Introduction

In the United States, many towns and municipalities use school boards as a steering factor to help make decisions about the school system in the town. We seek to simplify and understand these sometimes long meetings by applying the methods of Latent Dirichlet Allocation to summarize the topics covered in school board meetings. Our dataset consists of data from roughly 350 school boards from 2019 - 2022. In an original version of the dataset each row represented one meeting, however for this dataset called chunked_meetings.csv, we have separated each meeting's text into chunks of 300. This is because typically many different topics are covered in each school board meeting such that if we were to group by meeting we would not find interpretable findings. Thus we chunk our text data into smaller chunks so that it is more manageable and so we can more easily assign each row to various topics.

Methods

The first pre-processing step we took was applying a function to lowercase the text, remove punctuation and lemmatize the text. It does this row by row transforming each instance of raw text. Normally, we would perform more preprocessing with this function, but, we want to utilize bigrams and trigrams so we only perform the above to start.

Next, we apply `Phrases` class from the Python library Gensim to identify these bigrams, trigrams, and connector words such that "pledge of allegiance" is identified as one token. Only now do we remove stopwords (We wanted to keep stop words that are connector words) and single instances of letter.

Because we are going to use the Gensim library for our entire modeling process, we now have to create a Gensim formatted corpus. Essentially it takes each word in the corpus (our dataframe) and assigns it to a number. After that we limit the dictionary to only include tokens that appear in 20 documents and exclude tokens that appear in more than 65% of documents.

Now we are ready to begin creating our 6 models. We propose two types of text representations: bag of words and tf-idf which combined with various topic model sizers in LDA will yield 6 different topic models. They are listed below:

1. LDA with 50 topics and Bag of Word representation of the documents
2. LDA with 75 topics and Bag of Word representation of the documents
3. LDA with 100 topics and Bag of Word representation of the documents
4. LDA with 50 topics and tf-idf representation of the documents
5. LDA with 75 topics and tf-idf representation of the documents
6. LDA with 100 topics and tf-idf representation of the documents

Results

After fitting these models, our metric to determine the most interpretive model was using Gensim's Coherence Model. Then, using the "most" coherent model, manually interpret ourselves and label some of the most important topics.

Topic 5: Accounting

- There are a number of words relating to finances and resources in this topic. Fund, general_fund, account, supplies amount, interest, and bond are all words relating to accounting.
- Evidence: The first seven documents ordered by topic 7 probability are about either asset reports, purchases, or account reports. There are dollar signs, numbers, and an explanation of where district funds are going.

Word_5	Prob_5
fund	0.036625
total	0.018945
supplies	0.018786
account	0.018003
amount	0.017195
general_fund	0.014151
june	0.013978
report	0.012914
interest	0.010634
bond	0.010460

Topic 16: Initiative Analysis

- This topic has to do with goals and programs established by the district, and the measure of their completion. Goals, services, project, and systems all have to do with initiatives, and data and goals have to do with tracking those initiatives.
- Evidence: Each of the top 5 documents has to do with goals, and more specifically, self-evaluation is frequently referenced. Initiatives such as team-building activities and orientation are discussed, as well as whether data supports their implementation or expansion.

Word_16	Prob_16
data	0.019795
goals	0.017808
services	0.017398
project	0.017231
systems	0.011593
contract	0.011438
board	0.010868
classroom	0.008738
resolution_no	0.008593
construction	0.008564

Topic 21: Teaching

- This topic has words related to teaching. Teachers, grades, math, english, and students all reflect education.
- Evidence: The top 5 documents are of an interesting variety, but they all have to do with teachers and their relationship with students: the highest is a listing of teachers are appointments for extra hours, but the others include the proposal and approval of a special-education program with the student specifically in mind, an art festival that students attended, and the retirement of long-tenured teachers.

Word_21	Prob_21
teachers	0.019599
grades	0.016767
math	0.014951
students	0.013242
tenure	0.012126
english	0.012008
reading	0.011124
day	0.011005
will_be	0.010267
january	0.009387

Topic 40: Tupper Lake Central School District

- The top documents in this topic are all from the same dialogue which comes from the Tupper Lake Central School District and was about teachers resigning, school positions, and sports coaches. Certain school board member names occur frequently across the topi documents in this topic.
- We picked this to show the weakness of LDA, as it is possible for a minority conversation to end up dominating a topic just because it is dissimilar from everything else
- Evidence the phrase “Motion by Jason Rolley, seconded by Korey Kenniston” occurs multiple times in each document and thus occurs several times across the topic. Other phrases such as “Carried (4-0)” occur frequently. We can also see in the data that the top four documents that are the most composed of topic 40 occur in chunks 2-5, indicating that

Word_40	Prob_40
appointment	0.031905
teacher	0.028529
name	0.027324
resignation	0.022460
effective	0.018794
carried	0.014834
position	0.014514
school	0.012929
salary	0.012893
monitor	0.012570

Topic 55: Response to Motion

- In this topic there are multiple words related to agreeing with some sort of motion. Motion is the top word and words such as approve, seconded, aye, second and approval are in this topic.
- Evidence: The top 5 documents prevalence of topic 55 all contain approval of motions. It is filled with the speaker going through and asking for a yes or a nay, and, the majority is yes's, which confirms the topic being an approval to motions.

Word_55	Prob_55
motion	0.056733
approve	0.032863
board	0.030769
seconded	0.025797
as_presented	0.017754
aye	0.015881
second	0.013806
motion_made	0.011764
approval	0.011415
minutes	0.009454

Further Validation

If we had further time, we believe that word and topic intrusion would have been great sources of further validation of the model we had selected. Theoretically, we could select some number of our supposedly most interpretable topics for observation. We could then randomly generate a word from the corpus and insert that word into the top five words for each of our topics. Ultimately, we would want to measure the rate at which people with no prior knowledge of our model or of our corpus correctly identify the intruding word. If the rate was high, this would give us evidence that our topics were, indeed, interpretable. We could run a similar validity test at the topic level– we could give volunteers a document, and three topics that our model rated with a high probability along with one intruding topic with a low probability. If the intruding topic is detected at a high rate, we would be sure that our model was producing meaningful and relevant topics. However we do not have the resources available to generate such validation methods and thus we are left with labeling topics manually and using GenSim to look at model coherence.