpose of this project, we considered 150 speeches per year, between 2016 and 2019. These documents come in two formats: WORD and PDF files. Ultimately, these documents don't have a uniform structure or a standard format (for instance, a document can have purely a text content, while others might include pictures).

Methods

This multi-staged process begins by extracting each of the documents from a local directory folder within the National League of Cities Database. The first step is to iterate over each of the sub-directories in the Database and identify whether a document is a PDF or WORD file. The next step is to correctly extract the text information from each document. To achieve this task, we used two Python libraries: Docx and PyPDF for WORD and PDF files, respectively. In the case of the WORD files, we did not evidence any inconvenience and successfully extracted all the texts from each document. However, for the PDF files, the process was somewhat troublesome. While some files could be read without any problem, others were simply read as empty files or with strange symbols and characters. Approximately 70% of these documents were successfully read and, for the purpose of the project, we excluded those unreadable PDF files. It is noteworthy that we also used the library PDFMiner to perform the text extraction for the unreadable PDF files, but the results remained the same. None of these PDF files were encrypted.

The next stage was to identify the city, state, year and type of each file. Taking into account that each file does not have a standard form, the best way to identify this information was extracting it from the file name. As all the files are saved with the same format, and considering that this will not change in the future, Regular Expression techniques were used to extract the information from each file. Subsequently, the information was saved on a Dataframe table, which each column specifies the following information: city, state, year, file name and the speech.

Once we had our Dataframe ready, further preprocessing needed to be implemented before the modeling process. First, we iterated over each speech and applied Tokenization to convert the text into individual words. Additionally, during this step, we also removed punctuation and words that had less than 2 characters. Thereafter, we removed the Stop Words and filtered the entities. Finally, we created our own custom dictionary of irrelevant words that we considered necessary to extract to have a better accuracy in our model.

Model

As stated in the Introduction, we intend to use Latent Dirichlet Allocation (LDA) topic modeling in order to identify topics and major themes in mayors' speeches across the United States. LDA topics modeling is an NLP modeling technique that represents documents as probability distributions of words. For LDA topic modeling, a set number of topics is given and a mathematical algorithm is used to determine the probability distribution of words given the number of topics. Once the model determines the number of topics, the user reviews the words associated with the topic to determine the category to which the words belong. In our project we used Gensim, a computer programming package, to implement our LDA topic model. Within Gensim's LDA package there are a set of parameters that must be selected, monitored, and refined in order to get the most meaningful and coherent topics. After testing and evaluating our different topic outputs we decided the following parameters worked best for our purposes: 100 topics, 200 interactions, and 300 epochs.

Interface

A web application was created using Dash framework. The goal of the interface is to assist the NLC staff members identify important topics (policies). The idea is to show the process of Topic Modeling along the way so that the user can see what is going on in the backend. The functionality in the web application is broken down by tabs: File Upload, Data Preprocessing and Modeling. The File Upload functionality allows the user to upload either a WORD or PDF file. Similarly, the Data Preprocessing functionality allows the user to do pre-processing (lemmatization, stop word removal and entity removal) on the uploaded files. Finally, the Modeling functionality allows the user to run LDA Model on the uploaded files.

The LDA Model run on the data performs Topic Modeling, which returns a list of words for each topic. The results from the model help ease the burden of manual tagging. For better visualization of the results, graphs are displayed on the interface. PyLDAVis library is used to graph the results in a form of bubble graph, where each bubble represents a topic idea, and a bar chart that displays the list of words for those topic ideas. Similarly, scatter plots are created using plotly to

show topic distribution by year and region. The scatter plots show how the policies have changed over time and also how it varies across different regions.

Results

After running our model with our final parameters selection we received a coherence score of 43%. A coherence score is an accuracy measure within the Gensim program that measures the coherence of topics in a topic model. The coherence score is a score that takes into account the number of topics and the segmentation of words within and between topics. A coherence score of 43% is a strong indicator that there is significant room for improvement with our topic model. Another test for coherence is looking at the topics themselves to determine if they are meaningful and representative of the documents themselves. After reviewing our topics, about 30% were very coherent, 30% were moderately coherent, and 30% were unclear. Attached as Appendix A to this report is a table that shows examples of topics that fall within these categories. The topics in the high coherent topic section were easily labeled and the documents that were in the moderately coherent section were labeled after reviewing documents and comparing with human-taggings. The topics that fall into the unclear section were topics that were repeated and too vague to label.

When the model was run for each year, we observed that the years from 2016 to 2019 that similar distributions of topics. However, certain differences did appear both (i) in the presence and absence of certain topics and (ii)in the frequency of such topics. An example of the former is that topic of pedestrians appeared in 2019 but not in 2016. An example of the latter is that the topic of infrastructure appeared at a far greater rate in 2019 than in 2016.

Similar results were achieved when we ran models per geographic region. While the distribution of topics was quite similar, there were also some noticeable regional differences. For example, the topic of education was much more common in the Southern region than in the Northeast.

Discussion

We believe that our results were meaningful but our model still needs more fine tuning in order to be more useful to the NLC staff. After comparing our results with the human tagging, we realized we were able to capture about 50% of the tagging list. That 50% did not always match directly with their topics and often fell between the major topic level and the subtopic level. Going forward, we would like to try Hierarchical Topic Modeling to see if we are able to capture more of the nuances of speech.

Conclusion

This project showed that topic modeling holds promise in helping to ease the significant burdens associated with manual tagging of mayoral speeches. Nevertheless, our LDA topic modeling only helped us to achieve a score coherence of 43%, which we believe is not enough to ease the burden of manual tagging.

To increase precision, we think that might be useful to have a bigger base document and increase the number of epochs to obtain subtopics that are better related to each other. In addition, we also would like to try other modelling techniques, such as Hierarchical Topic Modeling, and perhaps include Deep Learning techniques.

References

Bird, S. (2016). *Natural language processing with python*. Place of publication not identified: OReilly Media.

Grinberg, M. (2018). *Flask web development developing web applications with Python*. Beijing: OReilly.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. India: Dorling Kindersley Pvt, Ltd.

Lane, H., Howard, C., Hapke, H. M., & Griffioen, A. (2019). *Natural language processing in action understanding, analyzing, and generating text with Python*. Shelter Island, NY: Manning Publications Company.

State of the Cities 2019. (2019, May 24). Retrieved from https://www.nlc.org/resource/state-of-the-cities-2019