Latif, Aneeka
IST 736
HW 8

Introduction

There are many forms of text media on the internet – magazine articles, news pieces, and literature to name a few. Many websites rely on user submitted or author curated tags to classify these text documents. Model classification based on tags or annotations is known as supervised learning.

In some cases, tags may not be available, or it may be a cheaper and more scalable alternative to use a method that does not rely on annotations. Topic modeling is a way of categorization based on the document text itself. This has very useful real-world applications with real time news or social media content classification which more often than not will not have content tags. Social media platforms could leverage topic modeling by choosing this classification over user provided content tags which could be missing or unreliable. Social media platforms could further use topic modeling to flag or remove harmful content per user preference.

Analysis and Models

About the Data

The dataset is comprised of 429 text files of political speeches given in the 110th floor debate by the House of Representatives. The speeches include both male and female speakers as well as both democratic and republican representatives. Each text file includes the entire speech as well as html formatting.

```
</TEXT>
</DOC>
<DOC>
<DOCNO>Mr. DONNELLY. (VETERANS' BENEFITS IMPROVEMENT ACT OF 2008 -- (House of
Representatives - September 24, 2008))</DOCNO>
<TEXT>
   Mr. DONNELLY. I want to thank the ranking member of the committee, my friend from my
home State of Indiana. Our districts actually touch up upon each other, and I'll try to be
brief and not use too much of his time.
   One key provision in this bill, as amended, would ensure that severely injured veterans
released from active duty are able to receive disability benefits immediately for injuries
that can be promptly rated while they wait to be assigned a rating for other injuries that
are not immediately ratable.
   The bill before us would codify temporary ratings for severely injured veterans who
have paid a high price on behalf of our country. The passage of this legislation will make
temporary ratings a right of our wounded warriors, instead of just an option to be
employed by the Veterans Administration.
   I want to take a moment to thank my good friend, the chairman of the Disability
Assistance and Memorial Affairs Subcommittee, Mr. Hall, for his work on this legislation.
I want to thank Chairman Filner and I want to thank Ranking Member Buyer for their
assistance and leadership on this issue as well.
   We have much work to do to continue to improve the way our disability claims process
works for injured veterans. However, S. 3203 represents real change that will directly
translate to improved service for those Americans who have fought and sacrificed on behalf
of our Nation.
   I urge all my colleagues to vote for this bill. I want to thank again the ranking
member for his graciousness.
</TEXT>
</DOC>
```

Figure 1: Example of raw data file

The 429 files were loaded into a corpus. Steps were taken to clean the data of formatting, special characters, numeric, and all other non-alphabetic characters. English stopwords from the NLTK corpus were also removed from the dataset. Additional stopwords relating to legislation and referential titles were removed as well, such as "mr" ,"representative","senate","legislation", and "house". The tokenized data was then loaded into a dataframe and passed to a model for topic classification.

Methods and Models

Latent Dirichlet Allocation

Latent Dirichlet Allocation is an algorithm used for topic modeling. LDA operates under the assumption that topics are combinations of words (tokens) and documents are mixtures of topics based on some

calculatable topic probability. Given a set of documents and a set of vocabulary, LDA calculates the distribution of words in each topic and the distribution of each topic for each document.

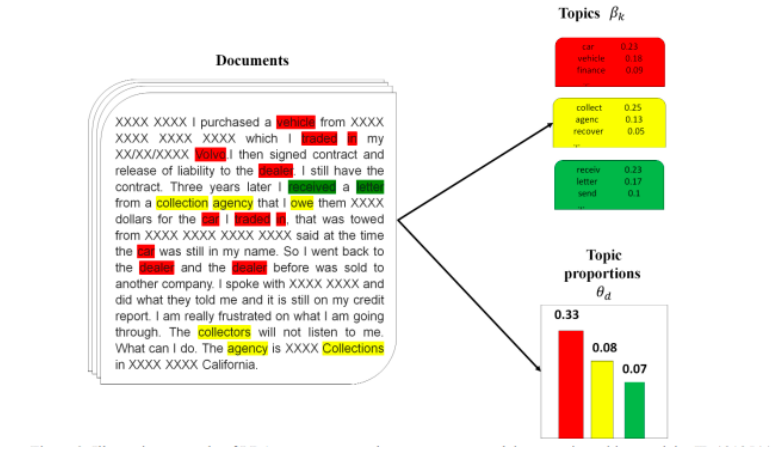

Figure 2: Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints (Bastani et al)

The first model ran on normalized vectorized data – the dataframe consisted of inversely normalized word counts per document with an ISO-8859-1 encoding and special characters removed. Vectorizing by n-grams did not make a significant difference in the results, so that parameter was not added. The LDA was ran with 10 iterations to find 5 topics.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
thank	country	energy	united	country
nation	states	first	states	energy
health	need	cobble	bilirakis	iraq
energy	government	united	thank	states
amendment	tax	work	work	thank
care	war	becerra	blunt	need
children	get	thank	rise	national
tax	doggett	national	america	committee
carolina	way	believe	iraq	amendment
cost	administration	fallin	program	work

There is some overlap between the topics but topic 1 seems to focus on health, energy, taxes and budget. Topic 2 is centered around taxes, wat, and Senator Doggett. Topic 3 includes energy once again, and Representative Becerra. Topics 4 and 5 are similarly about iraq and creating legislative programs/amendments.

Conclusion

In conclusion, topic modeling can be very useful for classification in instances where it is not feasible to use human annotation. The results of topic modeling can be less intuitive than other forms of classification, and often there is some overlap between topics.

Latif, Aneeka
IST 736
HW 8

Real world applications of topic modeling would be useful in any form of content recommendation. Topic modeling can be used to group together similar content that can't be rigidly described with a single category. This may not be as accurate as human annotation but would indefinitely be more scalable and therefore, more affordable than human annotation. While topic modeling may not be the most straightforward classification method, it is surely necessary for the future of text classification.