IST736 Text Mining - Homework 8

Martin Alonso

Introduction

The objective of this exercise is to use LDA to analyze to most trending topics of discussion on the 110th Congress floor debate. Using 400 files from the debate and the Mallet software, we'll explore the most common topics and trends that appear in the discussion.

Methodology

After installing Mallet on a Linux machine, the entire LDA dataset was saved into one directory and loaded into Mallet using the import-dir function. For this analysis, the additional input options given were to remove stop words, and focus on unigrams and bigrams for topic identification.

With these parameters set, the algorithm was designed to focus on the twenty most common topics discussed on the floor debate. However, many of the topics repeated themselves, having identified discussions on Health, Education, Foreign Policy and Immigration, National Security, and Economics.

Given that these topics were repeating themselves, the algorithm was altered, changing the number of topics from 20 to six, in order to obtain the most detailed topic-word relation as possible. These results were more satisfying as they concentrated the the first results into tighter topic bins, though, these were also harder to identify as the words were more spread out.

Results

Topic	DP	Words
0	0.25	housing text doc gentleman chairman docno bill frank financial mortgage massachusetts people amendment act insurance house home affordable program public
1	0.25	mr bill appropriations chairman amendment text docno doc gentleman million funding house related year agencies act science president research program
2	0.25	tax bill text mr budget docno doc energy year house billion act million representatives pay today economy farm relief income
3	0.25	trade text minnesota doc president docno workers jobs agreement american house people america colombia illinois peru free labor representatives percent
4	0.25	mr president house text doc docno congress iraq speaker care representatives people health american florida republican working budget members important
5	0.25	mr text iowa doc national speaker land bill federal state act house lands congress water county park representatives utah public
6	0.25	mr bill docno text doc house act time representatives legislation support chairman committee amendment government federal program national law yield
7	0.25	mr transportation text water doc docno coast house oberstar rail gentleman states infrastructure chairman public act safety guard committee federal
8	0.25	health care children insurance speaker medicare medical year program bill percent texas schip coverage system drug mental cost state house
9	0.25	intelligence security text doc docno act foreign homeland house carolina national protect bill mrs department united york south states community
10	0.25	mr american tax doc docno speaker text spending house budget majority government increase democrats money taxes people georgia federal percent
11	0.25	energy oil gas percent world natural years prices coal production fuel today america price drill country supply drilling produce nuclear
12	0.25	mr text doc docno speaker house representatives time mrs davis resolution service illinois madam great national office support life university
13	0.25	iraq war mr speaker american troops america military world today president day docno iraqi nation people life text al children
14	0.25	people don country back time make things put years congress fact united work good ve states thing money great lot
15	0.25	mr speaker house bill committee doc text docno time rule gentleman rules representatives members vote consideration majority friend floor act
16	0.25	support ms act legislation nation children american today colleagues education families speaker important congress rise students country work national people
17	0.25	veterans mr text doc docno military house war speaker care bill representatives va iraq chairman troops support defense service forces
18	0.25	texas states united border speaker mr law american people court justice crime america immigration docno doc illegal text drug federal
19	0.25	united states rights lee resolution jackson speaker world human people support government international foreign security ms israel text countries million

Table 1: 20-topic words

Looking through 20-topics, we find that there are many overlapping themes. Table 1 shows the first five results to be talking about, broadly speaking, housing, economy, energy, economy, and foreign policy. But, looking down the list, we find that the topics start repeating themselves. Foreign policy appears again as part of the 14th and 20th topic. Economy is again discussed in topic 15.

Furthermore, there are other topics approached which are not among the top 5, namely health, infrastructure, education, and immigration. Overall, there are close to 11 single topics within these 20. And, we could argue, that these could be narrowed down further more to reduce the number of

topics. For example, topic 18 talks about veterans. But it also references the military, which is mentioned in issues on foreign affairs.
With this initial analysis, we narrowed the number of topics to six, table 2, showing that the ideas discussed in the house are more synthesized than what they initially appear to be.

Topic	DL	Words
0	0.83333	mr house text doc docno bill speaker people representatives tax time gentleman american congress committee don budget majority money chairman
1	0.83333	energy oil people gas percent country years don world back america states time things make congress american prices year lot
2	0.83333	health docno doc text care president house mr speaker children people iraq congress american families ms country representatives veterans today
3	0.83333	mr bill act docno text doc house support chairman legislation committee representatives amendment time program national federal today energy year
4	0.83333	mr speaker doc docno text house representatives time american states nation great united national today day madam life people service
5	0.83333	iraq united states people speaker war security world american government mr rights president support text america law intelligence foreign act

Table 2: Six-topic words

The narrowed down topics suffer a bit from underfitting, however. We notice that all topics contain one form of ‘mr’, ‘president’, ‘docno’, ‘bill’, or ‘speaker’. They all, also, reference the people and the United States. But there is still a semblance of what the overall discussion is pointing to. Topic 1 is referring to taxes and the American budget. Topic two is on energy, namely oil and prices. Topic 3 discusses health with relation to veterans, and Topic 4 talks about war and Iraqi children, a topic broached again in Topic 6.
But we are also losing information on some other valuable topics. Education, immigration, and health are no longer discussed, with the latter only in reference to veterans. Gone are the discussions on infrastructure and the economy. Where we first suffered from overfitting the topic discussion, we now suffer from topic underfitting.
Ten topics should thus do the trick. However, as table 3 shows, the topics, again, repeat themselves.

Topic	DP	Words
0	0.5	docno text mr doc speaker house support representatives ms time nation american mrs colleagues women today rise resolution service national
1	0.5	mr text docno doc house president iraq veterans people congress speaker representatives american war care country troops make health important
2	0.5	health bill care children act mr support legislation program year speaker education house docno text doc time today years medicare
3	0.5	doc text docno housing tax bill act american today legislation trade house families america people representatives congress jobs support economy
4	0.5	mr house bill text doc docno speaker representatives gentleman committee time chairman tax majority amendment budget vote year spending act
5	0.5	energy oil gas percent world years prices natural today america fuel production country coal price drill supply mr drilling don
6	0.5	mr speaker docno border text doc states united america texas american law carolina house madam representatives congress south americans nation
7	0.5	mr iraq doc text docno house war security president bill support act united speaker states military intelligence troops representatives world
8	0.5	people back don country time states united american government things money congress make put years fact part good system state
9	0.5	mr bill doc text docno act house chairman amendment representatives committee national time support legislation gentleman program federal water speaker

Table 3: 10-Topic words

For starters, the military and the economy are discussed in three of the topics. Health and Education are combined into one single topic; infrastructure, energy, and immigration are each given a single topic. Overall, we find that the topics repeat themselves, but there is no single way of identifying single topics, mainly because there are words that repeat themselves but don’t add much value or context to each discussion.

Conclusions

Overall, we can identify several topics such as economics, health, education, foreign policy (both immigration and military affairs), and infrastructure. These topics overlap and are discussed so frequently, that it is difficult to pinpoint one single set of words that would clearly relate to a single topic.

This problem is exacerbated by the fact that we are able to remove stop words, but have not removed words that repeat themselves and have no meaning. Words such as ‘mr’, ‘bill’, ‘president’, and ‘speaker’ are common in all texts and topics, while adding nothing to discriminate among the topics discussed. If we were to remove these words altogether from the text training, we could better identify the topics discussed.

Furthermore, by not stemming or lemmatizing, we are also getting similar words weighing heavily on several topics. Adding these features could ameliorate our results, showing more clarity among the topics to better identify them.

However, given the tools at hand, we can safely say that we have successfully identified not only four main topics discussed, but four additional topics that were not outlined in the initial objective, but could may as well be part of the ongoing discussions held in the House.

References

- https://programminghistorian.org/en/lessons/topic-modeling-and-mallet