Sana Khan
Homework 8

Introduction

The internet contains an unimaginable amount of data, in all forms. From social media where users can post pictures, videos, and comments, to blogs and newspaper, we are surrounded by information. Everything in our lives can be recorded and posted on the internet to anyone else to view. With so much information available at all times, it can be overwhelming and difficult to make sense of it.

Having this much information at the tip of our fingers, how can brands or companies get our attention. They have to complete to get our attention or our business and need to have short headlines or catchy visuals to do so. With everything needing to be as short as possible to quickly achieve this goal, how can this be done?

Thankfully, there is an automated way to begin to understand large bodies of text, through topic modeling. Topic modeling works by discovering, identifying, and categorize the main themes or topics within a large collection of texts. Topic modeling is useful not only because it can organize and summarize large amounts of texts but can also help with improving information retrieval and content recommendation. This is especially helpful to consumers who may be searching for a movie or book and can find an accurate recommendation.

Analysis - Multinomial Naïve Bayes and SVM

Data Prep

There were four separate files that contained the data to be used for topic modeling. The folders represent female and male, republican and democratic house representatives that contain the floor debate of the 110th Congress. Each folder has multiple text files that represent the debate from a unique speaker.

To begin, each folders contents were downloaded and then loaded in based on the file path. Then each document was preprocessed by removing unwanted characters and stop words. Each document was then stored in a new variable which was then used in count vectorizer. Once the data was vectorized, it was added to a new data frame.

	act	administration	amendment	america	american	americans	believe	\
0	140	24	12	95	200	75	28	
1	206	22	81	57	101	38	38	
2	203	75	100	60	82	46	61	
3	64	8	11	10	41	17	10	
4	59	4	16	2	11	12	2	
..	...	...	...	...	...	...	...	
424	69	16	20	10	30	18	14	
425	77	5	50	14	48	17	8	
426	53	6	26	50	98	23	24	
427	81	12	74	60	88	55	61	
428	43	6	10	12	24	16	15	

	billion	budget	care	...	vote	want	war	way	work	working	world	\
0	20	12	186	...	44	126	33	32	94	65	29	
1	12	6	24	...	32	49	79	46	72	31	161	
2	65	46	50	...	39	118	243	64	139	50	31	
3	0	4	25	...	14	38	45	17	42	9	24	
4	5	3	2	...	13	16	23	3	25	8	7	
..	...	...	...	...	...	...	...	...	...	...	...	
424	17	2	7	...	6	25	14	28	20	6	29	
425	22	9	3	...	22	39	8	51	23	2	13	
426	32	33	20	...	24	36	24	55	51	27	29	
427	14	25	39	...	27	122	77	23	21	14	31	
428	12	2	10	...	23	27	55	17	22	8	28	

Some additional cleaning was done on the data frame, to remove any numbers or words with a length of less than two. The new data frame was then used the model to predict the topic.

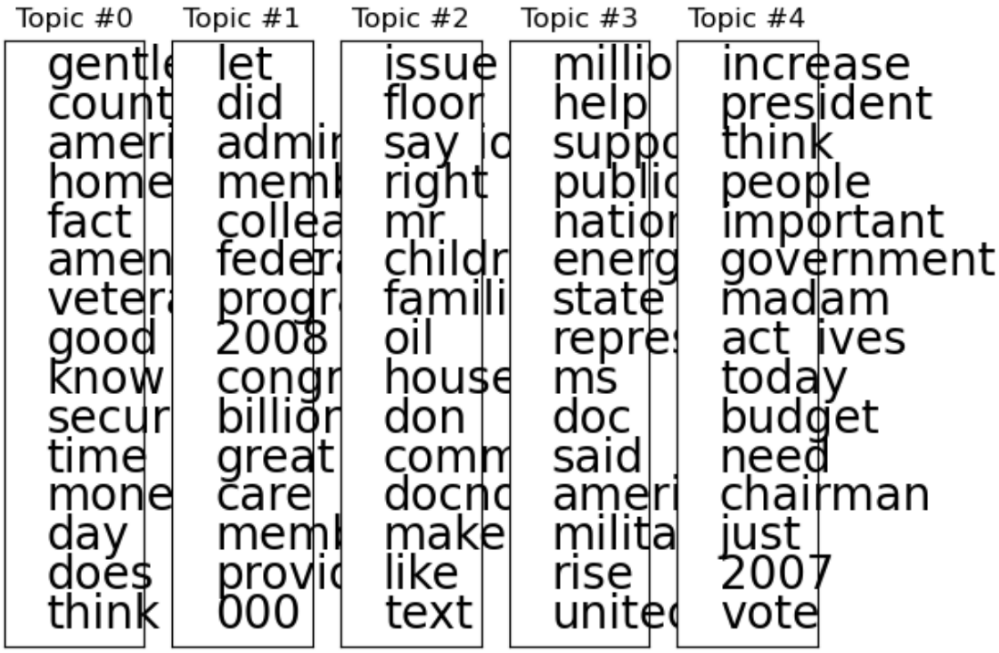
Model/Method

Latent Dirichlet Allocation (LDA) was the model used for this topic modeling problem. LDA is an unsupervised machine learning model, since we do not have a label that informs us of the topic. LDA has “assumptions” that topics will have similar words and it uses the grouping of those words to predict the topic. It works by taking the probability of a word belonging to a topic. The output it shows how many words appear for each topic in your text.

To use LDA, count vectorization is still performed and a new instance of LDA is instantiated on the vectorized data frame. This instance has 5 topics selected as a parameter to begin with. The following probabilities show the probability that a document is about a specific topic. The highest probability per row indicate which topic cluster the document is part of.

```
List of prob:
[[0.2        0.2        0.2        0.2        0.2       ]
 [0.1000267  0.5999045  0.10002375 0.10002311 0.10002193]
 [0.10002806 0.10002438 0.10002507 0.10002435 0.59989814]
 [0.10002679 0.59990448 0.10002376 0.10002308 0.10002189]
 [0.1000282  0.10002437 0.10002506 0.10002444 0.59989793]
 [0.1000268  0.59990426 0.10002379 0.10002316 0.100022  ]
 [0.59992059 0.10001993 0.10002056 0.10001997 0.10001895]
 [0.10002676 0.10002314 0.10002369 0.59990457 0.10002185]
 [0.10002603 0.10002244 0.59990777 0.10002252 0.10002124]
 [0.5999207  0.10001999 0.1000205  0.10001991 0.10001891]
 [0.10002675 0.1000231  0.10002372 0.59990455 0.10002188]
 [0.10002672 0.59990451 0.10002373 0.10002314 0.1000219 ]
 [0.10002823 0.1000244  0.10002512 0.10002439 0.59989785]
 [0.10002684 0.59990437 0.10002375 0.10002305 0.10002199]
 [0.10002818 0.10002436 0.1000251  0.10002439 0.59989797]
 [0.10002604 0.10002249 0.59990763 0.10002251 0.10002134]
 [0.10002676 0.59990446 0.10002375 0.1000231  0.10002193]
 [0.10002683 0.10002306 0.10002373 0.59990449 0.10002189]
 [0.10002597 0.10002245 0.5999078  0.10002246 0.10002132]
 [0.2        0.2        0.2        0.2        0.2       ]
 [0.10002675 0.59990445 0.10002378 0.10002315 0.10002188]
 [0.5999206  0.10002    0.10002051 0.10001997 0.10001892]
 [0.59992063 0.10001997 0.10002049 0.10001999 0.10001892]
 [0.10002674 0.59990445 0.1000238  0.10002308 0.10002193]
 [0.10002677 0.10002311 0.10002375 0.59990452 0.10002185]
 [0.10002607 0.10002247 0.59990774 0.10002241 0.1000213 ]
 [0.5999207  0.10001992 0.10002052 0.10001996 0.1000189 ]
 [0.10002602 0.10002252 0.5999076  0.10002248 0.10002138]
 [0.10002681 0.1000231  0.10002381 0.59990439 0.10002189]
 [0.59992056 0.10001999 0.10002058 0.10001997 0.10001891]
 [0.10002609 0.10002251 0.59990757 0.10002248 0.10002135]
 [0.10002678 0.59990447 0.10002373 0.10002309 0.10002193]
 [0.10002603 0.10002247 0.59990776 0.10002243 0.10002132]
 [0.59992067 0.10001996 0.10002047 0.10001996 0.10001893]
 [0.10002676 0.59990462 0.10002367 0.10002307 0.10002188]
 [0.59992067 0.10001993 0.10002047 0.10001996 0.10001897]
 [0.10002817 0.10002444 0.10002501 0.10002433 0.59989804]
 [0.10002682 0.59990424 0.10002385 0.10002314 0.10002194]
 [0.10002822 0.10002439 0.10002507 0.10002435 0.59989797]
 [0.10002674 0.10002302 0.10002377 0.5999045  0.10002197]
 [0.59992075 0.10001995 0.10002049 0.10001993 0.10001888]
 [0.10002603 0.10002242 0.59990778 0.10002249 0.10002128]
 [0.10002823 0.10002436 0.10002502 0.10002435 0.59989804]
 [0.10002817 0.10002429 0.10002506 0.10002441 0.59989806]
 [0.10002606 0.10002242 0.59990779 0.10002242 0.1000213 ]
 [0.10002609 0.10002246 0.59990762 0.1000225  0.10002133]
```

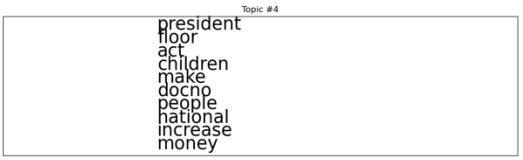
The top 15 words were then plotted to try to understand the topics.

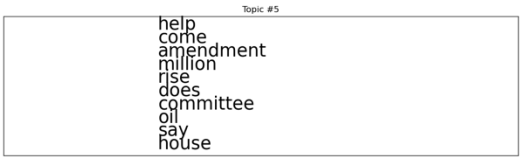


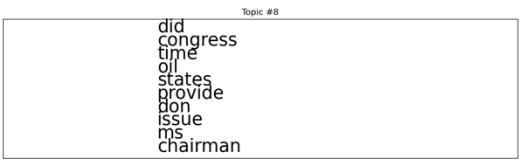
Based on the plot, we can make the following inferences.

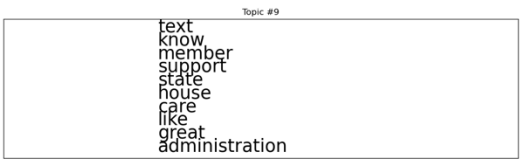
- Topic 0: Seems to be about American veterans and security, possibly discussing governmental programs and funding (e.g., "billion," "money") related to veteran affairs.
- Topic 1: This could pertain to federal administration or policy, as indicated by words like "administration," "federal," "congress," and "2008".
- Topic 2: The words suggest issues related to family rights or oil and energy policy debates on the "floor". Since this has a mix of topics, it may be beneficial to pull more words to understand the topic better.
- Topic 3: Appears to focus on military or national energy matters, with words like "military," "national," and "energy" prominent.
- Topic 4: Relates to government budget matters, possibly in a specific year ("2007") and involving presidential and congressional actions or votes on budget increases.

If we increase the number of topics to 10 we see better defined categories that are easier to infer what the topics are about.









- Topic 1 Focuses on budgetary concerns within the United States, possibly involving the government, the defense sector ("dod" could refer to the Department of Defense), and financial aspects ("billion").

- Topic 2: Appears to cover legislative aspects, with words like "law," "tax," "resolution," and "2007" indicating a focus on legal frameworks and perhaps financial legislation in a given year.

- Topic 3: Relates to taxation and legislation, with "tax," "country," "fact," and "resolution" suggesting debates or discussions around fiscal policies.

- Topic 4: Centers around presidential actions or policies affecting children and national issues, with the possibility of financial implications ("increase," "money").

- Topic 5: May involve discussions or amendments related to legislation, with a financial angle suggested by "million," "rise," and "committee."

- Topic 6: Looks to be centered on voting, possibly within the context of an election year ("2008"), and includes discussions about the administration and thanking colleagues, which is common in political discourse.

- Topic 7: Touches on contemporary issues ("today"), security, and national topics, with "need," "nation," and "security" indicating a focus on urgent or significant matters.

- Topic 8: Could be related to legislative procedures, with "did," "congress," "issue," "house," and "ms" (possibly referring to another member of Congress)

- Topic 9: It difficult to determine the topic of this text, as there are not key identifies. This may be a good example of showing more words per topic to get a better understanding.

Conclusion

Topic modeling is a powerful tool for summarizing and understanding large collections of text data by identifying the underlying themes or topics without needing to read through all the material manually. For a human to have to read each file and access the topics, it would take countless number of hours. However we can use topic modeling to quickly predict the topic. While it may not be perfect, it does a great job of providing an idea of what the topic can be, and with some fine tuning can become better.