Developing Natural Language Processing Tools for Sharing of Indigenous and Community Knowledge

Alex Chow
Huaiwen Dong
Yingqiu Kuang

UBC Data Science for Social Good 2019
Data Science Institute
National Energy Board
Boeing Vancouver
CIFAR
Mitacs

We would like to deeply thank the hən̓q̓əmin̓əm̓-speaking xʷməθkʷəy̓əm (Musqueam) and səl̓ilwətaʔɬ (Tsleil-Waututh) peoples, the Sḵwx̱wú7mesh Sníchim-speaking Sḵwx̱wú7mesh Úxwumixw (Squamish Nation) and the hən̓q̓əmin̓əm̓- and Halq'eméylem-speaking Stó:lō peoples on whose ancestral, traditional and unceded territories we write and submit this report.

Acknowledgements

We would like to first thank the indigenous groups who contributed their knowledge at these court hearings, without whom this project would not have generated as many insights as it did. We hope that the analyses we made and the tools we built can help further indigenous knowledge and build towards a more equitable society. In particular, we would also like to thank the Musqueam, Tsleil-Waututh, Squamish Nation, and Stó:lō peoples on whose ancestral, traditional and unceded territories we write and submit this report.

We also thank the Canadian Energy Regulator (CER) (formerly known as the National Energy Board (NEB)) and the University of British Columbia Data Science Institute's (UBC DSI) Data Science for Social Good (DSSG) Program for the opportunity to work on this project. We would also like to further extend our appreciation specifically to Andrew Strople and Ryan Hum, our project coordinators from CER, for their patience and assistance on this project, as well as Dr. Raymond Ng and Dr. Kevin Lin at the UBC DSI for their mentorship and guidance. Additional thanks to the other DSSG research fellows for their feedback and comments throughout the span of this project.

Abstract

Empowering indigenous participation in Canada's energy economy has been atop Ottawa's political agenda in recent years. In collaboration with the Data Science Institute at the University of British Columbia and the National Energy Board of Canada (now restructured as the Canada Energy Regulator), this project seeks to develop and utilize Natural Language Processing (NLP) tools to analyze a dataset that consists of 5,622 public court hearing documents over the past 60 years. Our deliverables are threefold. First, by pre-processing those documents, we extract information on important stakeholders engaged in these hearings, including indigenous communities, oil and gas companies, government agencies, and all other social organizations. We also simultaneously extract and compile fully cleaned conversation files for subsequent, complicated NLP analyses. Second, we focus on two primary NLP tools – topic modelling and sentiment analysis – and apply them to the model training, both at the individual transcript's level and at the aggregate level. In so doing, we derive seven primary themes (categories) and excavate 73 specific, latent topics from the past transcripts. For each transcript, we also identify their dominant category(s) and dominant topic(s). Finally, two web applications are created. They are powerful tools to store and visualize all our findings. More importantly, these two applications are designed with the capabilities to customize and present the findings to meet users' diverse preferences and needs.

Introduction

Energy industry is of critical importance to the Canadian economy. Canada is currently the sixth-largest crude oil producer and the fifth-largest natural gas producer in the world. In 2017, the energy sector made up 9.2% (CAD$ 175 Billion) of the country's Gross Domestic Product. Including indirect jobs, the energy sector creates 900,000 jobs, employing an estimated 4.9% of the national workforce. Meanwhile, Canada is also a large net exporter of energy. Energy products made up 17% of Canada's total exports and were valued at CAD$ 71.4 Billion in 2017.

However, the competitiveness of oil and gas industries in the country has been shattered by growing public concerns over the recent years. On one hand, pipelines have become a global focus for climate activists. According to Environment Canada, oil and gas account for the largest source of Canada's greenhouse gas emissions. On the other hand, throughout years, prices in energy markets are highly sensitive to fluctuations in demand and supply, and the continued volatility has caused significant unpredictability. Canadians need informed, excellent, and engaged regulations to ensure their safety, protect their lands and communities, prevent market inefficiencies, and most importantly, reflect their interests in decision-making. The most recent federal Bill C-69, introduced in early 2018, made it mandatory to weigh climate changes as a factor in any proposed pipeline project. It also widened public participation in the review process and imposed more requirements for consulting indigenous communities affected by pipeline construction.

The National Energy Board of Canada (NEB), now the Canada Energy Regulator (CER), has been playing a vital role in providing those public goods. Since 1959, the NEB has regulated over 73,000 pipelines across Canada and 1,462 kilometers of international power lines. The important work under their mandate includes oversight of pipeline construction and environmental protection, damage prevention and emergency response, expanded offering of energy information (markets and supply, sources of energy), and adjudication of applications before the Board. Furthermore, they strongly believe in the importance of listening to and understanding indigenous knowledge as a fundamental aspect of safety and environmental protection. The inclusion of stakeholders and indigenous peoples ensures engagement, transparency, as well as unbiased and accessible performance in their decision-making processes. Therefore, for any application for a major pipeline or power line project, the NEB holds a public hearing before a decision is made. Public court hearings allow participants – including the company proposing the project, directly affected persons, and other persons with relevant information or expertise – to express their point of view and present evidence for or against the project. The NEB aims to collect all the information it needs to make a transparent, fair, and objective recommendation or decision.

The NEB's relentless effort towards greater inclusivity in its decision-making process helped create a treasure trove of texts. Over the past sixty years, the court hearings have produced 5,622 court hearing documents and the organization has been attempting to compile, categorize, and publicize them in REGDOCs, the NEB's online document repository. Their ambitious aims, however, come with challenges. It is virtually impossible for a human to read through these tens of thousands of pages of documents; nor is it feasible to distil shared concerns from the various indigenous representations by traditional textual analysis alone.

In this project, we turn to Natural Language Processing (NLP) tools. We have taken advantage of the latest developments in such new digital prowess to find new and efficient solutions. This project is hoping to answer the following questions:

- Over the past sixty years, who are the active participants in public hearings held by the NEB?
- What are the stakeholders' primary interests and major concerns in these public hearings?
- What are their attitudes towards these issues discussed in public hearings and how have their perceptions evolved over time?
- How can we preserve and reconstruct the stories of indigenous communities across Canada?

The following report consists of four sections. The first section introduces the dataset and our approach to pre-process those documents. The second section offers a brief overview of some theoretical foundations supporting NLP tools, especially in topic modelling and sentiment analysis, and explains how these tools are applied to our model training. The third section discusses the construction of the two web applications and elaborates on the usage of these interactive platforms to present customized findings from our models. The final section concludes with some interesting patterns shown in our deliverables and some suggestions for our future work.

Description of Dataset

The dataset consists of 5,622 public court hearing documents from 1959-2018. Each year, anywhere between one to ten different projects are brought before the board. Court hearings are carried out for the project proposer and various stakeholders to provide testimony for or against the project; this testimony might include the potential benefits and risks of the project, cost-benefit analyses, risk assessments, and proposed solutions to mitigate assessed risks, among others. Court hearings for each project may be completed in a day or span several days. Regardless, the transcripts from these hearings usually span multiple transcripts, each of which may be anywhere between 30 to 200 pages long. These transcripts make up the bulk of the documents in this dataset (approximately 5,100 of the documents are transcripts from the past 60 years) and are the focus of this project. The remaining documents include letters and other written-in testimony to these court hearings, which are in-turn brought up in the transcripts. Due to their less structured form and shorter content, we exclude these documents from our analysis and focus on the hearing transcripts instead.

Each hearing transcript begins with a cover page, followed by a participant list, evidence/testimony list, and then the transcript of the hearing proper. The metadata provided at the start of the transcript is relevant for us to extract information about each hearing such as when the hearing was held, who the participants were and what organizations they were representing, and what indigenous groups were represented.

Data Cleaning and Pre-Processing

Through the process of creating a prototype for preliminary analysis of this dataset, we chose to focus on a smaller subset of the data to improve processing and analysis times. As such, we further narrowed our focus to the hearing transcripts from the years 1994-2018 for the following reasons: First, our preliminary conversations with NEB revealed that 1994 was when indigenous groups were first represented in court proceedings. Thus, we expect that the

data from years after 1994 to be more congruent with our aim of identifying indigenous concerns and stories. Second, our preliminary data exploration revealed that the data from earlier years (particularly between the 1960s and 1970s) required more pre-processing before it was ready for text analysis. Particularly, many of these older transcripts were scanned and of poorer quality visually and as such, could not be easily read by our programs without Optical Character Recognition (OCR) software, which would have further increased our processing times. Third, the format in which participant lists were compiled varied across different years. By limiting our focus to just the later years, we were able to isolate only three different formats in which these participant lists were recorded, which made for easier extraction by both isolating the participant lists from the rest of the transcript, as well as separating the different participants and organizations into separate objects. Here, we trained a simple model to recognize which one of the three formats the participant list was recorded in, and the model would then proceed to use the appropriate functions for the identified format to separate out the participants and organizations.

Our subsample consists of 1,516 hearing transcripts from 25 years (1994-2018). These transcripts were used for our preliminary analysis of participants and organizations over the time period. We also modelled the latent topics, conducted the sentiment analysis, and visualized our findings on these transcripts. Prior to this analysis, we processed the data in four steps. First, we isolated the participants list and the transcript, removing data that was irrelevant to our analysis, such as the cover page, evidence list, and miscellaneous information. Second, we identified which participants and organizations were represented in each transcript. As mentioned earlier, since the participant list tended to follow one of three different formats based on the year that the data was collected, we simply had to write code that would recognize which of the three formats the participants list was in and follow up by identifying the participants and their representative organizations in the relevant format. Indigenous organizations and communities were also identified using a dictionary of indigenous groups as well as a keyword search. This information was saved in an excel sheet with each row representing a transcript, and columns that contained lists of participant-organization pairs, individual participants, total organizations, non-indigenous organizations, and indigenous organizations.

Third, we processed each hearing transcript by reading each transcript into an individual excel sheet. We removed descriptions of the court hearings (such as mentions of "commencement of the hearing" or "breaks") and extracted only the speaker turns (i.e. each uninterrupted statement from an individual). The final product of this step was that each hearing transcript, originally in a pdf format, was converted to an individual excel file with two columns: the speaker and their respective line of dialogue. Each row represented one speaker turn in chronological order. We would later apply the topic modelling and sentiment analysis, trained on the entire dataset, on each of these transcripts to identify the latent topics and sentiments in each transcript and how they have changed over time. We were also then able to match these topics and sentiments with the speakers in each transcript, who were then matched with the participant list from step two to identify trends for organizations over time.

Finally, we combined all the hearing transcript data from step three into one large excel sheet. Each row continued to represent a speaker turn, but chronological order no longer mattered since the file contained all transcripts from 1994-2018. The purpose of doing so was to create the large database required to get a sense of what the common latent topics were in the dataset. Using this large excel sheet, we then trained the LDA model that was later applied back to each individual transcript. The next section will delve into the step-by-step approach to this model training.

NLP Tools on Model Training: LDA Modelling and Sentimental Analysis

Natural Language Processing has been experiencing rapid growth over the past two decades. Technologies and tools based on NLP have become increasingly widespread. Those simple programming techniques not only enable us to automatically extract key words and phrases that sum up the style and content of an unstructured text; they also help to extract meaningful patterns and actionable insights from large quantities of raw textual data. The application of NLP techniques in this project focuses on two primary tools: (unsupervised) Latent Dirichlet Allocation (LDA) modelling and (unsupervised) sentimental analysis by VADER Lexicon.

Latent Dirichlet Allocation (LDA) modelling

In text analytics, topic modelling has been widely adopted to extract various diverse concepts or topics present in the documents. There are several popular topic modelling approaches in the field, including Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) modelling, and Nonnegative Matrix Factorization (NMF) modelling, each of which involves using specific statistical and mathematical techniques. LDA modelling, often known as a probabilistic statistical model, relies upon latent Dirichlet allocation to discover connected latent semantic structures in text data that yield topics and concepts. Given our limited human capital and time, we chose to focus on LDA modelling – in particular, the unsupervised LDA modelling approach – to extract and collect as much information at this preliminary stage.

LDA models in this project are trained in two different levels. The first level of analysis is to extract the key topics in each hearing transcript. We choose to train our models on each of the individual transcripts; in each model, we first manually decide the number of topics present in the single transcript and then look for the optimal number of topics through multiple rounds of experiments. To improve the accuracy, each topic generated in the model comes with ten keywords and the most representable topic sentence. (For more information for what this model looks like, see section 1.1 LDA Topic Modelling Visualization)

The second level of analysis, by contrast, applies LDA models on an aggregated level. We adopt an innovative approach -- a two-step LDA modelling, in order to find the common topics that tie all the transcripts together. The workflow in the aggregated-level LDA modelling is shown in **Figure 1.**

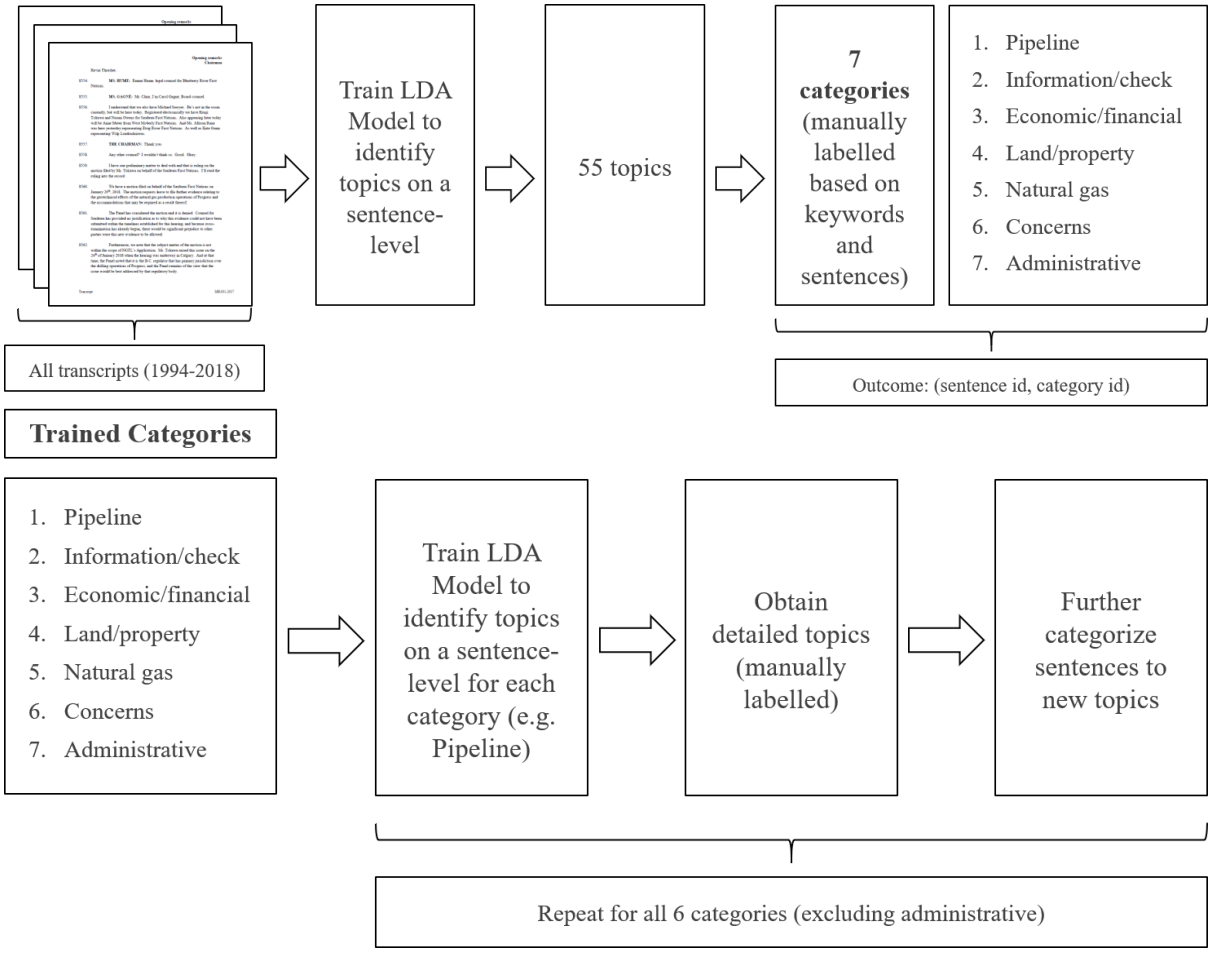

Figure 1. Workflow of aggregated-level LDA modelling

In the first step, we train the model on all the transcripts that we have. The modelling process is identical to our workflow for the individual level of analysis, we are able to derive 55 themes. After rounds of trials and experiments, we believe that 55 is a small enough scale, so we analyze them and group them into 7 primary categories. We also incorporate coherence measures in topic modelling techniques to further distinguish between good and bad estimates of themes in the model (for coherence score graphs for our models, please see Figures A1 – A7 in Appendix). This process involves much intellectual labour, as we have to read through the key words and the top one to two hundred key sentences to determine how best to categorize each of the 55 themes.

The outcome of our analysis is displayed in **Table 1**. Having researched relevant scholarship, we believe that the seven categories nicely summarizes the key aspects of the NEB's daily duties and responsibilities. Pipeline construction and natural gas development have to do with infrastructural projects. Economic and financial matters encapsulate the work related to project developers and the economic implications of development. Land and property are important aspects of pipeline projects and the area development they entail. Concerns aggregates the comments and questions that the parties involved raised during the hearings. Information includes the analysis, study, and assessment that are critical to the resolution of the concerns. Administrative procedure is also found a frequent topic. It covers the procedural dimensions of the hearings. At this point, our model assigns each sentence in all the transcripts to a dominant category.

pipeline	information_check	economic_financial_matters	land_property	natural_gas	concerns
1. risk scenario 2. price, credits, and rates 3. reserve areas and lands 4. oil spill and environment 5. water and life 6. community and people 7. regulations 8. construction: equipment 9. community engagements 10. construction: assumption, consideration, requirements 11. markets and pipeline industry 12. forecast: extension, facilities expansion 13. contractor and company business 14. pipeline system and work 15. shipper, toll, and transportation 16. pipeline routes/location 17. construction: application 18. assessment: alternatives and data	1. measurement check 2. local plans and specie risk analysis 3. pipeline safety management and water 4. concerns: land, community and traditional knowledge 5. board jurisdiction 6. evidence accuracy check 7. words/arguments clarification check 8. discussion/fact/report/document check 9. service check: principle and regulation 10. question check 11. data check: market/capital, toll, cost, price 12. plant and effect on nature	1. customer needs and requirements 2. firm service: contract and shipper 3. toll and cost 4. cost-of-service rate and standards 5. project application and board decision 6. service evaluation: information, different views, and case comparison 7. supply: market, competition, and forecast 8. capital and investment return 9. firm service: volume and capacity use 10. firm service: plant and facility 11. Proposed change and result 12. supporting evidence and data 13. pipeline project and cost	1. land, property, and caribou protection 2. field production and development 3. traditional use studies 4. winter scenario: production, price and ecosystem 5. spill, watercourse, and fisheries 6. production in area: water and fish 7. pipeline design: area assessment, and environment commitment 8. area identification: pipeline routes and habitat 9. pipeline running area, path, and infrastructure 10. area study: emission, specie and community life 11. harvest and watercourse 12. pipeline design in area: pressure, toll, and system design 13. future market and area development	1. producer and contractor 2. plants and production capacity 3. gas flow, market, and costs 4. system design and utilization	1. market and price determination 2. treaty rights and environment 3. financial risk and return 4. learning traditional oral knowledge 5. people and local community 6. economic opportunity: expansion and capacity 7. assessment: chemicals, emission, and water 8. social-economic assessment: consideration and impacts 9. issues by landowners and shippers 10. comments on firm service: toll, assets, transportation 11. pipeline construction: issues, opinion, clarification, position, uncertainty 12. consultation: process and scope 13. impact assessment: methodology and company-led

Table 1: General categories and specific themes derived from aggregate-level LDA modelling

In order to further extract more specific topics under each category, we further train the model on all sentences under each of the seven single categories. This allowed us to generate new sub-themes, and we manually group these sub-themes into topics and assigned topic labels for each of these new topics. Our repeated model training within each of the six major categories, excluding the category of administrative procedure gave us much more substantive topics than the 55 we had when we ran the model across all, unfiltered sentences. As Table 4 shows, we manage to get a total of 73 topics. They include: 18 specific topics under pipeline, 12 under information check, and 13 each under land and property, concerns, and economic and financial matters. We also derive 4 substantial topics under natural gas. Compared to the 55 topics we had in step one, these topics are much more concrete and coherent under each category. Each sentence in all the transcripts is assigned to a dominant category and a dominant topic under its own category.

Sentiment Analysis by VADER Lexicon

This project also deploys VADER lexicon, an unsupervised technique in text analytics to predict the sentiment present in our transcripts. VADER stands for Valence Aware Dictionary and sEntiment Reasoner. It is a specially curated lexicon with a rule-based sentiment analysis framework that was built for analyzing sentiment from social media resources. The VADER lexicon includes a list of positive and negative polar words with the associated scores. We then assign scores to the text documents to compute the sentiment.

Similar to our approach in topic modelling, we also conduct sentiment analysis in two different levels. For single hearing, we choose to compute the compound sentiment score in every sentence of the transcript and categorize it into three kinds of sentiment: positive, concern, and neutral. In the meantime, at an aggregated level, we seek to summarize the sentiment of stakeholders, especially indigenous communities across Canada. This is our proposed solution to comprehend their primary interests and major concerns over pipeline construction and other important infrastructural projects. To do this, we first calculate the sentiment score for each sentence in all the transcripts; by matching sentences with their speaker information, we then calculate the proportion of such sentiment in each of the stakeholder groups.

These two primary NLP tools have helped us derive interesting results. First, each sentence in our entire sample is assigned a category and topic. Second, by calculating the distribution of the topics and categories in each transcript, we develop an understanding of the key themes in each transcript. Third, we align the timing of each transcript. This allows us to see how the categories and topics are temporally distributed on the timeline, and how the concerns evolved; Fourth, we match sentences with their speaker information, therefore linking the themes to the indigenous communities to capture their narratives and concerns. Finally, the outcomes are socially relevant at a larger level: if we compare the thematic patterns in these transcripts with the politics and institutional changes in reality, we may be able to see the impact of policy changes.

Data Visualization

1. Individual Transcript Visualizer

1.1 LDA Topic Modelling Visualization

A conversation data file which can be either a pre-formatted conversation data or a plain text file copied from a pdf file is required to be uploaded for analysis. The former format would be an excel sheet as described in our pre-processing stage, while the latter would be a plain text file that when entered into our application, converts the data into the former format. The number of topics identified in the LDA model is decided by the user since each transcript may have different content structure. A table with an overview of keywords and the number of sentences in each topic is available. Two visualization graphs assist with the judgement on the number of topics parameter. The size of the circles in the topic distribution chart reflects the number of sentences in each topic, and the distance between circles is a two-dimensional projection of pairwise Hellinger distance[1] between probability distributions of topics. The topic progress graph clusters the conversation progress into topics. Generally, a good model will have a well-spread distribution chart and a well-clustered progress graph. Sentiment graph, word frequency graph, and sentence filter are also available.

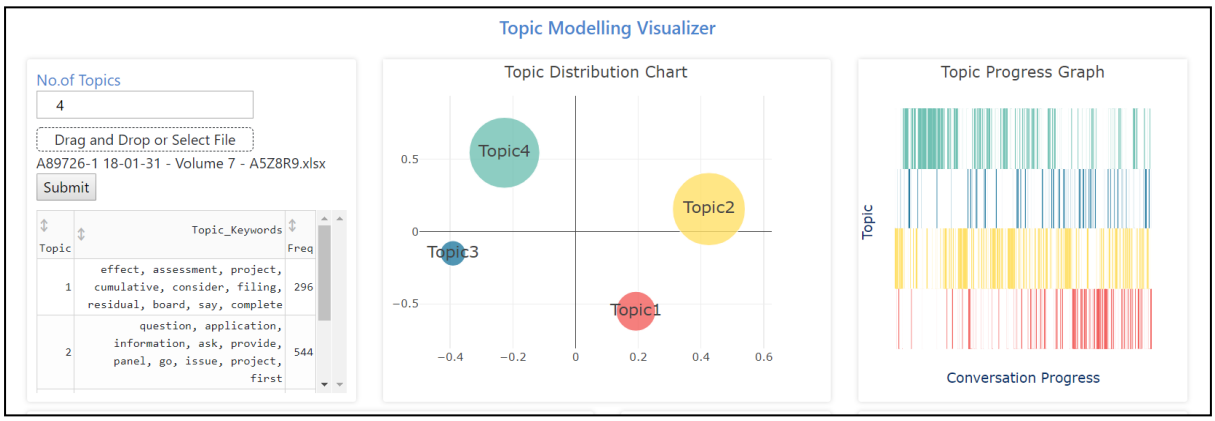

Figure 1: User Interface of Single Transcript Visualizer: Summary Table, Topic Distribution Chart and Topic Progress Graph

1.2 Frequency Bar Charts and Vader Sentiment Analysis

In order to explore details in each topic, three different frequency bar charts are produced:

The word frequency bar chart illustrates the top 10 most commonly appeared words in each topic. The overall frequency in the entire conversation (grey portion) and topic-specific frequency (colored portion) are contrasted for each word. Generally, a representative word will have more colored portion in its bar which corresponds with higher frequency in one topic compared with other topics.

The most frequent speakers for each topic can be generated as well. For each topic, sentences from a particular speaker are aggregated, the general sentiment distribution of this collection of sentences is analyzed by Vader sentiment analysis method which is built on a

[1] Similarity between two distributions: $H^2(P,Q) = \frac{1}{2}\int\left(\sqrt{dP} - \sqrt{dQ}\right)^2$.

dictionary with labelled emotional intention and strength of more than 7000+ English words. The methods developed a scoring system based on the dictionary, use of punctuations, capitalization of words and property of connecting words (Hutto & Gilbert 2014). According to the compound score calculated from this scheme, we labelled each sentence as support (score > 0.05), neutral (-0.05<score<0.05) and concerns (score < -0.05).

Finally, an indigenous group frequency chart for each topic is used for reference purpose. The rationale behind this is similar to the word frequency charts.

1.3 Sentence filter

As results from LDA models have limited explanatory power, original sentences from the transcript might be needed to make comprehensive conclusions. A sentence filter with fields in key word, speaker and indigenous group realizes this function.

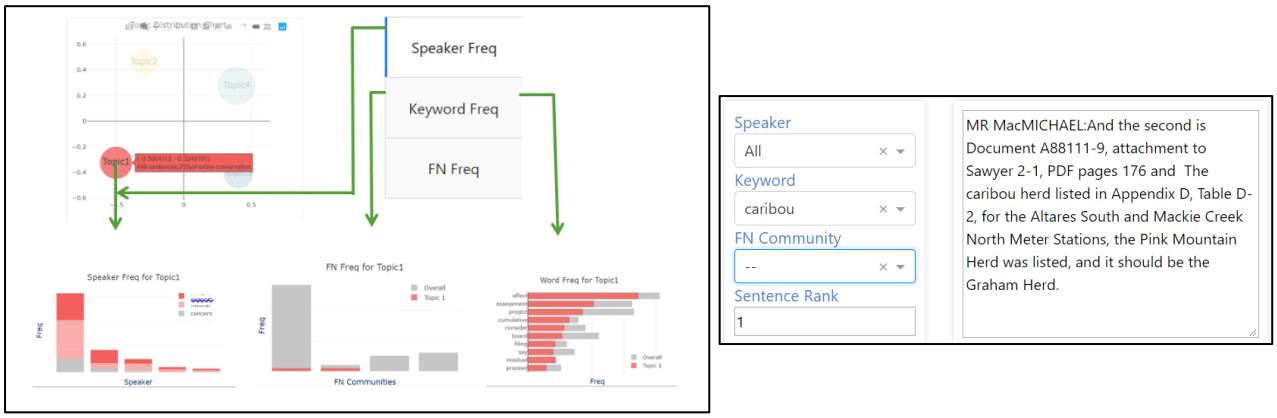

Figure 2: User Interface of Single Transcript Visualizer: Sentiment Graph, Word Frequency Graph and Content Filter

2. Multiple Transcript Visualizer

2.1 Transcript Name Filter

The multiple transcript visualizer begins with a file name filter with fields in year range, transcript type, company and organization and indigenous group as well as topic categories and subtopics generated from the two-step LDA topic model. A map of the geographical location of the majority of indigenous groups can be used for reference.

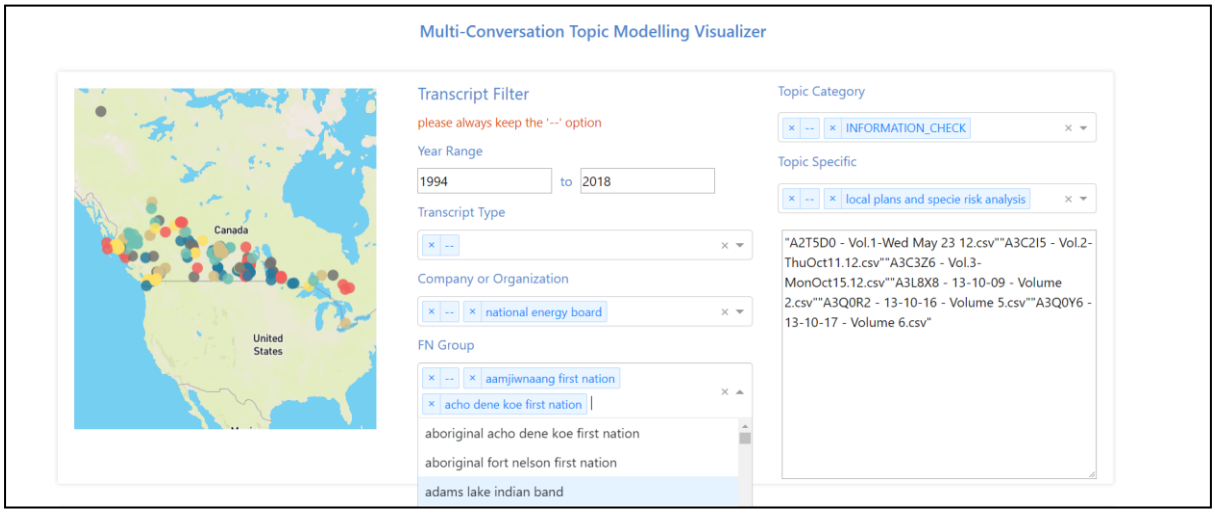

Figure 3 Map and Transcript Filter

2.2 Story Teller

Multiple transcripts can be uploaded, and relevant indigenous groups will be identified. The application will then produce the conversation contents from corresponding speakers which users can navigated through and download.

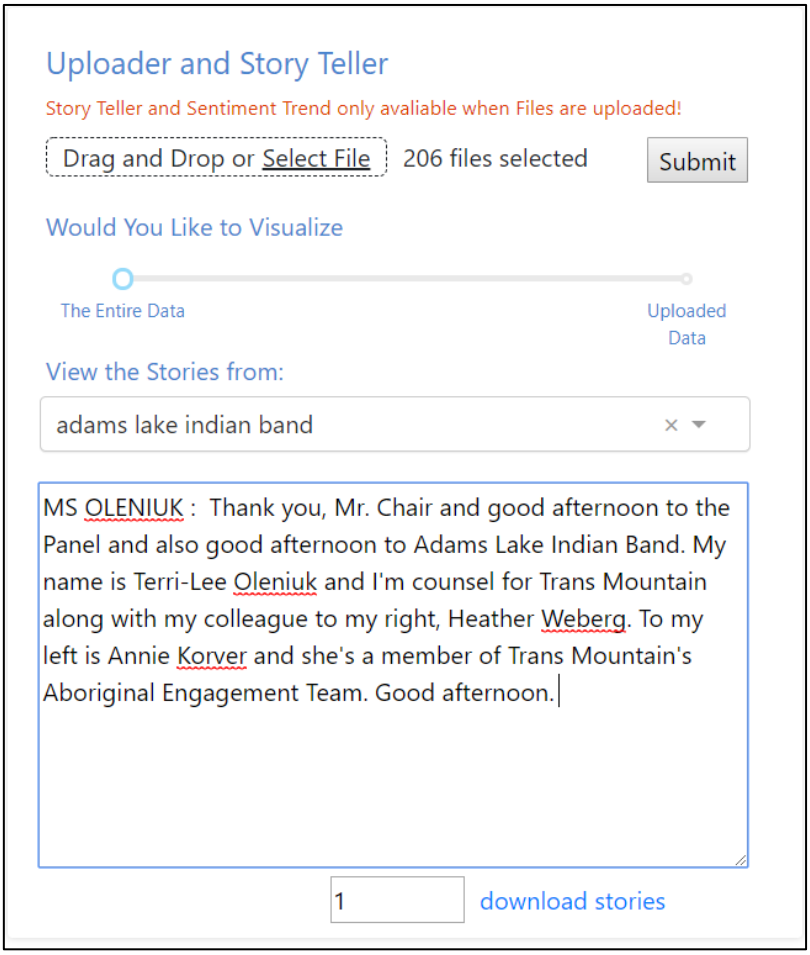

Figure 4 Uploader and Story Teller

2.3 Topic Frequency Time Series and Organization Specific Analysis

Multiple time series line charts for the frequencies of the 55 topics in 6 categories can be searched with any combination of topics by selecting topics in a dropdown menu as shown in Figure 5.

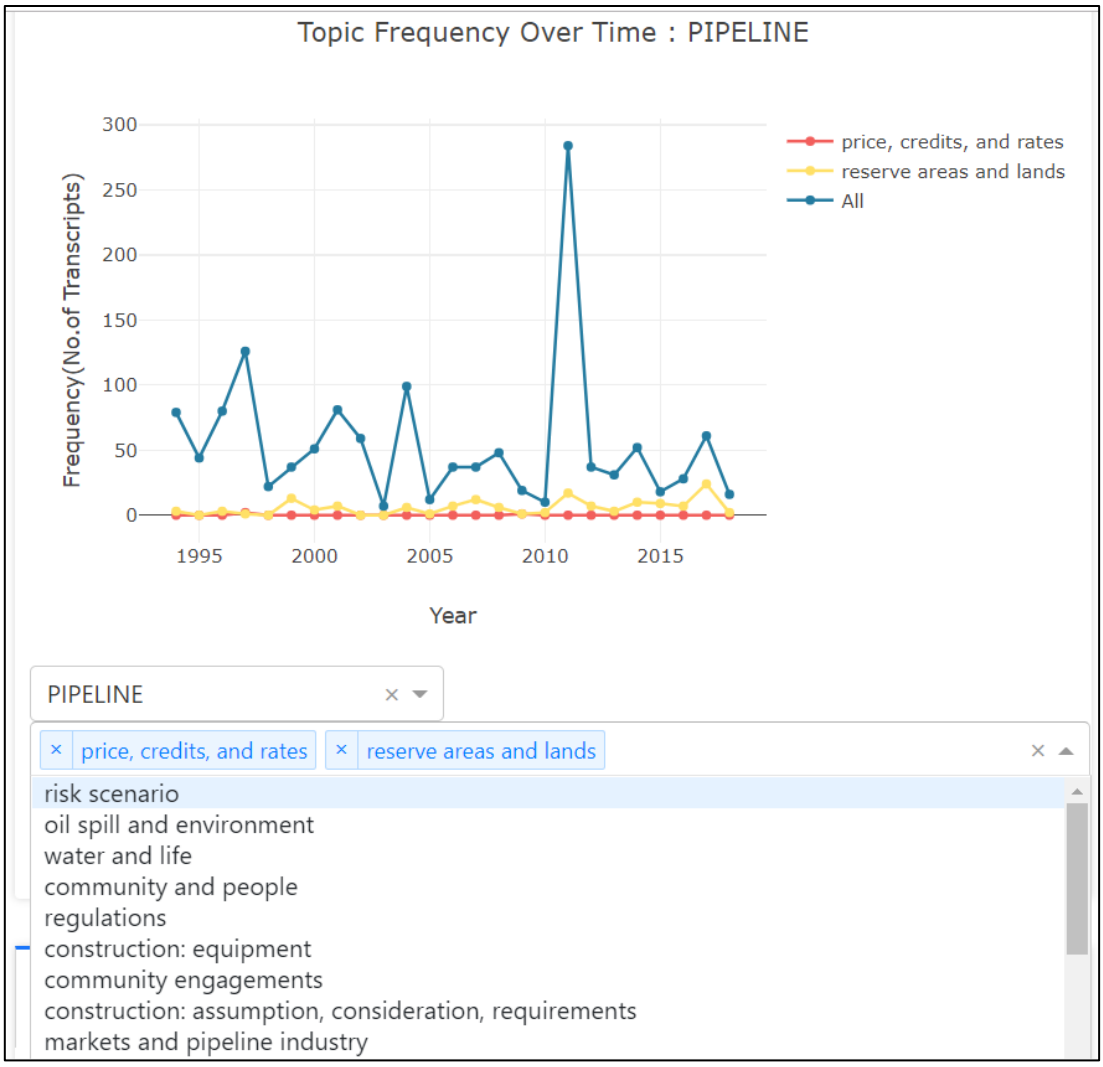

Figure 5 Topic Frequency Time Series

Organizations are separated as indigenous groups and other organizations, and overall frequencies of the appearance of topic 15 organizations in both types are visualized in bar charts. The remaining organizations can be viewed in the dropdown menus.

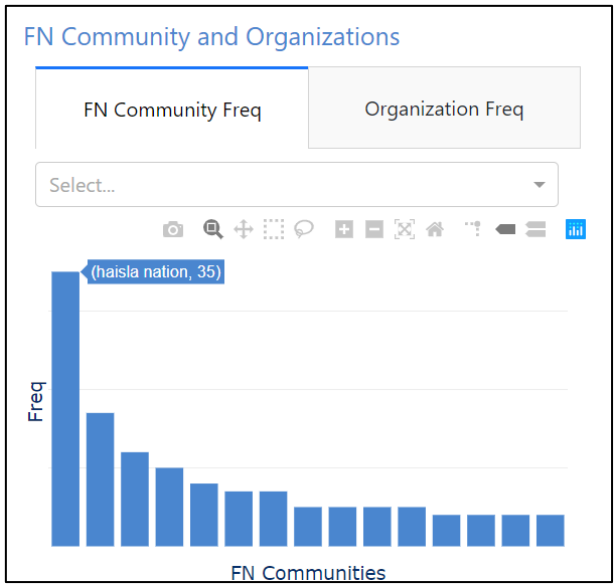

Figure 6 Organizations Frequency

Topic distribution time series and Vader sentiment analysis stack bar charts for each organization are available as shown in Figure 7. The Vader sentiment analysis unit here is speaker turns instead of sentences.

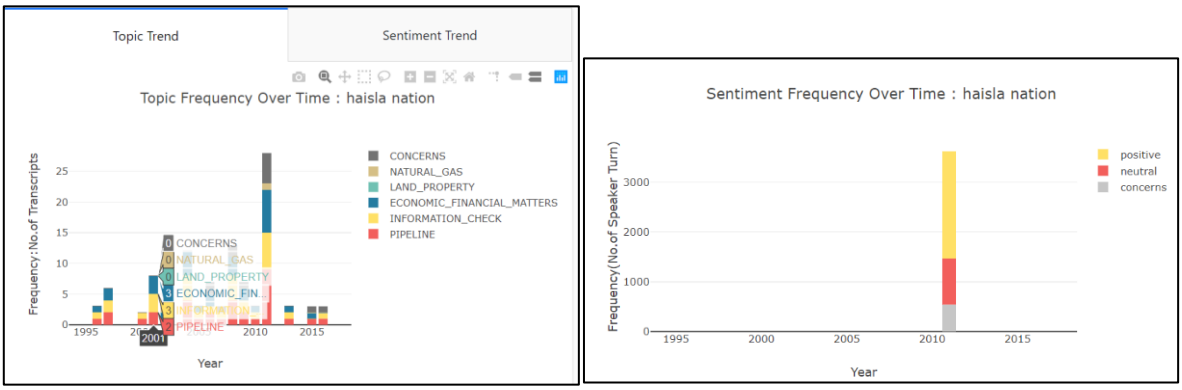

Figure 7 Topic frequency and sentiment analysis over time for individual organization

Conclusion

NLP provides a promising future for processing large quantities of text data. We hope that this prototype has provided a glimpse into the potential of computational methods for not only processing data that would take incredibly large amounts of human labor to do through conventional means, but also making sense of them in ways that no single human can do due to the sheer size of the data. With the right NLP tools, we are able to generate these meaningful summaries and themes of this enormous text database rich in information and potential. Furthermore, by combining these with the right user-friendly applications, anybody with an internet connection will be able to access this powerful database. This includes both the decision-makers at NEB and public members interested learning more about pertinent social topics today such as energy production or indigenous issues.

The topics that we have modelled by transcripts, participants, organizations, and across time will help improve the workflow of the different teams at NEB and in turn, improve the efficiency of their decision-making process. An example of how the applications will be incorporated into NEB decision-making might look like this: a new project might be proposed on certain indigenous lands. Using the multi-transcript visualizer, NEB may now filter out transcripts that only include the indigenous communities whose lands the proposed project may be carried out on. They will be able to quickly view, on a macro-level, how many times these groups have been represented in court hearings and in turn, what their main concerns (themes) have been. NEB will also be able to see how these concerns have changed over time, if at all. Perhaps in a particular year, there was a spike in indigenous concerns over the risk of an oil spill. NEB may then select a transcript from that particular year and enter it into the individual transcript visualizer. Here, they have a more in-depth look into the concerns from that particular project. They can also get a better sense of the sentiments of the indigenous groups using the sentiment analysis chart. By identifying the specific concerns and general sentiment of the indigenous groups, NEB will have the information they need when they go on the ground to consult these groups. From our conversations and consultations with the various teams at NEB, this is the most promising potential of the applications we have developed – they will no longer be going into indigenous communities without any prior knowledge that previously would have been lost within thousands of pages of transcripts. They will now be able to identify specific concerns that have been brought up in the past, and check with indigenous communities if these concerns have been met. If so, this gives NEB more confidence in approving the projects, and if not, it gives them the information necessary to hold the relevant organizations responsible.

Future Work

Our two web applications are now under further development at NEB. The data team at NEB will be able to make improvements to the existing user interface of the two web applications to make it ready first for internal use, and in the future, for public use. NEB will also be able to pre-process the older transcripts and either allocate the themes and topics that we generated from the smaller subset of data to these transcripts, or, for a more holistic and complete analysis, repeat the workflow that we have described in this report to retrain the LDA topic modelling on the full transcript database and allocate the new themes and topics to all the transcripts.

Due to the limited timeline of this project, we have also only scratched the surface of the potential of NLP tools. NEB could also rely on and develop more complex NLP tools, such as the previously mentioned LSA and NMF topic modelling techniques. We expect that

experimentation and exploration into these more advanced NLP tools will bring about huge benefits for better understanding this rich database of information.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. "Indexing by latent semantic analysis." *Journal of the American society for information science* 41, no. 6 (1990): 391-407.

Hutto, Clayton J., and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In *Eighth international AAAI conference on weblogs and social media*. 2014.

Appendix

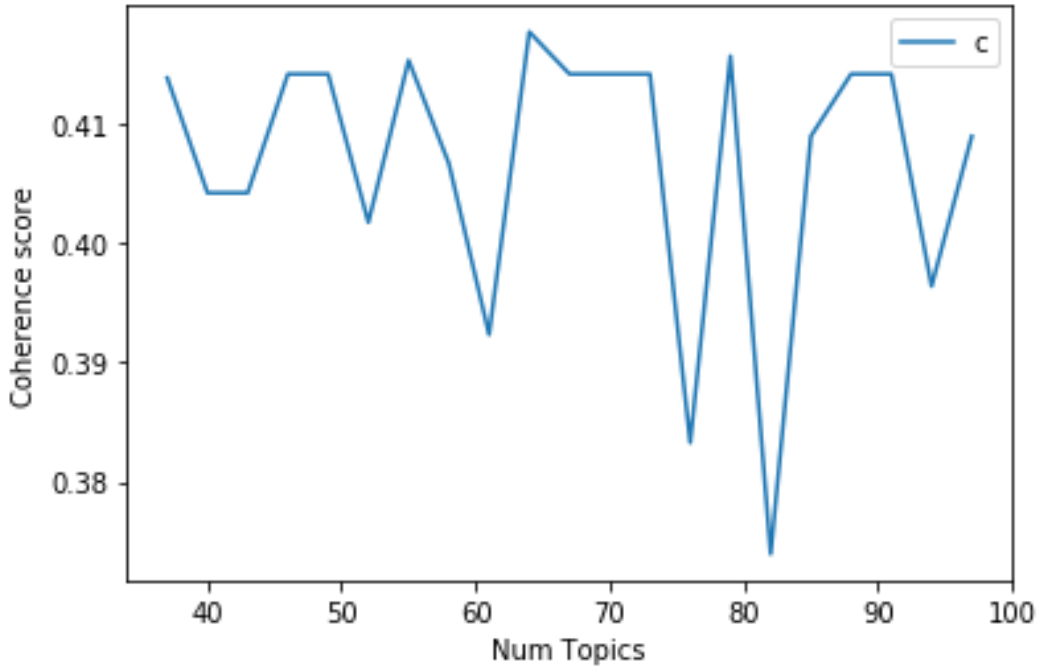

Figure A1: Coherence Score Graph: Initial 55 Topics

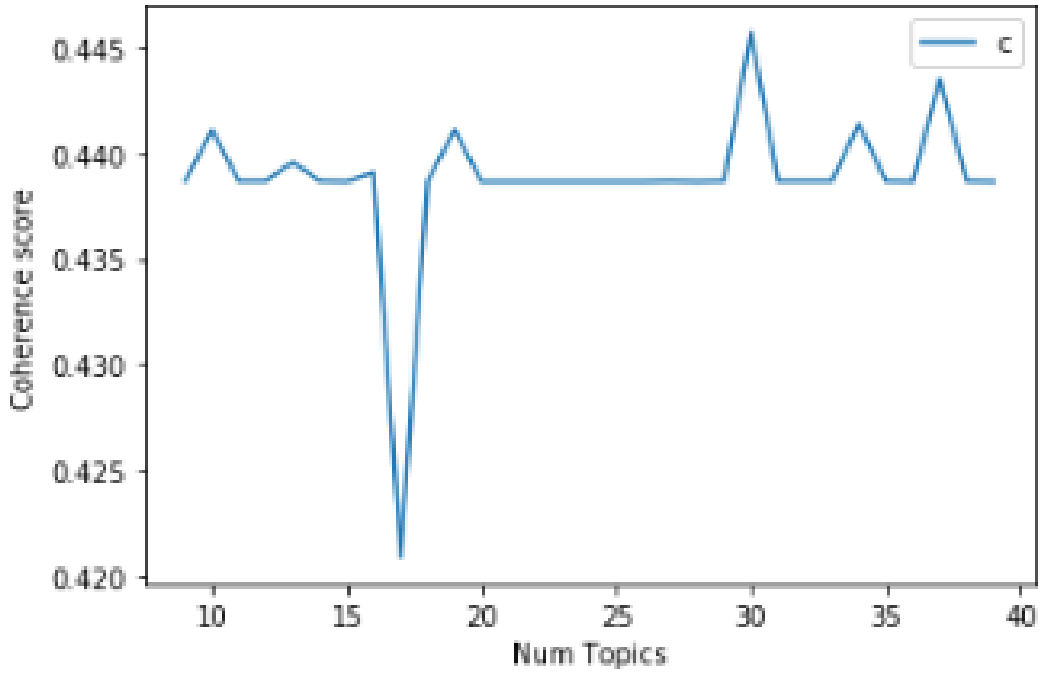

Figure A2: Coherence Score Graph: pipelines

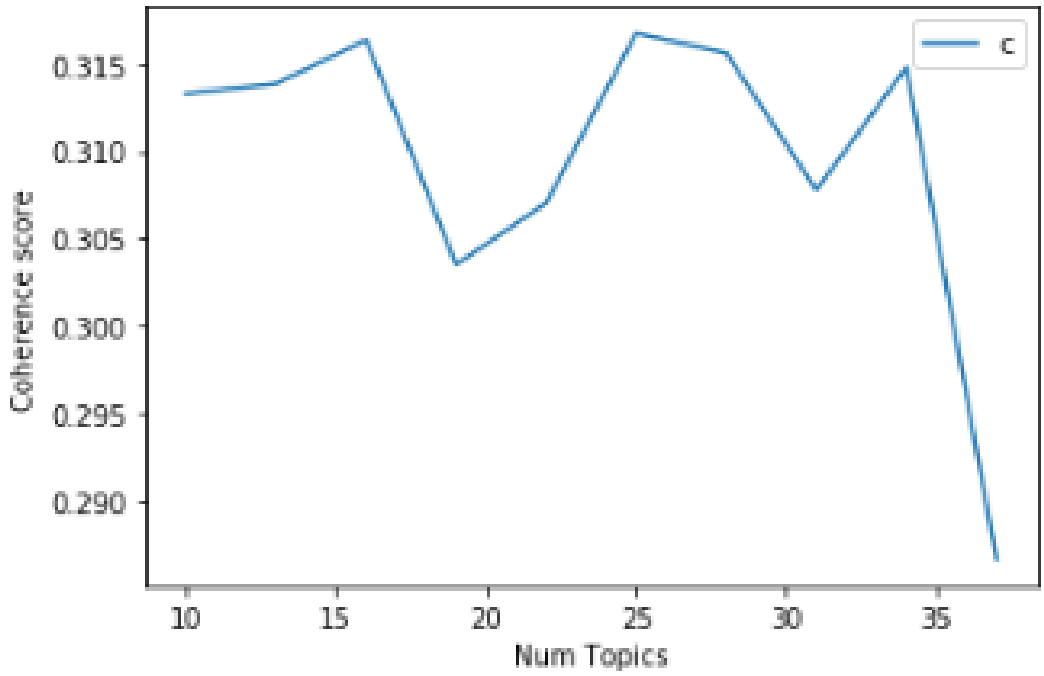

Figure A3: Coherence Score Graph: information_check

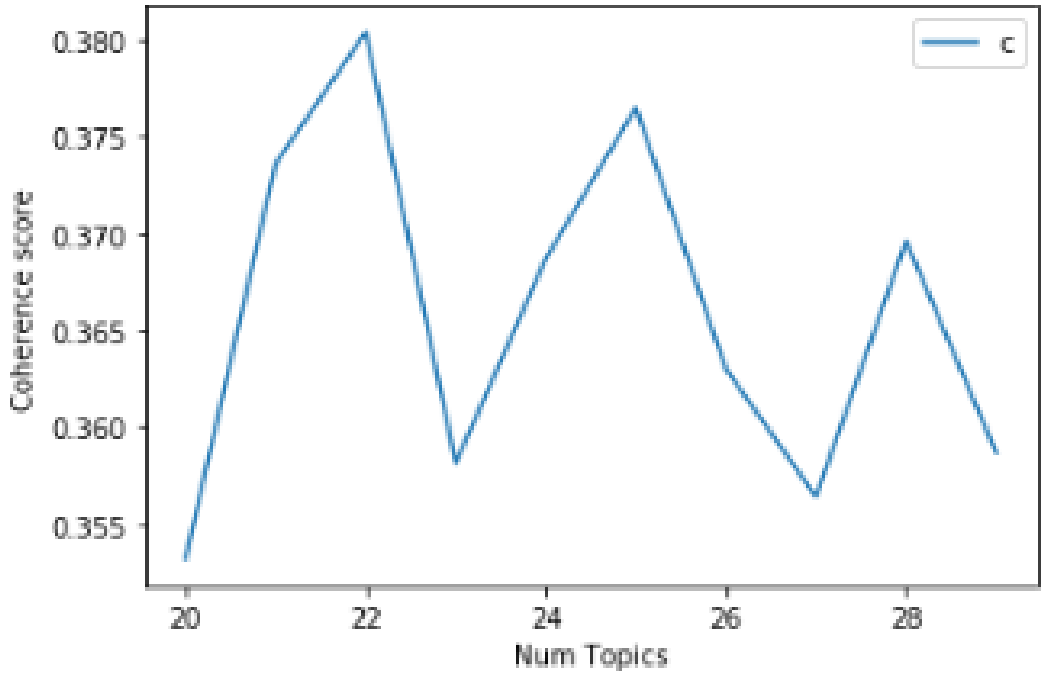

Figure A4: Coherence Score Graph: economic_financial_matters

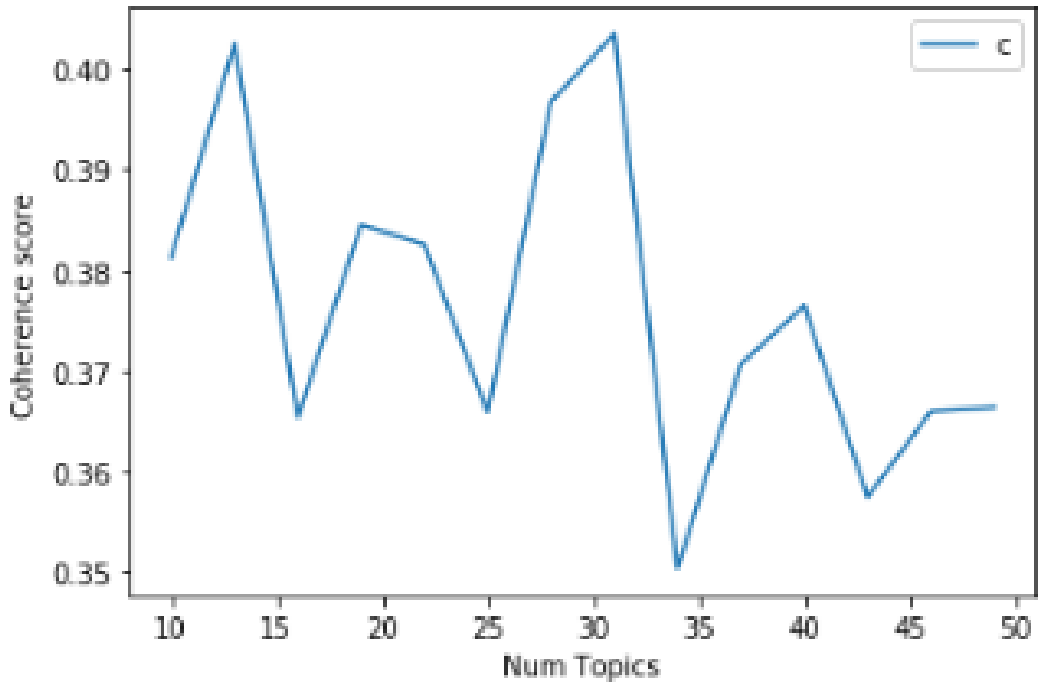

Figure A5: Coherence Score Graph: land_property

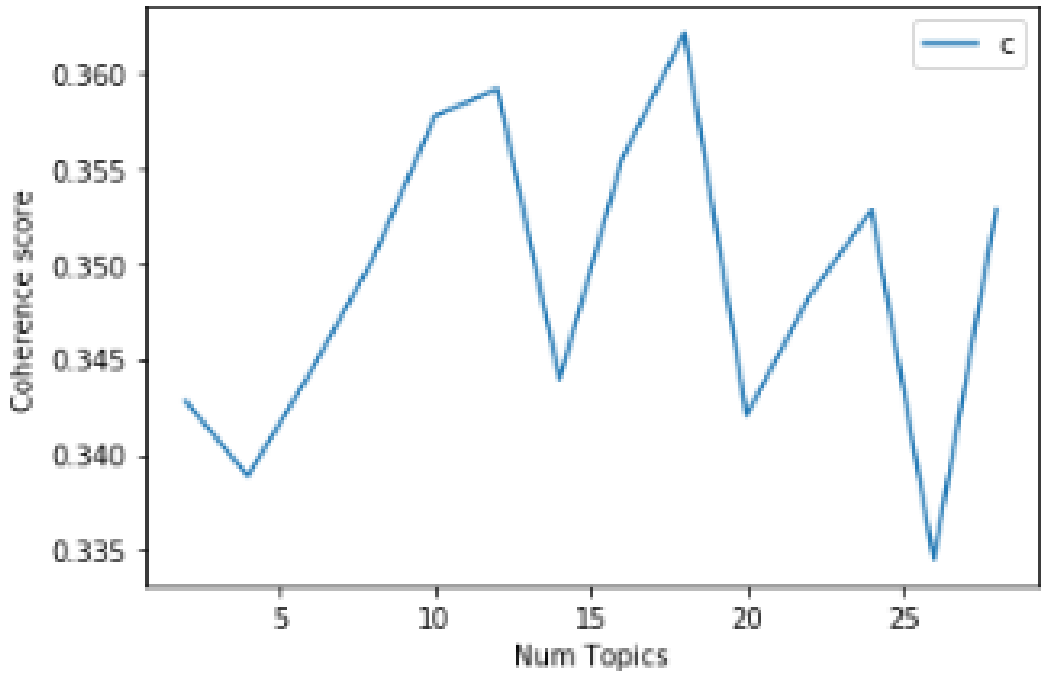

Figure A6: Coherence Score Graph: natural_gas

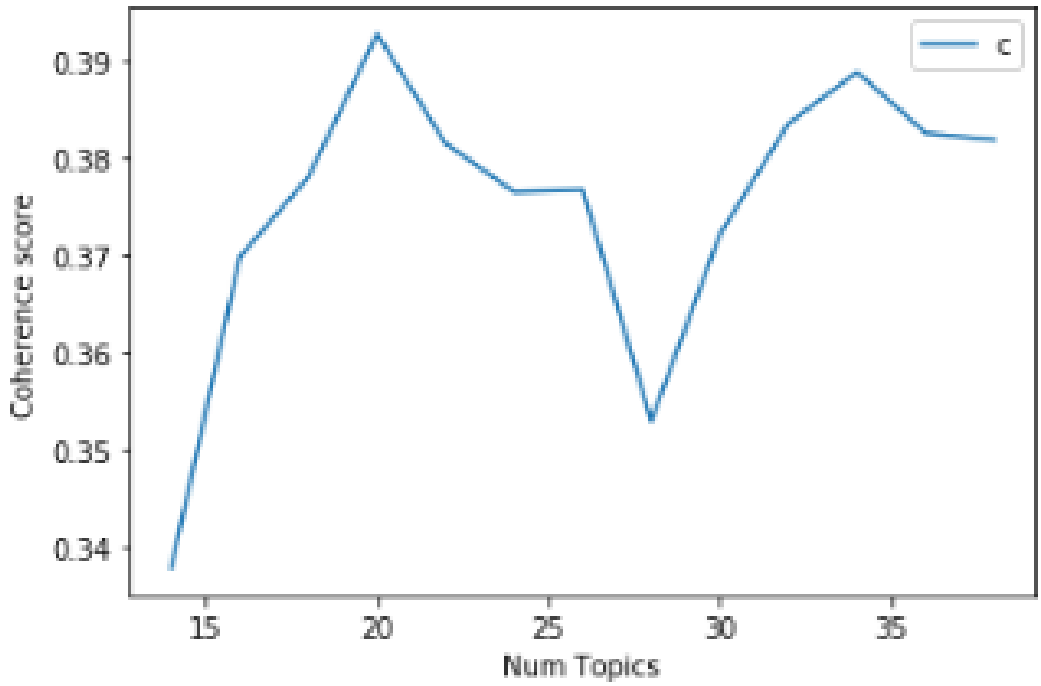

Figure A7: Coherence Score Graph: concerns