Darrell Nelson II

Text Mining (IST - 736)

09/04/2019

HW08: Topic Modeling

Introduction

Large amounts of data are collected every day. With this huge boom of data production and utilization, it becomes difficult to interpret what we are looking for. So, we need tools and techniques to organize, search, and understand vast quantities of information. Topic modelling provides methods to organize, understand and summarize large collections of textual information. It helps in:

- Discovering hidden topical patterns that are present across the collection
- Annotating documents according to these topics
- Using these annotations to organize, search and summarize texts

Topic modeling is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents. It can be described as a method for finding a group of words (i.e a topic) from a collection of documents that best represents the information in the collection. It can also be thought of as a form of text mining – a way to obtain recurring patterns of words in textual material.

Analysis and Models

About the Data

Dataset is comprised of political text files separated into four subfolders. The folders were separated out by sex (male/female) and party affiliation (democrat/republican) based on their author. For this experiment only the female democrat folder was used due to computer processing issues. The text files are read into a Python IDE (Spyder) and converted into a list. This list acts as a corpus with each item in the list accounting for a different text. Stop words were then removed, data was then vectorized and lemmatized before being placed into a dictionary which contained the frequency count of each word accompanied by the actual word. The dictionary is then converted into a corpus and feed into the Latent Dirichlet Allocation model.

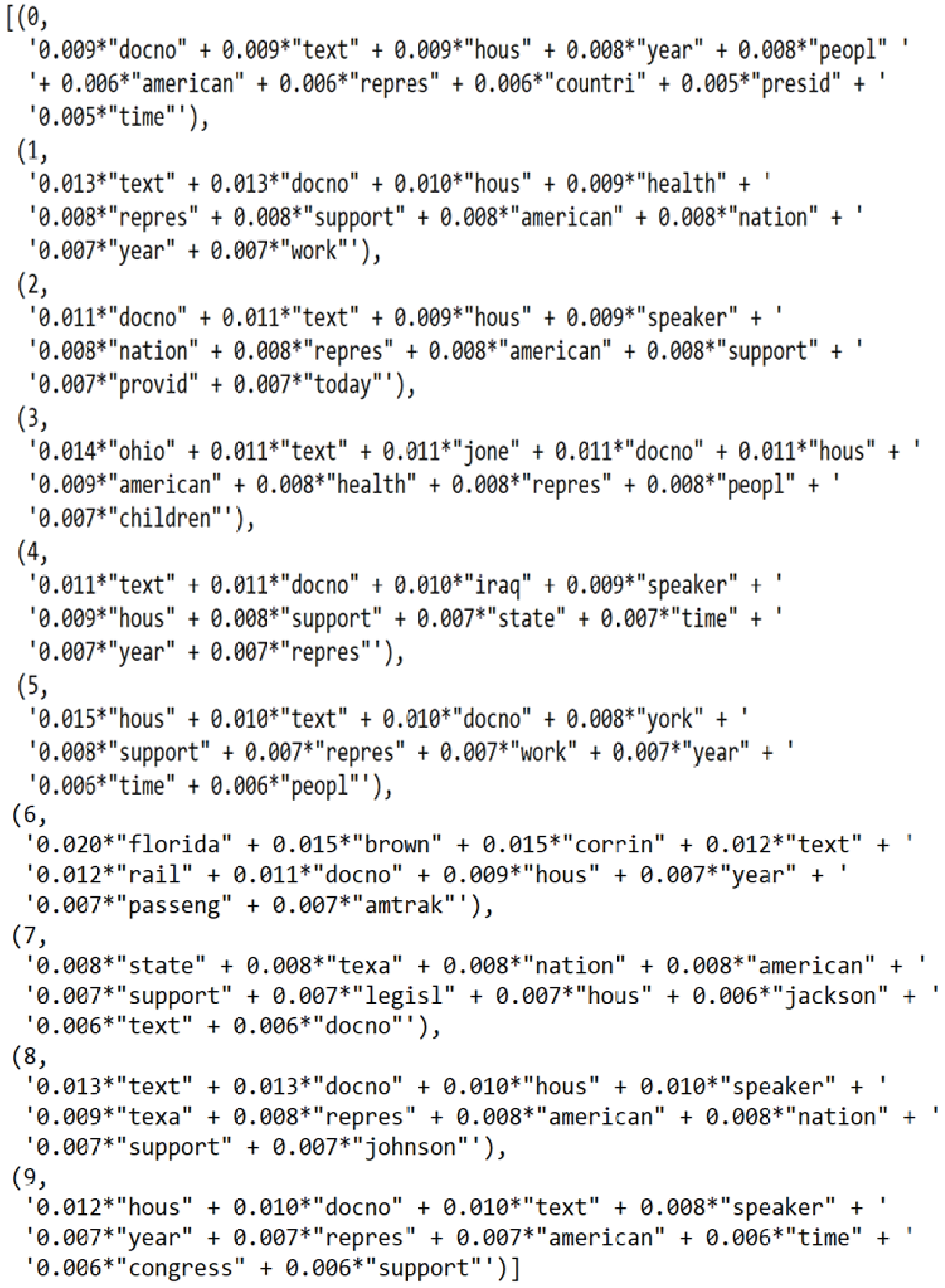
Models

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a statistical model used in unsupervised machine learning. It allows sets of observations to be explained by unknown groups based on their feature similarity. It is mainly applied in natural language processing where the observations are documents, features are the words in the documents, and the unknown groups are the different topics the documents could be about. LDA takes a fixed number of topics and associates a set of documents with those topics based on the frequency of the words in said documents.

Results

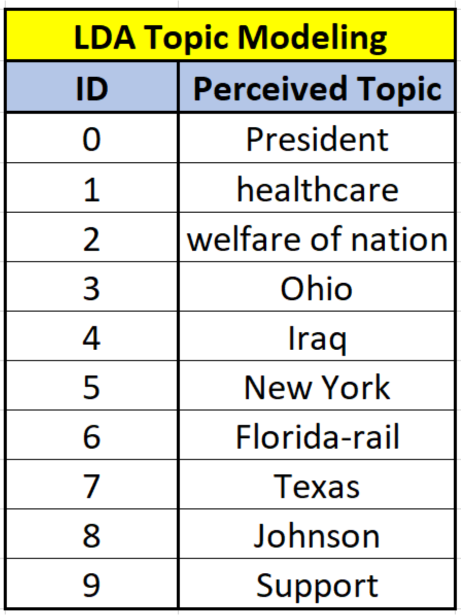
The LDA model was configured to produce 10 topics from the 50 documents in the female democratic section. Raw output below:



```
[(0,
  '0.009*"docno" + 0.009*"text" + 0.009*"hous" + 0.008*"year" + 0.008*"peopl" '
  '+ 0.006*"american" + 0.006*"repres" + 0.006*"countri" + 0.005*"presid" + '
  '0.005*"time"'),
 (1,
  '0.013*"text" + 0.013*"docno" + 0.010*"hous" + 0.009*"health" + '
  '0.008*"repres" + 0.008*"support" + 0.008*"american" + 0.008*"nation" + '
  '0.007*"year" + 0.007*"work"'),
 (2,
  '0.011*"docno" + 0.011*"text" + 0.009*"hous" + 0.009*"speaker" + '
  '0.008*"nation" + 0.008*"repres" + 0.008*"american" + 0.008*"support" + '
  '0.007*"provid" + 0.007*"today"'),
 (3,
  '0.014*"ohio" + 0.011*"text" + 0.011*"jone" + 0.011*"docno" + 0.011*"hous" + '
  '0.009*"american" + 0.008*"health" + 0.008*"repres" + 0.008*"peopl" + '
  '0.007*"children"'),
 (4,
  '0.011*"text" + 0.011*"docno" + 0.010*"iraq" + 0.009*"speaker" + '
  '0.009*"hous" + 0.008*"support" + 0.007*"state" + 0.007*"time" + '
  '0.007*"year" + 0.007*"repres"'),
 (5,
  '0.015*"hous" + 0.010*"text" + 0.010*"docno" + 0.008*"york" + '
  '0.008*"support" + 0.007*"repres" + 0.007*"work" + 0.007*"year" + '
  '0.006*"time" + 0.006*"peopl"'),
 (6,
  '0.020*"florida" + 0.015*"brown" + 0.015*"corrin" + 0.012*"text" + '
  '0.012*"rail" + 0.011*"docno" + 0.009*"hous" + 0.007*"year" + '
  '0.007*"passeng" + 0.007*"amtrak"'),
 (7,
  '0.008*"state" + 0.008*"texa" + 0.008*"nation" + 0.008*"american" + '
  '0.007*"support" + 0.007*"legisl" + 0.007*"hous" + 0.006*"jackson" + '
  '0.006*"text" + 0.006*"docno"'),
 (8,
  '0.013*"text" + 0.013*"docno" + 0.010*"hous" + 0.010*"speaker" + '
  '0.009*"texa" + 0.008*"repres" + 0.008*"american" + 0.008*"nation" + '
  '0.007*"support" + 0.007*"johnson"'),
 (9,
  '0.012*"hous" + 0.010*"docno" + 0.010*"text" + 0.008*"speaker" + '
  '0.007*"year" + 0.007*"repres" + 0.007*"american" + 0.006*"time" + '
  '0.006*"congress" + 0.006*"support"')]
```

```
Perplexity:  -7.459630416997474
```

These 10 topics were then assigned labels that best coincide with what subjects the politicians may be talking about.



LDA Topic Modeling	
ID	Perceived Topic
0	President
1	healthcare
2	welfare of nation
3	Ohio
4	Iraq
5	New York
6	Florida-rail
7	Texas
8	Johnson
9	Support

For example, TopicID 1 has “health” and “support” as two of the main grouped words. Therefore, it is justifiable to assume this topic is based on healthcare. This type of logic was utilized for all 10 topics to create the ‘Perceived Topic’ column in the table above.

The perplexity function is a statistical measure of how well a probability model predicts a sample. It takes the trained LDA model and the given theoretical word distributions and compares them to the actual topic mixtures or distribution of words in the documents. The function rates models on a negative scale (the more negative the better). Perplexity without context is meaningless, however. There must be something to act as the baseline and since each experiment is unique the perplexity must be chosen based on running multiple LDAs on the same dataset with a different number of topics. The topic size with the lowest perplexity is chosen as the winner. With only one iteration of testing being done in this experiment it is not possible to determine if this model is optimal or how close it is to the optimal solution.

Conclusion

In conclusion, LDA is a great way to cluster text data into different groups with little human annotation. It is quite time intensive in terms of computer processing, but a dedicated server and automation can keep the results coming even after the data scientists clock out. With this in mind, it is an extremely useful tool to capture the sentiment of a particular population. It can also be used to gauge audience engagement and customer feedback. With so many companies willing to pay through the nose to not only understand but to manipulate the masses into thinking their product is better than the competition it makes logical sense for LDA and topic modeling to be a skill in the arsenal of any established data scientist.