Nicholas Louis Brown
Professor Yu
IST736
November 6, 2018

Topic Modeling Text Documents Using Latent Dirichlet Allocation

Intro:
Latent Dirichlet Allocation (LDA) is a popular choice for gaining insight into unstructured text data. Using LDA I analyzed a dataset consisting of text containing the floor debates of the 110th Congress. By using LDA I hope to extract the major topics discussed by our elected leaders to get a better idea of what is going on in Congress. My predictions for this experiment are hopeful. I feel that LDA is a very good choice for analyzing this kind of data and I believe we will see common US political issues if the topic modeling is successful.

Methodology:
Though the data was organized by party and gender I was more interested to see the discussion at large so I decided to analyze the entire folder. I began the analysis by reading in the text files, I got multiple errors due to sklearn attempting to create the model by reading the file as utf-8 throwing multiple errors. To solve this problem, I read the data in using the encoding setting “ISO-8859-1”. Finally, the settings I adjusted for LDA included the number of topics and number of features. I ended up keeping the number of features at 1000 and the number of topics to 20. Any smaller topic size than 20 and I felt many topics were left out leaving mostly irrelevant material. In addition to pythons sklearn I also used mallet to do some deeper digging. Models were evaluated primarily by hand examining the outputted topics from the models.

Python Sklearn Results:
Using python and sklearn I was able to do a simple preliminary LDA to create some topics which could guide the deeper dive using mallet. Below is a table showing the 20 topics generated from 1000 features. I feel these topics contain a lot of political themes however they are not very focused and contain a lot of ‘filler’. For example, line one contains both topics I expected like ‘oil’, ‘energy’, and ‘tax’; and topics I did not expect or feel necessary like ‘increase’, ‘care’, or ‘children’. Further down are the results from my attempt at using mallet to try and “trim the fat” from these results.

```
Topic 0:
oil energy tax increase money spending budget billion care children
Topic 1:
housing georgia frank massachusetts money financial mortgage jersey affordable yield
Topic 2:
tax budget majority spending money increase appropriations taxes billion democrat
Topic 3:
texas military border law veterans rights department foreign human forces
Topic 4:
energy minnesota land lands farm yield park conservation resources river
Topic 5:
care 30 sure group murphy republican talk veterans connecticut talking
Topic 6:
ohio blue money billion budget debt coalition tax fiscal got
Topic 7:
care medicare children insurance medical drug cost texas law things
Topic 8:
mrs energy children tax families virginia budget care increase pay
Topic 9:
ms florida children energy families administration care democrats veterans budget
Topic 10:
education energy children college students families programs rule consideration yield
Topic 11:
york rule rules consideration providing yield energy friend food tax
Topic 12:
children veterans care ms families military education energy administration life
Topic 13:
intelligence hoyer senate foreign energy bipartisan legislative week information passed
Topic 14:
children energy ms oil care yield families life florida tax
Topic 15:
carolina department yield water transportation balance appropriations agencies homeland programs
Topic 16:
trade tax workers jobs agreement families energy relief free care
Topic 17:
iowa oil got things energy law life children little king
Topic 18:
energy oil gas natural prices price coal fuel really production
Topic 19:
texas energy tax money majority research children billion things oil
```

Mallet Results:
I used mallet after sklearn to better understand the differences that tuning the model can make. Using mallet, I was able to take a deeper dive into the workings of the model. Right of the bat mallet felt more powerful as it worked through the command line. Again, interpreting the output was difficult but I wanted to see how similar the model would behave when compared to scikit learns implementation. For the mallet model I set the number of topics to 10. For whatever reason mallet took significantly longer to process the data than scikit learn. In my case I was somewhat disappointed by the output from mallet. I feel it included many more “trash words” which I attribute to the stemmer and vectorizer settings. Despite my disappointment the mallet output was still somewhat accurate including words like ‘immigration’ and ‘security’.

```
nbrown@Nicholas-MacBook-Pro [~/desktop/ist736/hw8] > cat sample-keys.txt
0       0.90943 bill house text doc docno speaker committee time act representatives gentleman amendment chairman members vote yield rule appropriations rules program
1       0.63693 docno doc text bill act chairman house national representatives amendment support federal water h.r time legislation program committee state transportation
2       0.35909 energy oil gas percent years it's world country people natural prices don't that's today production time america fuel we're price
3       1.19284 bill act docno text doc support legislation house health today speaker h.r children representatives families colleagues important program congress education
4       0.60841 people doc docno text house iraq president congress american country representatives make speaker war it's care don't time back years
5       0.69218 doc docno speaker text house representatives time veterans support today mrs great service madam american day life national resolution war
6       0.63375 iraq united war people states text world doc docno security u.s support government speaker american resolution president military human foreign
7       0.35613 housing people health time care insurance program year bill system medical don't country frank medicare money back massachusetts states congress
8       0.3004  texas people states united speaker border law american iowa america country congress jackson-lee court government federal king rights immigration justice
9       0.41338 tax american budget text docno doc spending people house speaker money increase taxes government federal majority congress percent representatives democrats
```

Conclusion:
LDA is very effective for topic modeling unstructured text data. By feeding the algorithm relatively uncleaned data the limitations and capabilities of the algorithm are much more apparent. LDA is rather effective at selecting topics which may make sense however it is evident from both tests that the algorithm’s output can be bogged down and muddled by extra words and failure in the vectorization and stemming settings. Based on the experiment results I would certainly consider using these algorithms for other applications and call our LDA experiment a mild success.