Modeling Political Influence From Media Appearances

Chase Diaz

Problem Description

The goal of this project is to model the influences shaping a given local-level politicians' legislative agenda from their media appearances. To accomplish this, Python code was developed to scrape YouTube search results for a given politician and extract meaningful information from each media appearance. **This includes highlighting the outlets they are most frequently appearing on, the topics and legislation they are frequently discussing, and describing the extent of their reach**. This scraper enables users to automate the data collection to quickly gain these insights for any politician, which would otherwise be extremely time-consuming to complete manually.

To tackle this problem appropriately, I first conducted manual analysis focused on three local-level politicians before attempting to automate the solution: Kathy J. Byron of Virginia, Liz Krueger of New York, and Ana Maria Rodriguez of Florida. This provided background knowledge as to which outlets these politicians were most frequently making appearances, which businesses and sectors had impacts in their districts, and an overall story as to what their influences are and why. The manual analysis was extremely helpful because it gave the ability to verify that the data retrieved from the scraper was accurately capturing the influences of a given politician.

Methodology

YouTube was chosen as the primary focus for capturing media appearances since it's the most comprehensive data source for videos. In order to model influence, I felt that for any given politician we should be able to identify the outlets or channels they are most frequently appearing on, the topics they are most frequently discussing, and to identify appearances where climate legislation or topics are discussed. The YouTube scraper automates several tasks in just a few minutes that would otherwise be extremely time consuming to manually complete:

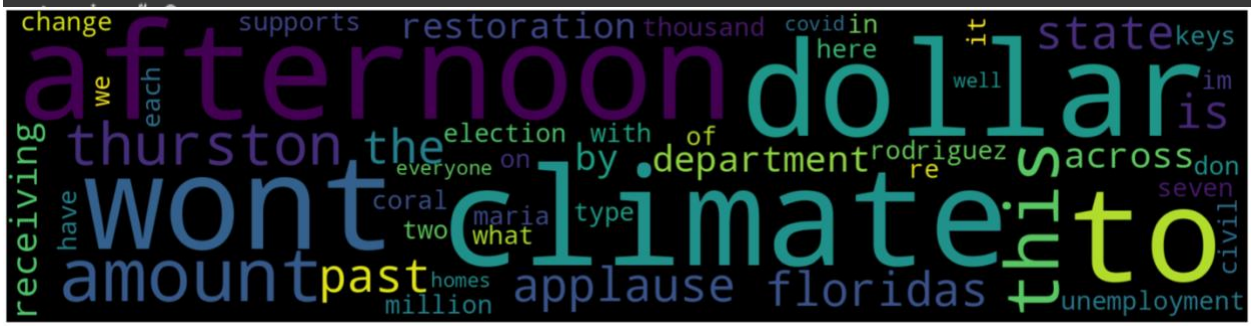
1. **Retrieving media appearances:** For a given set of politicians, retrieve the top 25 search results for each politician. The scraper returns a DataFrame where each row contains a single video search result with data for each video and the channel it was posted from (id's, title, description, and statistics). This piece provides the ability to provide a dataset of media outlet appearance for a given politician.
2. **Identifying unknown influences:** By downloading and parsing the captions from each video (where available), we can identify the most common entities discussed. The most frequently discussed topics are political influences we want to capture. To accomplish this, I used the spacy library to implement a Named Entity Recognition (NER) model to recognize frequently discussed entities in the captions of videos for a single politician.
3. **Searching for specific influences:** Based on a list of keywords given as a parameter, this dataset tells the user which videos contain the given keywords. This is useful because users can quickly identify which videos mention certain topics or legislation of interest.

Additionally, I have provided users the ability to quickly generate insights from data the scraper has collected.

1. **Video + Channel Stats:** the number of times a politician has appeared on each YouTube channel. This allows users to quickly identify which channels a politician is appearing on the most, and how many views and subscribers that channel gets. I also provided a function that will generate two plots. The first visualizes total views for each of the channels they appear on, and another visualizing the number of views for the top 10 most viewed videos from the search results.
2. **Topic Model:** Using the Top2Vec Python library, the data scraped from YouTube was implemented into a topic model, an automated clustering of text into specific topics. To generate the topic model for a single politician, captions from each YouTube video were provided into the model as training data. The model then clusters words and phrases used in similar ways to specify topics, and returns a list of topics and associated keywords. For each topic, it lists the words that were frequently appearing and used in the same context.
3. **Topic Model Word Cloud**: A visual of the popular words for a topic recognized by our topic model. The words used more frequently and in similar contexts will be visualized larger than other words in the visual. The larger the word in the visual the more it contributed to specifying the topic.

	politician	count of videos on channel	channel_title	video_view_count	channel_subs	total_channel_views
0	florida senator ana maria rodriguez	5	CBS Miami	5293.0	500000	476253814
1	florida senator ana maria rodriguez	4	WPLG Local 10	1582.0	645000	691577884
2	florida senator ana maria rodriguez	3	Elect Ana Maria Rodriguez	1478.0	8	1478
3	florida senator ana maria rodriguez	2	EWTN	0.0	802000	284524242
4	florida senator ana maria rodriguez	1	American Conservation Coalition Action	52.0	492	16545
5	florida senator ana maria rodriguez	1	Avivamiento	437435.0	703000	318275549

```
topic # 1
Words: ['climate' 'afternoon' 'dollar' 'wont' 'to' 'amount' 'this' 'thurston'
 'applause' 'floridas' 'is' 'state' 'the' 'past' 'department' 'across'
 'receiving' 'by' 'restoration' 'unemployment' 'it' 'rodriguez' 'maria'
 'thousand' 'supports' 'coral' 'in' 'on' 'each' 'million' 'keys' 'im'
 'type' 'we' 'here' 're' 'have' 'seven' 'two' 'election' 'what' 'change'
 'with' 'don' 'civil' 'of' 'everyone' 'covid' 'homes' 'well']
```



The following datasets produced from the scraper give insight into the influences each politician considers when choosing their media outlets. The scraper can retrieve and transform this data

into a cleaned dataset in minutes, whereas manually this effort would take a considerable amount of time and effort.

1. Search Results: a dataset of the top 25 YouTube search results for each of the 10 politicians. For each video, the scraper provides the title of the video, description, name of the channel, video and channel statistics such as views and subscribers, and a link to the video. This can allow for further investigation into each politician's media appearances. A random sample of this dataset is shown below.

	video_title	description	publish_date	channel_title	channel_id
14	2020 Virginia Chamber Legislative Awards Program	The Virginia Chamber of Commerce recognizes th...	2020-12-04T18:07:30Z	Virginia Chamber of Commerce	UCvl-O1JjFsdvqr0UKD2efTA
20	Nebraska Legislature advances Let Them Grow Ac...	Nebraska lawmakers advanced the bill that woul...	2023-03-23T21:24:35Z	KETV NewsWatch 7	UCczdililpYSmiF074WMCNiA
15	NYS Senator Krueger on State Parks Legislatio...	May 28, 2010: New York State Senator Liz Krueg...	2010-05-28T20:39:10Z	NYSenate	UCVJzgJ1u9YANbvHDHryZO2g

	politician	data_source	video_id	video_view_count	comment_count	video_link	total_channel_views	channel_subs
	virginia delegate kathy byron	YouTube	MUFNAyFSv_4	66.0	0	https://www.youtube.com/watch?v=MUFNAyFSv_4	26345	112
	nebraska senator jen day	YouTube	PaB8oPuJVqE	NaN	NaN	https://www.youtube.com/watch?v=PaB8oPuJVqE	77167108	70200
	new york senator liz krueger	YouTube	s0ulaZB9iWI	21.0	0	https://www.youtube.com/watch?v=s0ulaZB9iWI	2774082	4530

2. Keywords Found: Using specific keywords such as "climate", "emissions", and "renewable energy" users can identify keywords appearing in each YouTube video. Example shown below.

	video link	keywords found	politician
45	https://www.youtube.com/watch?v=Bs6fxIRM5xw	agriculture, energy	florida senator ana maria rodriguez
62	https://www.youtube.com/watch?v=MUFNAyFSv_4	renewable, agriculture, energy, climate, envir...	virginia delegate kathy byron
23	https://www.youtube.com/watch?v=P5EWxMk3mMI	pollution, environment	texas chairman jon niermann
51	https://www.youtube.com/watch?v=4O4LNc0dV_M	climate change, climate	massachusetts senator karen spilka
1	https://www.youtube.com/watch?v=LHfxyVDWSvI	environment	alabama representative scott stadthagen
74	https://www.youtube.com/watch?v=Ov4mC7HxqgU	climate change, pollution, agriculture, climate	kansas representative ponka-we victors
29	https://www.youtube.com/watch?v=yoVQikgcW9Y	nuclear, pollution, agriculture, energy, envir...	texas chairman jon niermann

3. Entities: A few of the commonly recognized entities the NER model produced from our search results from 10 randomly chosen politicians were "exxon" by Texas Chairman Jon Nierman, "marijuana" by Ohio Representative Jason Stephens and "lb147" (Nebraska legislation) by Nebraska Senator Jen Day. These frequently recognized entities indicate influences commonly discussed in their appearances. Top recognized entities are shown below.

	count	politician
exxon	33	texas chairman jon niermann
jane seu	32	nebraska senator jen day
marijuana	25	ohio representative jason stephens
joseph	25	florida senator ana maria rodriguez
lb147	22	nebraska senator jen day

Discussion of Results

From the 3 chosen politicians used as manual analysis for a baseline comparison against the results of the YouTube scraper, there is a pattern as to how each politician chooses to make appearances on media outlets. The most common outlet for making appearances are local news networks. This makes sense because many of their constituents will be exposed to their local networks. The local news networks YouTube channels also tended to have the highest number of views and subscribers across the different channels our politicians would appear on.

In the case of one of the politicians, Florida Senator Ana Maria Rodriguez, an investigation into her influences can be conducted from the data we've gathered from our YouTube scraper. She has appeared on Hispanic Catholic networks such as Eternal Word Television Network and Avivamiento. Senator Rodriguez represents a political district in South Florida that includes the Florida Keys, where the majority of her constituents are Hispanic and hold Catholic religious views. The generated topic model confirms these potential influences by identifying keywords that are relevant to this demographic such as "abortions", "unborn", and "faith" clustered into a single topic.

In addition, the topic model generated a second topic with environmental keywords such as "climate", "coral" and "restoration", aligning with the interests of her constituents who rely on a healthy environment for the local economy. These topics are frequently discussed in her interview appearances. The Named Entity Recognition (NER) model further confirms the influences on Senator Rodriguez, as commonly recognized entities were "jesus", "miami", "cruise", and "latin america". These identified influences through our manual analysis are confirmed by our automated solution, which can accurately capture and model her influences.

This political media outlet scraper enables users to gain valuable insights into politicians' influences and agendas. The results from the 3 chosen politicians not only confirmed influences identified through our manual analysis, but also revealed previously unknown influences. This leads me to believe the data generated from the scraper is significant, since it can yield more information than a manual analysis, while spending a significantly smaller amount of time to do so.