LDA Topic Modeling

Scott Snow – March 18, 2019

IST 736 – Text Mining

Introduction

One of the primary tasks in data mining is classification. When it comes to collections of documents, there are many ways to classify each one. They can be classified by sentiment, language complexity, length, etc. While many of these are simple, one that is complicated is topic classification. Specifically, it is difficult when the topics aren’t known before hand and must be derived from the documents themselves. This is because a computer doesn’t know which words are central to a document. It can however compare words that appear in many places and compare them to those that don’t. The words that don’t appear in many places can then be used by a human to infer a general title for each category.

The main benefit to topic modeling is to gain information about the source of the data. This information can be gathered in real time such as the topics of tweets being made on a given day, or they can be historical such as looking at an American history textbook to see how prevalent controversial segments such as slavery or wartime use of nuclear weapons are. Both types of information can be used to better a specific brand or product or better humanity.

A close up of a logo

Description generated with high confidenceAbout the Data

This specific data set comes from the minutes of the 110th House of Representatives. It contains a set of monologues made by each house members. The set of documents are file sorted into four categories: male/Democrat, male/Republican, female/Democrat and female/Republican. The distribution of these for categories is shown to the right. The graph shows that the house was very male dominated. This cohort of males seems to be evenly split between republican and democrat. Females on the other hand were primarily democrat. Female democrats outnumbered their republican counterparts more than two to one. There are 428 total documents

Model Discussion

The model used to obtain the documents was Latent Dirichlet Allocation. In simple terms, LDA looks at the probability of term co-occurrence given a specific set of topic terms. Then each document is given a probability for each topic which is based off a list of common co-occurring terms. By combining the log of each likelihood, a numerical summary of a specific allocation can be obtained. The primary reason for doing this is to choose the number of topics that results in the highest accuracy. In this process, the values 20, 30, and 40 were chosen to begin with. After a curve formed, 35 was chosen. After more points where chosen around 35 and 35 still had the highest average, the output was made with 35. This however resulted in a lower value. To be sure, more outputs were created with nearby values. 35 still proved to be the most effective number of topics.

Based on preference and due to an issue and tokenizing bigrams in console, a vectorization that included only unigrams and removed English stop words was used. No stemming or lemmatization was done. Additionally, each document is in a mark-up format. Since these only take up a small portion of the words used, the documents were not modified. Once the list of key terms was obtained, labels were assigned to each topic that most closely matched the set of terms. Only one labeler was used but a more robust method would either to be use multiple labelers, or to have workers classify the label based on accuracy.

A screenshot of a cell phone

Description generated with very high confidenceResults

Most of the documents had a max likelihood for a topic of around .2. Some were higher with a highest likelihood around .7. On the left is a distribution of these likelihoods.

A picture containing text

Description generated with high confidenceA close up of a logo

Description generated with very high confidenceA picture containing text

Description generated with high confidenceA close up of a logo

Description generated with high confidenceOf topics, democrats, both male and female, discussed Health Care the most. Female republicans seemed to be primarily concerned with tax policy and budgetary issues. Male republicans were the most diverse group in terms of topic focus. Below is a breakdown for each group by topic.

Conclusions

A benefit from this analysis is deriving the goals and motivations for each political party and for each gender within those parties. The 110th congress was in session between January 3, 2007 and January 3, 2009. During this time, health care was a big issue. The efforts by the democrats to pass a large-scale health bill seemed to be in motion during this time and of course was finalized with the Affordable Health Care Act. Additionally, the housing market was a concern during this time and so was one of the larger focuses for some groups. Additionally, it is informative to compare what was important at the time to what is important now. A controversial topic for debate currently is gun control and border issues. Gun control is not even a topic in this analysis and border issues is only a minor topic compared to some of the others. For the issue of gun control, most people know why that has emerged as an important issue. For the topic of border control however, it would be interesting to look at this analysis for each congress since the 110th to see if border issues grew in relevance or emerged from obscurity during this current iteration of congress.

Topic modeling continues to be a challenging endeavor. This is partially because some lists of keywords can be too arbitrary. Consider some of the labels, “Patriotism,” “Nationalism,” and “Speakers from Florida.” These topics are arbitrary, and it is unclear to the reader what prompted this label and whether a reader would give a similar label if presented with the data. Ultimately, topic modeling and data mining by extension are not meant to be the absolute solution. Merely they are tools that help uncover information that would not be obtainable by normal means.