Towards a Clear Understanding of Canadian Internet Policy Consultations

cybera

cybera.ca | @cybera

Executive Summary

The CRTC's regulatory proceedings can often provide important insights into how Canadians are accessing the internet. Materials submitted to the CRTC are made publicly available but with the current size and structure of the CRTC's database, these files are difficult to locate, aggregate, and analyze. This makes it cumbersome for anyone looking through these files to understand the key positions each party is taking, and what their impact is on the CRTC's regulatory decisions.

In this report, Cybera describes its efforts to provide a framework and application for aggregating, organizing, and analyzing the vast amount of data submitted to the CRTC. Working with the particularly rich source of documents submitted to the consultation on "Basic Telecommunications Services (2015-134),", Cybera has built a proof-of-concept policy browser. Our goal is to make CRTC consultations more digestible to all Canadians. The policy browser is available at policy-browser.data.cybera.ca along with all our scripts and methods.

Highlights of the project include:

- Developing a proof-of-concept and publicly available policy browser that allows users to browse and search documents submitted to the CRTC 2015-134 consultation
- Making available an open source framework that can be adopted by anyone to analyze additional CRTC consultations. This framework aggregates, processes, and imports submitted documents into a graph database. The database can then be used to support a local and customizable version of the policy browser.
- Running a proof-of-concept analysis of the CRTC 2015-134 consultation using various natural language processing techniques. Through this analysis, Cybera identified patterns in the language used by various intervenor categories, and also compared the submissions of individuals who participated in the consultation. For example, advocacy organizations discussed affordability from the perspective of Canadian consumers, while telecommunications companies discussed affordability from the perspective of running a business. Individual submissions tended to focus on more personal topics, such as the impact on daily needs and personal circumstances.

Introduction

Cybera has participated in two Canadian Radio-television and Telecommunications Commission (CRTC) public consultations: regarding wholesale telecommunications services (CRTC 2013-551)[1] and a review of basic telecommunications services (CRTC 2015-134)[2]. Based on our experiences with interacting and accessing public submissions on the CRTC website during these consultation processes, we identified an opportunity to make the process and its results more transparent to everyday Canadians and policy makers. The goal of this project, funded by the Canadian Internet Registration Authority Community Investment Program, is to apply data science tools and techniques to public policy consultations submissions in order to create a general purpose framework that helps identify relevant information. This will help Canadians gain knowledge on the positions of intervenors, with a high-level goal of understanding how these positions inform policy.

As a proof of concept, we used the CRTC 2015-134 Review of Basic Telecommunications Services consultation as our test case to determine how to aggregate documents submitted to the CRTC, build a submissions browser, and analyze submissions in order to showcase what types of insights can be gained. The idea is that the open-source tools and scripts, along with the framework created by this project, can be applied in the future to other public CRTC consultation processes (past or present) or adapted to public consultation processes outside the CRTC's domain.

CRTC 2015-134 was chosen due to its subject matter. The CRTC's objective was to determine whether certain services, including broadband internet access, should be classified as basic telecommunications services by the regulator. In addition, CRTC 2015-134 received a particularly large number of public submissions. Our analysis focused on two main areas of the consultation, 1) Whether or not broadband internet should be classified as a basic service, and 2) What positions are being taken on the affordability of broadband internet services.

Our data science approach involved applying and assessing how modern data engineering tools and natural language processing techniques can be used to automate extraction of relevant information, context and sentiment from public submissions.

In the following sections, we will highlight the policy file browser we built and the analyses conducted on the submissions made to CRTC 2015-134.

[1] CRTC 2013-551 submission, accessed Feb. 23, 2018.
[2] CRTC 2015-134 submission, accessed Feb. 23, 2018.

Policy Browser

The Policy Browser is hosted at https://policy-browser.data.cybera.ca

The Policy Browser serves different purposes, depending on whether it is being used to help manage data associated with a public process, or whether it is being used by a researcher or member of the public looking to review data already collected. Much of the functionality of the browser itself is focused on the latter.

Users can browse documents submitted to the public process based on date or organization. They can also browse queries that have been saved to the local database, along with segments of text that match those queries. While displaying these matching results, the Policy Browser tries to give as many hints as possible on its context, without showing the same complexity of results that can be found by doing direct queries on the underlying Neo4J graph database - the graph style database used in this work. Consumers of the data can also do their own free text searches, using the underlying Solr query engine.

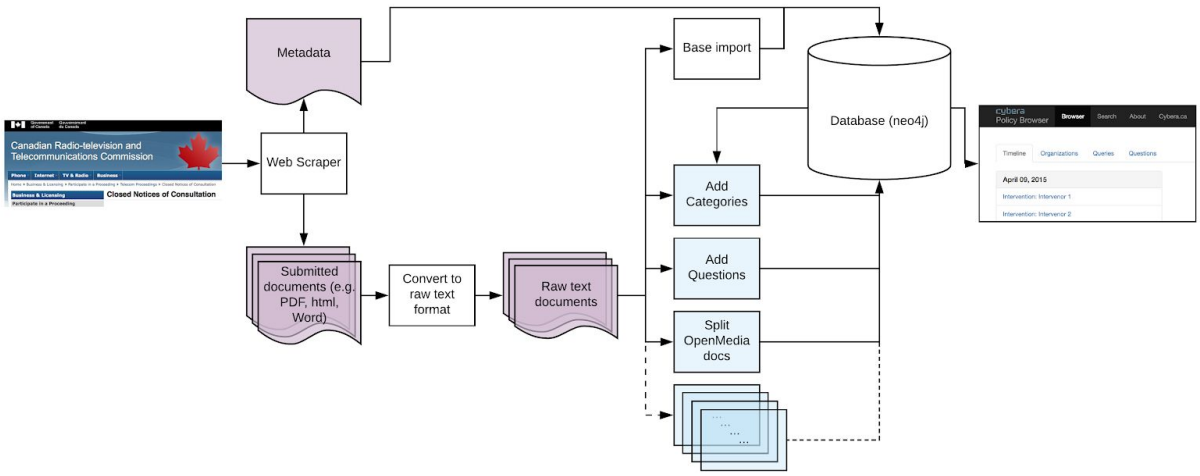
Finally, for questions that we (or future administrators) have already explored, there are summary pages that categorize matching segments of queries specific to a particular question, and show how many of these text segments exist for various organization categories. The browser makes CSV files available for download by anyone interested in doing further analysis in tools like R, Tableau, or even Excel.

Data administration functionality built into the Policy Browser allows users to save query results back to the database as new segments related to their original Documents. It also allows users to associate those queries with particular questions, which will continuously improve the quality of the CSV files that consumers can download.

Administrators with access to the hosting server are able to use a transformations framework, allowing the data to be gradually enriched as opportunities are identified. For example, a document that turns out to really be a collection of several individual responses can be split into multiple documents. The Policy Browser application itself follows a lightweight model-view framework, which makes it easy to add in new ways to view and navigate information.

More complete documentation for users of the Policy Browser can be found on the About page[3] . Information on how to add to, and work with, the underlying scripts and frameworks can be found in the code repository's README file[4].

[3] https://databrowser.data.cybera.ca/about
[4] https://www.github.com/cybera/policy-browser/blob/master/README.md



The Policy Browser was developed with a focus on one specific consultation: the CRTC public process 2015-134. It therefore defaults to that specific public process. Additional public processes can be added and accessed by changing the **ppn** parameter in the URL. It is currently not possible in the browser itself to navigate to a different **ppn**, but this would be an easy addition for a developer to make.

Many of the underlying data structures are quite specific to CRTC processes in general, so it would likely take more work to adapt the policy browser to document sets originating from outside of the CRTC. However, the scraper and browser can easily be adapted to other CRTC consultations, and instructions are included in the code repository for this.

Converting various document formats to text is done without knowledge of where those documents were downloaded from, with a naming scheme based purely on the content of the document, which means this can be run on other consultation documents. The basic transformation scripts will also work on other consultation documents. However, scripts specifically made to address problems specific to consultation 2015-134 would not function on other consultations without proper modification.

Our hope is that the Policy Browser helps other researchers in two key ways:

1. By decoupling the organization of the documents from the retrieval of the documents (while retaining information derived from their initial scraped location), it becomes easier to navigate the documents in new and novel ways. We have presented a few ways of doing that navigation, including organizing: By the times that the documents were submitted, 2) By the organization that submitted them, 3) By queries that match sections of of the documents, and 4) By what questions they may answer. (Note that navigation is also aided by converting all documents to a common text format so that they can be

browsed on a single webpage, as opposed to separate PDFs, DOCs, or other artifacts that must be individually downloaded.)

2. The browser provides a means to do the initial breakdown of a large amount of unstructured data (we counted approximately 16 million words in the 2015-134 documents) into more manageable pieces. We also provide a way to easily download these curated text snippets via CSV. In this way, researchers trying to understand what is going on in a consultation can work on two separate but complementary levels at the same time:
    a. By finding and imposing structure to the information in the documents (by querying for individual text segments and relating those to specific questions)
    b. By using contemporary data science techniques to analyse the smaller, more focused sets of text produced

Analysis

Documents Overview

In total, there are 23,186 documents in the CRTC 2015-134 consultation analysis. This is equivalent to approximately 65,000 pages, or 216 novels,of published material. These documents range from full interventions consisting of dozens of pages written by legal teams at various organizations, to single paragraph submissions made by individual Canadians. Of note, a large number of these documents were aggregated by Open Media, which ran a campaign to solicit Canadians' input to the CRTC 2015-134 consultation. Open Media collected approximately 20,281 submissions, which were provided in PDF-form and split up as part of our document analysis.

In total, documents were submitted by 21,260 intervenors, the majority of which were submissions aggregated by Open Media. There were also many submissions made by individual Canadians (529 in total) and organizations (166[5] in total). In addition, 289 submissions were collected by ACORN, an independent national organization for low- and moderate-income families.

We categorized each organization based on their similar and/or distinct backgrounds. The categories used were:

Category	**Sample Organization**
Advocacy organizations/Consumer advocacy organizations	Deaf Wireless Canada Committee, i-CANADA
Chamber of commerce/economic dev agency	Chambre de commerce des Îles de la Madeleine, Qikiqtaaluk Corporation
Government	Municipal and provincial governments
Other	Canadian Federation of Agriculture, NWT Association of Communities
Network operator - Cable companies	Rogers Communications, Shaw Cablesystems G.P.
Network operator: other	CanWISP
Network operator: Telecom Incumbents	Bell, TELUS Communications Company

[5] Note that this does not include organizations as listed in Phase 2 of the intervention, which specifically sought feedback from individual Canadians.

Rural/remote community association	Thetis Island Residents Assoc
Small incumbents	tbaytel

It should be noted that some of the numbers in this section numbers differ from the raw numbers obtained on the CRTC website. This is because our analysis and import mechanism were focused on automating as much of the process as possible. This includes extracting metadata from file submissions to produce labels such as the name of the submitting organization, date of submission, phase of submission, etc. Because these numbers are based on automated extraction and categorization, the totals have some degree of uncertainty associated to them.

Affordability of Internet Services

Introduction

For this analysis, we investigated the arguments surrounding affordability of internet service as they apply to each organizational category of intervenors. Our goal was to determine whether the language used by each intervenor category would imply differing opinions surrounding the issue of affordability. The disparity in affordable internet service is an important subject as it relates to the digital divide between rural and urban regions across Canada. These issues are further amplified as Canada's economy and workforce become more digital, and those who cannot afford access to the internet at reasonable prices are left behind in terms of the skills required to meaningfully participate in the digital economy.

Methodology

Doc2Vec

To locate and isolate segments of text within the entire corpus of documents, we utilized a neural network paradigm known as doc2vec[6,7,8]. Doc2vec is a machine learning algorithm that can identify the semantic meaning or context of various lengths of text. This can be useful to find lengths of text that express a similar sentiment, or discuss similar topics. It works by transforming each unique length of text — either sentences, paragraphs, or entire documents — into a numerical vector representation of the text. Once each length of text has been vectorized, these vectors are used to train a shallow three-layer neural network (input, hidden and output), allowing the network to 'learn' the context expressed in each length of text. Once the neural

[6]Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. http://arxiv.org/pdf/1405.4053v2.pdf.
[7]Xin Rong. word2vec Parameter Learning Explained. https://arxiv.org/abs/1411.2738.
[8] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pages 1188–1196, Beijing, China.

network is trained, it is then possible to find the lengths of text that are similar based on their vector inner product from the output of the neural network. In other words, doc2vec provides an excellent tool for searching for concepts or ideas within a corpus of text, independent of the length of the search text and the length of text segments used to train the model.

We divided every document into individual sentences, rather than paragraphs or documents, as it provided an artificially richer training set that is more straightforward to query. The neural network used had a hidden layer of 300 neurons and was trained over 20 epochs (an epoch refers to a full training cycle on the training set). Before training, each sentence was cleaned by removing special characters and numbers. Sentences that were less than fifteen characters in length were also removed from the raw text documents. After this cleaning process, we applied a "porter stemmer" to reduce conjugated words to their basic form (to reduce ambiguity between words). We additionally removed all stop words (e.g. common words that do not provide much insight into the text, such as "the") from the sentences used to train the network. Pre-cleaning the documents in this way simplifies training of the neural network, as it reduces both ambiguity and the size of the corpus the network needs to learn.

Once the network was trained, we used it to find segments of text that were semantically similar to the statement:

> "Affordability of broadband internet access"

We then utilized a Monte Carlo simulation to ensure that our search term had adequate coverage of the entire document space before we began our analysis. With this method, we were able to identify approximately 2,400 segments of text that potentially contained our subject matter. After initial segment extraction, we used pre-built word embeddings from text2vec to filter our resulting sentences further by only keeping sentences with words that are similar in context to "cost, affordability, price and expensive"[9]. This allowed us to focus our analysis on highly relevant segments of text.

Analysis

Once relevant sections of text were isolated and filtered to contain only words relevant to the idea of affordability, it was possible to visualize the relationships between frequently used language using a bigram network. In this context, a bigram is simply a pair of words that appear concurrently within the text itself. The bigrams that appear most frequently can then be visualized as a connected network, showing the relationships between the most frequent bigrams. The advantage of a visualization such as this is that the most frequently used language is shown as connected clusters, making it possible to determine a general overview of the common themes within a large body of text. We should cautiously note, however, that by

[9] Jeffrey Pennington, Richard Socher, Christopher D. Manning. GloVe: Global Vectors for Word Representation. 2014. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://www.aclweb.org/anthology/D14-1162.

using this kind of reductionist methodology, the context surrounding the words is easily lost. It is therefore important to avoid interpreting these results positionally without manually inspecting the relevant sentence(s). These results provide an easily digestible content summary.

In this discussion we showcase two of the most notably contrasting bigram pairs between different organization categories: those of advocacy organizations, and those from telecom incumbent groups. Interested parties can see a more complete discussion of the differences between intervenor category groups as part of our open source GitHub repository[10].

Figure 1 depicts the relationships between the most common bigrams by advocacy groups, taken from approximately 1,300 segments of text. The different clusters potentially represent developing narratives or themes within the corpus of machine identified segments. Interestingly, the second most frequent bigram by advocacy organizations is “afford(ability)” and “internet”, which appears inside a cluster of connected bigrams alluding to a conversation relating to internet affordability (which also includes a discussion of consumer oriented concerns). For example, the bigram pairs including the word “price” would imply a discussion of how advocacy groups either define the current state of consumer internet prices, or are discussing what they feel would be ideal. Essentially, this network shows that advocacy groups’ discussion of affordability is consumer-centric. A notable secondary cluster exists around the word “fund(ing),” wherein there are frequent bigrams surrounding words that would imply a discussion of either funding mechanisms for advocacy groups to operate, or funding mechanisms related to infrastructure. The secondary cluster may indicate that the discussion of affordability by advocacy organizations is not only consumer centric, but may also contain a discussion either of funding to provide infrastructure and services, or providing funding assistance to consumers. Further manual inspection of relevant text segments would be needed to more accurately identify the context alluded to by the bigram analysis.

Figure 1 appears to demonstrate that advocacy organizations are having a conversation about affordability as it relates to themselves and the stakeholders they represent. Certainly, this result is not surprising given the general motivations of consumer focused advocacy groups. Doc2vec appears to be effective at locating even slightly relevant elements of text surrounding the idea of affordability, and as such, conceptually relevant segments of text that discuss ideas such as funding mechanisms and costs are also captured. As the subset of text came from advocacy organizations, we are finding all conceptually similar statements on affordability as it relates to advocacy groups. Thanks to this general approach, we see that advocacy organizations are discussing the facets of affordability in a way we would expect, i.e. what affordability means to them, both from a consumer and a funding point of view.

[10] GitHub Notebooks

Figure 1. *This figure displays a network of word pairs (bigrams) that appeared more than 15 times in segments of text submitted by advocacy organizations around the topic of "affordability". In this figure, the opacity of the lines connecting word nodes are proportional to the frequency of the pair -- the darker the line, the more frequent the bigram. We note that, due to the word truncating (stemming) process, some words may be improperly conjugated from how they appeared within the text. For example, the word "fund" primarily appeared as "funding" within the text itself.*

In contrast to the bigram network created from the language used by advocacy organizations is the network created using the language of telecom incumbents surrounding "affordability". The bigrams in Figure 2 were extracted from 254 segments of text using an identical process to what was used to create Figure 1. When displayed as bigram networks, Figures 1 and 2 show that the language used by telecom incumbents around "affordability" is readily distinguishable from that used by advocacy groups. For example, Figure 2 shows clusters around words like "cost", discussing topics such as "cost-studies", "telecom-costs", and "final-costs". The most frequent word pair in these segments of text is "target" and "speed", which would imply that there is significant discussion around the costs associated with the internet services they

provide from their business perspective, either operationally or from a revenue point of view. The bigram network shown in Figure 2 seems to contrast that of Figure 1, in that the language around affordability is no longer consumer centric, but instead uses language that is more business oriented.

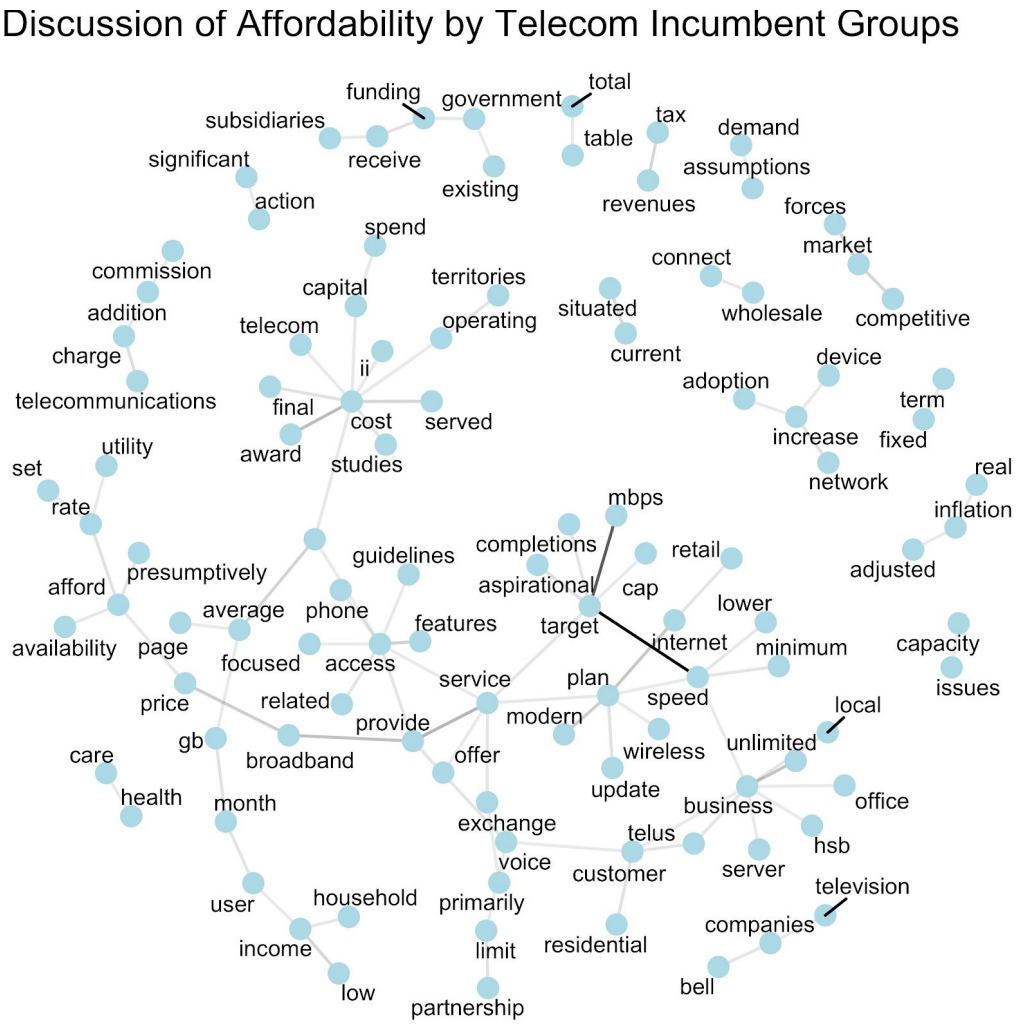


Figure 2. *This figure depicts the bigram pairs that appeared more than three times in segments of text taken from documents submitted by telecom incumbent groups around the topic of "affordability". The more frequent bigrams that appear in this figure are noticeably distinguishable from those that appear in Figure 1.*

While the language used by telecoms is distinguishable from that used by advocacy groups, the results in Figure 2 are not considerably different from those of Figure 1. Both figures provide a top-down summary of relevant topics and frequently used language used by telecoms and advocacy organizations. The language of telecom incumbents shown in Figure 2, is a representation of a discussion surrounding what affordability of internet services means from the perspective of a service provider. The most frequently used terms by telecoms seem to

focus on the cost of providing a service, rather than the cost of receiving a service. Using the expanded view points provided by Figures 1 and 2, it appears that telecom incumbents submitted a different discussion surrounding affordability than advocacy groups.

Conclusion

By using doc2vec as a search tool to locate sentences that contain content semantically similar to the "affordability of broadband internet access" topic, it was possible to identify other segments of text from the CRTC documents with some mention of affordability. These results were refined using text2vec word embeddings in order to further isolate the language used by differing invested organizations. We found that the language used around "affordability" is considerably different depending on the group having the discussion. Each group focused on what affordability means to their specific interests and objectives. Advocacy organizations discussed affordability from the perspective of Canadian consumers, and telecommunications companies discussed affordability from the perspective of running a business.

In essence, we have discovered that "affordability" takes on different meanings to different organizational categories. Doc2vec, in combination with a secondary word embedding filter, performed admirably as a text mining methodology for identifying a general overview of the discussion within important sections of text. Unfortunately, this methodology does not automatically identify opinions from text. For that task, human intervention is still required, either by manually going through each document, or by creating a training set for more sophisticated machine learning techniques.

While our analysis does not answer whether or not internet access in Canada is affordable, it does reveal potentially important discussion points from every group surrounding the idea of affordability from a larger scope. Essentially, this analysis has provided us with a "birds-eye-view" of the different issues surrounding affordability from the perspective of each individual group and provides an automated approach for analyzing large amounts of unstructured data in order to gain an understanding of which segments or submissions to investigate further. What we can state with certainty from this analysis is that the answer to the question of affordability of internet access in Canada really depends on whom you ask. Interestingly, the CRTC Commission noted similar findings in its Telecom Regulatory Policy CRTC 2016-496 decision, wherein on the question of affordability of broadband internet services, it found that "Parties were divided on the need for regulatory intervention regarding prices for broadband internet access services[11]."

[11] https://crtc.gc.ca/eng/archive/2016/2016-496.htm Paragraph 190, accessed Feb. 23, 2018.

Broadband Internet as a Basic Service

Introduction

At the heart of the CRTC's 2015-134 Notice of Consultation was the goal of determining the future of Canada's telecommunications services, and how essential broadband internet services are to that future. Ultimately, the CRTC ruled in Regulatory Policy 2016-496 that broadband internet is, indeed, a basic service. In the interest of understanding which intervenor groups were in favour of defining broadband internet services as a basic telecommunications service, we used the same text mining process applied to the affordability question described above. In this case, we were hoping to identify and extract the position of each organization on an updated broadband basic internet service objective, and analyze the language they used to discuss this subject.

Methodology

For the basic service question, we followed a similar analysis pipeline as the affordability question, but with the following search term:

> "Broadband internet services should be considered a basic telecommunications service"

This search found approximately 2,700 unique and potentially relevant segments. We again further refined these results using text2vec, filtering down to words that were similar in context to "defined", "essential", "basic", "universal", and "mandated", with a relatively strict similarity criterion requiring words to have a cosine similarity score of at least 0.8. However, in this case, the frequently used terms related to the basic service question were fairly uniform between groups (partially as a consequence of the question being repeated in each organization's answer). An unfortunate consequence of our organizational bigram similarity is that it does not reveal many distinguishable trends (the bigram network visualizations are available in the GitHub repository[12]). In this case, we used a metric known as term frequency–inverse document frequency (TF-IDF) to measure the "term importance" for each bigram. The TF-IDF process ranks the most frequently occurring words, while also offsetting unimportant high frequency terms such as "the". The higher the TF-IDF score, the more potentially relevant a word/group of words is to the text.

Analysis

In Figure 3 we see that the important terms vary between each intervenor group, with respect to segments of text that have been machine identified as relevant to the basic service question. For example, important bigram terms used by advocacy organizations appear more directly similar to bigrams of the basic service question itself. The other organizational groups have decisively different important word bigrams, indicating that they have other important points to

[12] GitHub Repository Notebooks Link

consider, including the potential impact of a basic service ruling on their core business services. The combined doc2vec, text2vec, and TF-IDF approach provided a means of identifying subtle but important differences in opinion in intervenor groups that were not immediately obvious after only applying text2vec filtering. This approach is, however, not without issue, as suggested by the high TF-IDF score assigned to the bigram “laffont universal” in the Network Operator: Telecom Incumbents group. (This is a result of doc2vec including a citation of a paper by French economist Jean-Jacques Laffont that was referenced by a number of Telecom Incumbents. While there is an argument to be made about the relevance of his work, a citation should not be regarded as “important language” used by this group.)

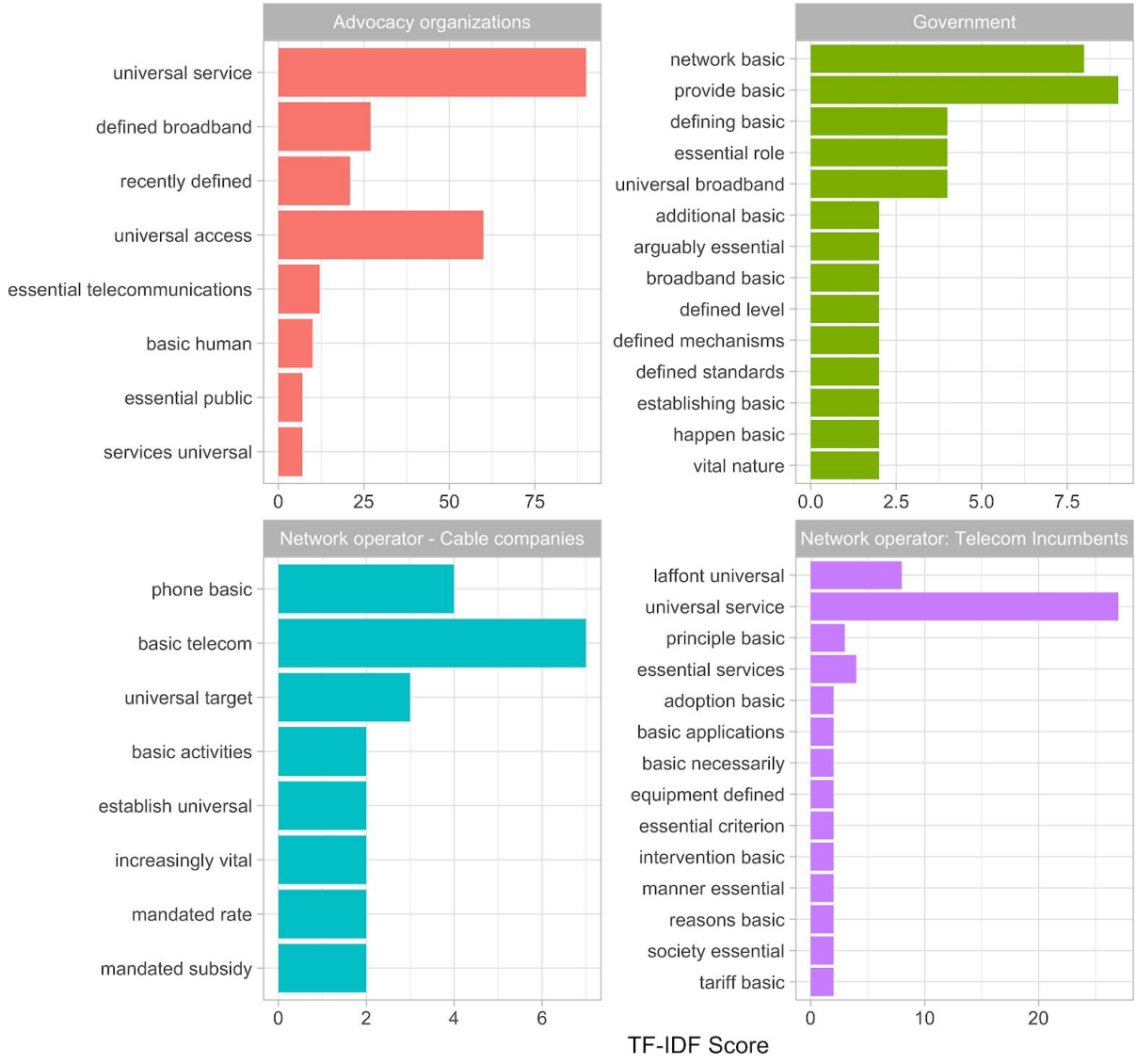


Figure 3. *These figures depict the TF-IDF score of bigrams from machine selected segments relevant to the basic service objective question from four invested groups. In practice, the most important terms to the selected segments are outlined here. In this context, a higher TF-IDF score is more likely to be indicative of a term of greater importance.*

Conclusion

In contrast to the question of affordability, the organizational discussion of the basic service objective seems to be, at the surface level, the same conversation. This is not necessarily surprising, as the basic service objective question leaves less room for interpretation or personalization than the question of affordability. Each group discusses the basic service objective in terms of the consequences such a definition will have for them. As a result, traditional bigrams are a less informative tool. However, by applying the TF-IDF score to the word segments, it is possible to find the most important bigrams in the group of text. Using this metric, it was possible to understand important topics surrounding the basic service objective "at a glance". This has revealed important word combinations which require further manual analysis to interpret. Due to the nature of the responses to the "broadband internet as a basic service" question, the text mining techniques we applied appear to have only provided a very limited general overview of the discussions taking place, and is not a substitute for manually inspecting the documents.

Overall, our results appear to corroborate those reported by the CRTC in Telecom Regulatory Policy 2016-496, in which they note that "Almost all parties in this proceeding, whether individuals, TSPs, governments, or non-governmental organizations (e.g. accessibility groups and consumer associations), submitted that Canadians need broadband internet access services to participate in Canada's digital economy[13]." The CRTC also noted that only "A small number of parties, such as Saskatchewan Telecommunications (SaskTel) and TBayTel, submitted that while fixed broadband internet access service is important, the Commission should not establish this service as a basic telecommunications service[14]." Although our text mining approach did not appear to identify a small number of dissenting opinions among the large amount of data, additional refinements or natural language approaches could be applied to assist in refining our approach.

Analyzing submissions from 3,000 Canadians
(OpenMedia, ACORN, and phase 2 interventions)

Introduction

While a large proportion of documents submitted to the CRTC were submitted on behalf of organizations, the consultation also presented an opportunity for individual Canadians to contribute during the second phase of the consultation. The CRTC presented the results from

[13] https://crtc.gc.ca/eng/archive/2016/2016-496.htm Paragraph 25.
[14] https://crtc.gc.ca/eng/archive/2016/2016-496.htm Paragraph 28.

this phase 2 questionnaire in its *Let's Talk Broadband Findings* report[15]. This part of Cybera's analysis focused on analyzing the free-form submissions presented to the CRTC in phase 2 of the consultation, in addition to free-form submissions made to OpenMedia and ACORN. Submissions from these organizations are more technically challenging to analyze due to their volume and unorganized nature, and were therefore excluded from the Let's Talk Broadband report and a review article published on the Basic Service Objective proceedings[16].

Origin of Submissions

OpenMedia sought support for its open letter to the CRTC on its website: unblockcanada.ca. The site provided participants with the option to either submit a standard letter to the CRTC, or create their own submission. In total, OpenMedia aggregated 20,281 individual submissions, of which 17,854 consisted of the standard letter, and 2,427 were unique submissions. Submissions made to OpenMedia covered all geographic areas in Canada (see Figure 4).

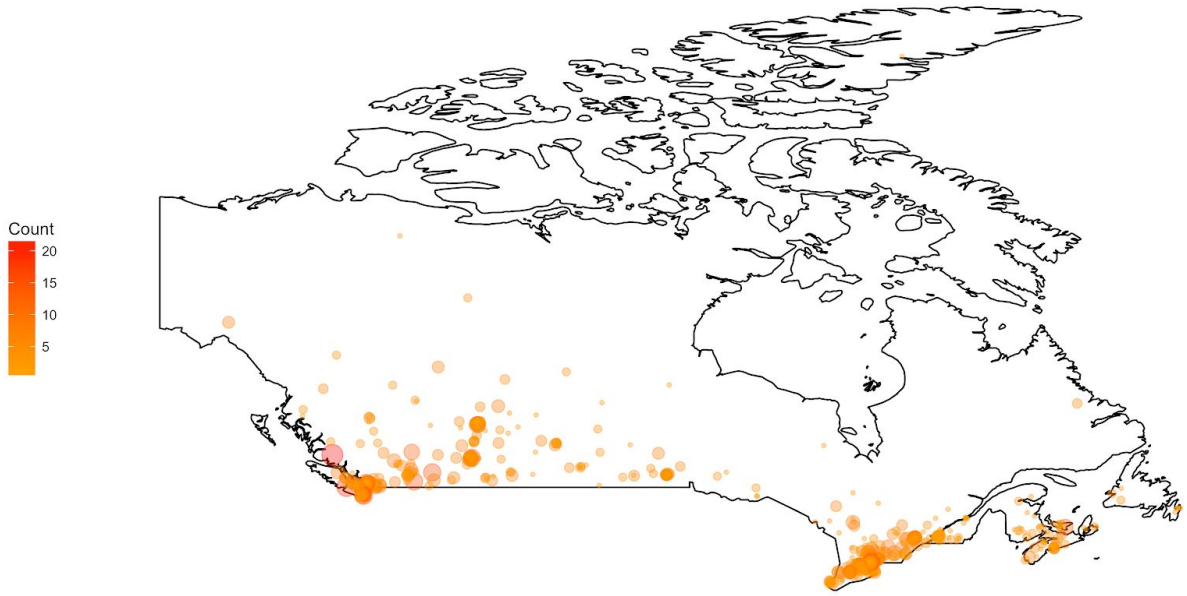


Figure 4. *Map representing where the 20,281 submissions by individual Canadians came from as submitted by OpenMedia.*

[15] Let's Talk Broadband
Findings Report. 2016. EKOS Research Associates Inc.
http://epe.lac-bac.gc.ca/100/200/301/pwgsc-tpsgc/por-ef/crtc/2016/030-15-e/report.html.
[16] Rajabiun, Reza. 2017. The Rise of Broadband as an Essential Utility and Emergent Concepts in Universal Access in Advanced Economies: Perspectives from Canada, 28th European Regional Conference of the International Telecommunications Society (ITS): "Competition and Regulation in the Information Age", Passau, Germany, July 30 - August 2, 2017.
https://www.econstor.eu/bitstream/10419/169494/1/Rajabiun.pdf.

ACORN Canada (Association of Community Organizations for Reform Now) is an independent national organization for low- and moderate-income families. In preparation for its submission to the CRTC, it asked members to fill out questionnaires consisting of multiple choice and free-form questions. ACORN aggregated 289 submissions. The questions from this form that were of particular interest to Cybera were:

- "How do you feel about the current pricing of high speed internet?"
- "Why is online access important to you?"
- "Please share how your life would change if you could easily afford home high-speed Internet."
- "Please share anything else relevant."

As described above, our analysis of submissions by individual Canadians also included the free-form submissions made to the CRTC as part of the second phase of its consultation process. Our analyses extracted text from 466 submissions out of a total of 529 reported by the CRTC[17].

Methodology

For the analysis of the free form text responses provided by individual Canadians, we used similar term-frequency (TF), bigram, and trigram analyses as described in the previous methods sections. Further, topic modelling techniques were applied as well, such as latent dirichlet allocation (LDA). For the topic analysis, five topics were identified for each of the submission categories. (For further details on how the analysis was conducted, see Text Mining in R[18].)

Analysis

Analyzing the text from the submissions made by individual Canadians, there are several distinguishing features depending on the source of the content. However, there are also consistent patterns seen across all three groups: discussions on the importance of affordable internet access and internet speeds. Looking more closely, analysis of word and n-gram (co-occurrence of words) frequencies reveals more subtle differences. OpenMedia submissions include many discussions about large telecom companies, competition, and community access; whereas ACORN submissions discuss low-income family issues and internet as a necessity for work and everyday life. Finally, submissions made directly to the CRTC during the second phase of the consultation often directly refer to large telecom providers like Shaw, Bell, and Telus, but also diverge more into discussions of other telecommunications services.

Differences are also seen as a result of the topic modelling analysis (Table 1). OpenMedia submissions focused on the quality of internet connections, speed, and affordability. By contrast, ACORN's submissions have a distinct narrative around affordability and the impact of

[17] Interventions Phase 2. CRTC. https://services.crtc.gc.ca/pub/ListeInterventionList/Default-Defaut.aspx?en=2015-134&dt=i2&lang=e&S=C&PA=t&PT=nc&PST=a.
[18] Text Mining in R. Chapter 6 Topic Modelling https://www.tidytextmining.com/topicmodeling.html.

internet service on day to day life. Direct submissions to the CRTC discuss internet speeds and telecom providers, as well as topics that would affect certain subsets of the Canadian population, such as rural internet.

Creating bigram-webs, as described in the above sections, can help provide a better understanding of how each of the word pairs are related to each other. Figure 5 (below) focuses on text submissions to the CRTC's phase 2 interventions.Clusters can be seen around words like "internet", "service", "data caps", and "upload/download speeds", highlighting some of the narratives emerging in these submissions.

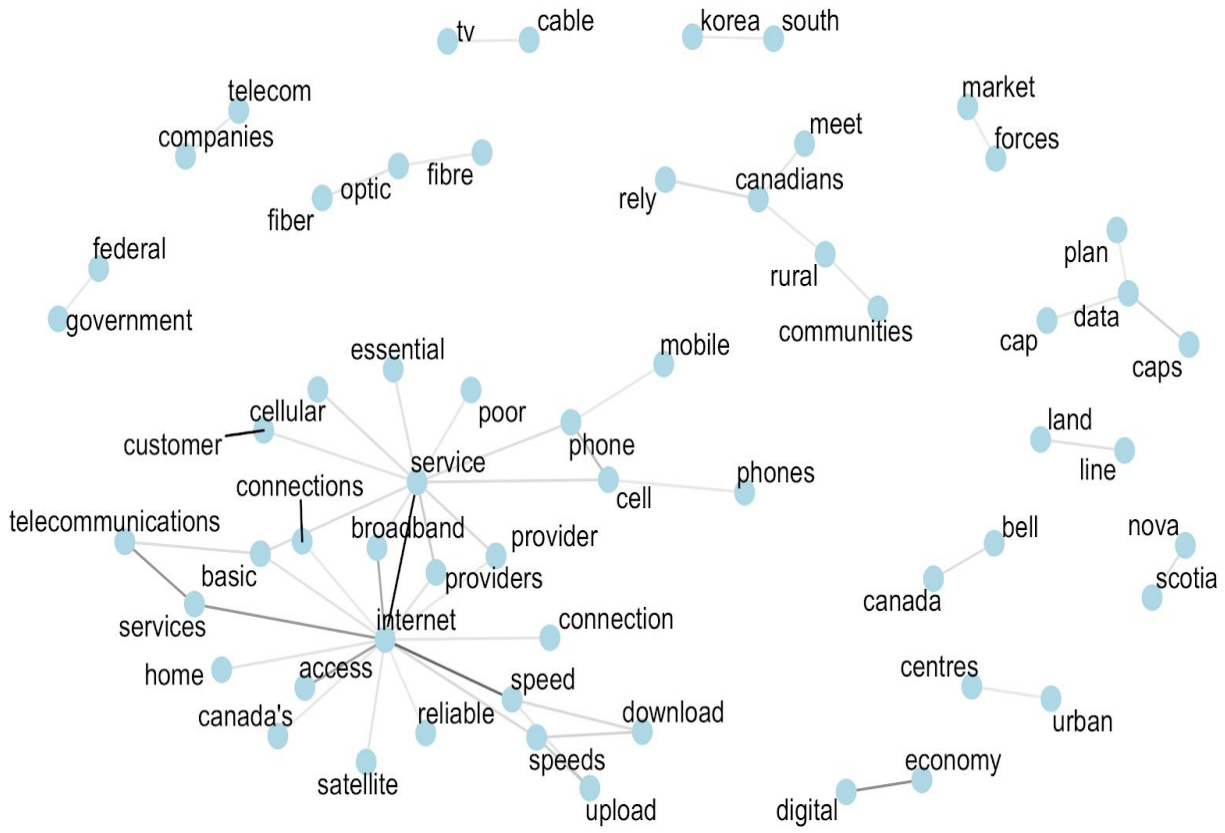


Figure 5. *Bigram network graph based on submissions made by individual Canadians during phase 2 of CRTC consultation 2015-134.*

Table 1. Topics identified from individual submissions made to the CRTC

OpenMedia	ACORN	Phase 2 submissions
Telecom companies	Internet access world	Internet data speed
Internet access affordable	Internet access information	Internet access speed
People pay country	Internet access afford	Internet access crtc
Speed price competition	Internet food information	Internet broadband digital
Internet access world	Internet food afford	Internet access data

Conclusions

A wealth of information was captured from individual Canadians during the 2015-134 CRTC Consultation. Analyzing the more than 3,000 submissions in detail provides a unique opportunity to gain a better understanding of what Canadians think and want out of their internet service. Our text mining analysis, using topic modelling and bigram network graphs, identified similar topics to what the CRTC reported in its policy decision 2016-496. For example, it noted that ACORN's members raised concerns about the affordability of internet services, along with some of the tradeoffs that are needed at times to have reliable internet access. Similarly, the CRTC commented on the frustration observed by individual responders during phase 2 of the consultation over their concerns with data limits, as reflected in the cluster seen in our bigram graph (Figure 5). While we primarily presented qualitative results here, further analysis and academic research on specific topics (e.g. "What should be considered basic internet service speeds?") would help produce quantitative outcomes as well.

Question and Answers During the Consultation

Introduction and Methodology

During CRTC consultation 2015-134, there were four rounds of questioning that allowed intervenors to ask for additional information from others. The questioning took place over approximately six months between August 2015 to February 2016.

The goal of this analysis is to determine how intervenors interacted with each other, how often, and which groups interacted with each other most frequently.

This analysis was performed on data that had been manually aggregated from the original CRTC consultation data and provided by Dr. Catherine Middleton (Professor, Ted Rogers School of Management, Ryerson University). Dr. Middleton's research group (in particular Hamzah

Kobari) manually inspected the data and extracted the relevant metadata showing which intervenors asked questions, who the questions were addressed to, and the date of the questions and responses.

Analysis

During the course of the consultation, a total of 338 questions were asked over four rounds, with round one being most active (containing 205 or more than 60% of the questions). Rounds two, three, and four were markedly less active, with 47, 76, and 10 questions, respectively.

The most active intervenor group was Network operator: other, followed by Advocacy organizations and government as shown in Table 2. Notably, there were only four questions submitted by individuals, meaning the questioning process of the consultation was heavily dominated by organizations.

Table 2. Questions asked by intervenor category during CRTC 2015-134

Intervenor Category	Number of Questions asked
Network Operator: other	90
Advocacy organizations	72
Government	70
Network Operator: Telecom Incumbents	47
Network Operator: Cable Companies	45
Small Incumbents	6
Individual	4
N/A	2
Other	2

Patterns and lines of questioning can be displayed using chord diagrams. Chord diagrams show where the questions originated from, and which party they were directed to. The starting width of the chord represents how many questions were asked by that party to the recipient group. The ending point shows how many questions were asked in the other direction.

For example, Figure 6 (below) shows how many questions were asked by groups from the Government category to the other categories. Following the purple band from *Government* to *Telecom Incumbents* reveals an imbalance in questions originating from *Government* to *Telecom Incumbents,* compared to the opposite direction. An interactive version of this chord diagram can be accessed here (please right-click the Download button to 'save link as').

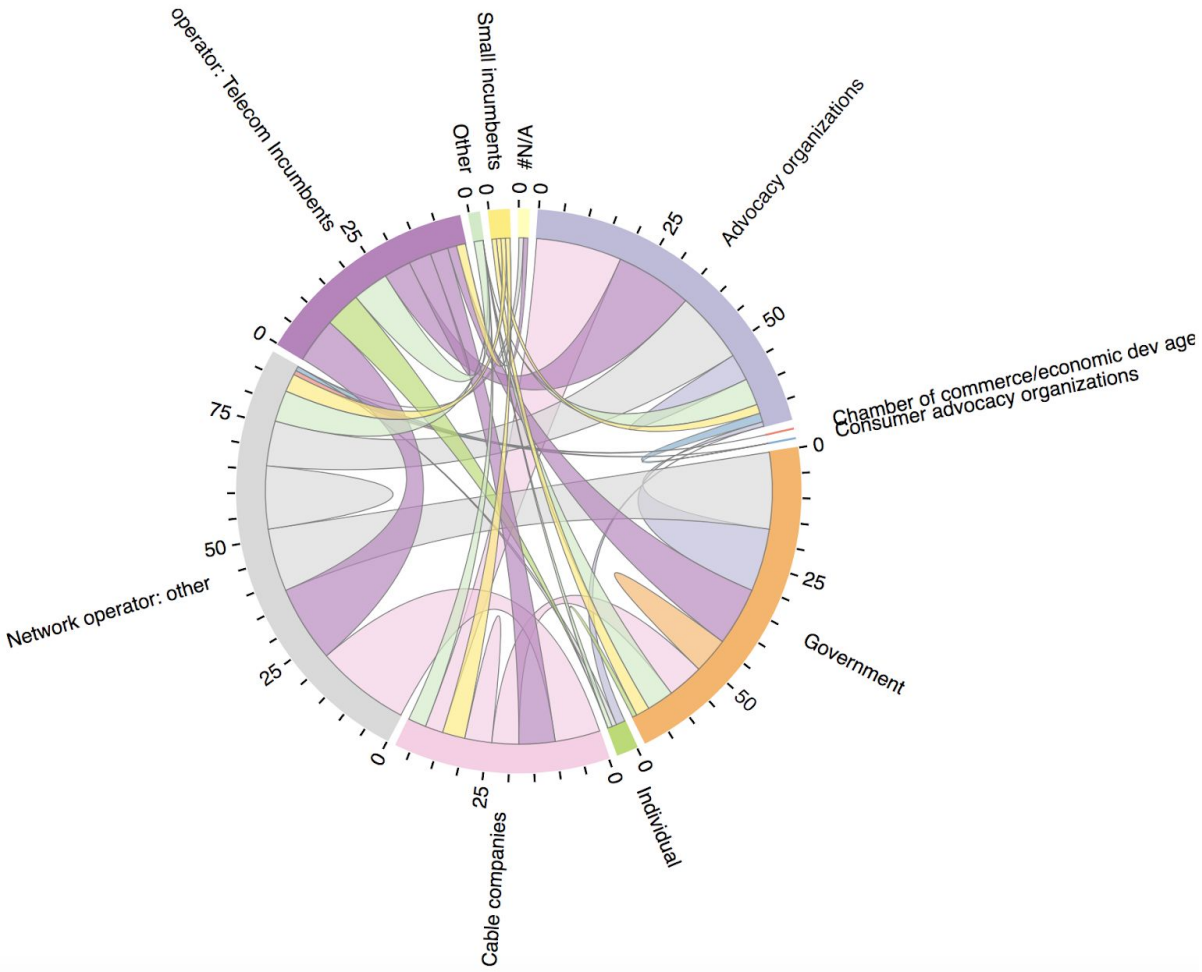


Figure 6. *Chord diagram showing the flow of questions between intervenor categories (e.g. Government, Cable companies, etc) during CRTC 2015-134. Each section represents a category and how many questions were asked by that category. Chords connecting categories represent how many questions were asked between categories.*

Analyzing the data on a per organization basis shows that the most active participants were the CRTC, which asked the most questions at 53, followed by the Canadian Network Operators Consortium, Rogers Communications, the Affordable Access Coalition, and Telus Communications. In total, only 20 intervenors participated in the questioning round, all of which our outlined in Figure 7 below.

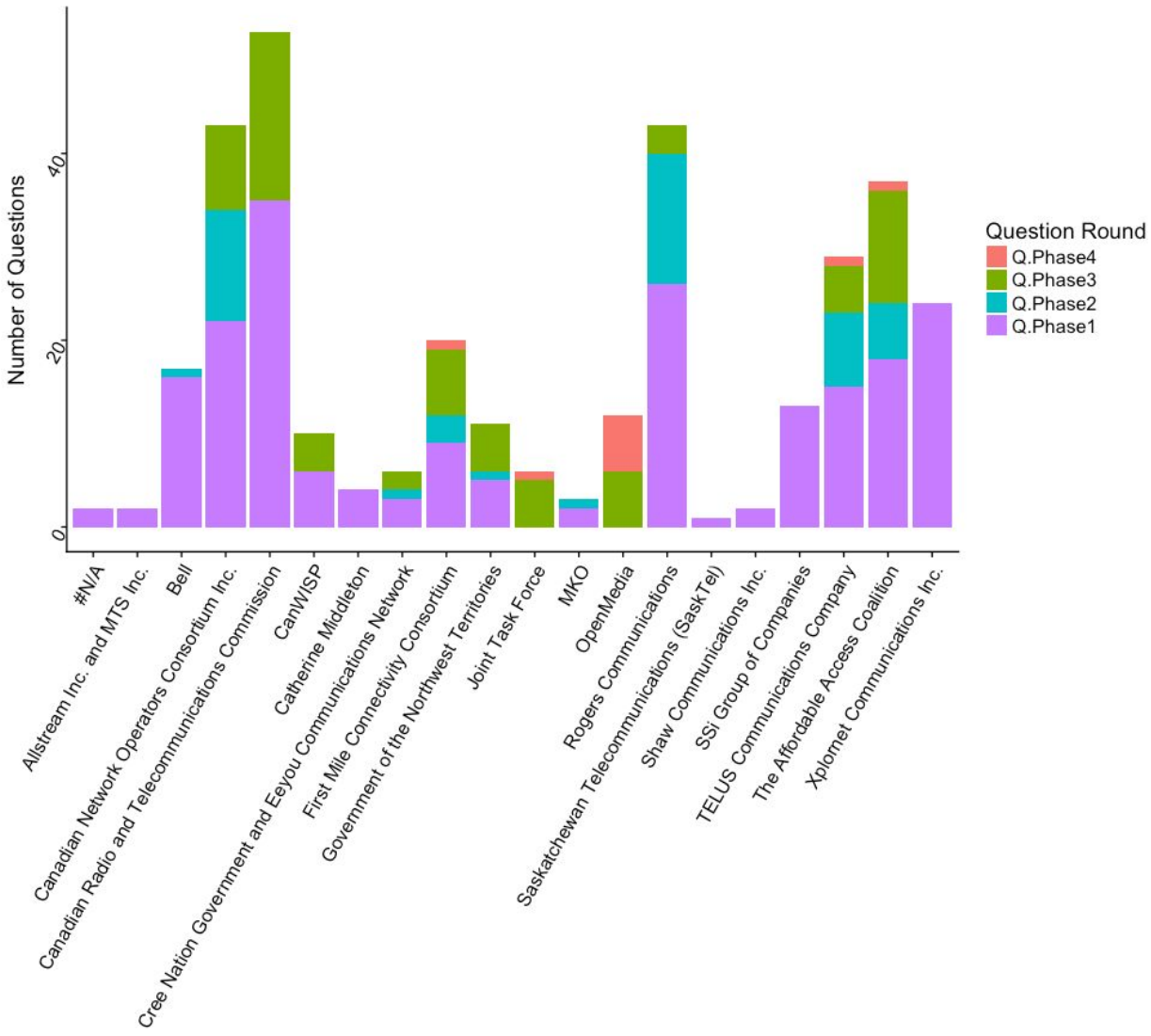


Figure 7. *Intervenors that participated in the questioning rounds of CRTC 2015-134. In total, 20 organizations took part.*

Analyzing how many questions each individual organization was asked reveals a larger group of organizations. In total, 61 intervenors were asked at least one question. Bell was asked 32 questions, which was the highest number of questions received and exceeded Telus (which received 19 questions) by almost 70%.

Combining the number of questions asked and answered reveals that there are certain organizations who asked or answered far more questions than vice versa. This was assessed by subtracting the number of questions asked from the number of questions answered. The CRTC displayed the largest imbalance of questions asked against received, with 0 questions received and 53 asked. (This is because the process does not allow for questions to be asked of the CRTC.) Notably, three of the top five organizations with a questions-asked surplus are

Government or Advocacy organizations. On the other hand, the five organizations with the biggest deficit in terms of questions received versus asked are all network operators.

Conclusions

The analysis of the four rounds of questioning by intervenors revealed interesting dynamics. It is clear that the questioning rounds were largely used by organizations as opposed to individuals, as only about 1% of all questions were asked by individual intervenors. With regards to questioning, the CRTC clearly leveraged the questioning rounds to extract more information from intervenors. At the same time, network operators received the highest number of questions, receiving more than 60% of all questions.

Recommendations

Cybera developed methodology for automatically downloading data related to one or more consultations conducted by the CRTC for further analysis. We believe there is tremendous value in analyzing the documents submitted during CRTC consultations to both individual Canadians as well as academics. During our work, we have been in touch with three separate research groups that have expressed interest in using our tools for further analysis.

While this approach represents one method for aggregating and processing the data that was submitted to the CRTC on public record, we also have several observations on the intervention process that could help simplify the reviewing of documents and improve the accuracy of the analysis.

Document Formats

- PDF is not an optimal format for text extraction. For example, files with footers and paragraphs overlapping multiple pages make it virtually impossible to separate the footer from the paragraph text.
- Zip archives require additional logic when downloading the documents, and also carry files inside the archive that follow different naming schemes and conventions than the rest of the submissions.
- Documents bundling statements from several individuals:
    - Several PDF documents contain 3,000+ pages, which look to be rows of a spreadsheet printed to a PDF file. Even a simple change, such as making the original spreadsheet available, would be much more amenable to analysis.

File Browsing

- The current CRTC site requires that every file be downloaded in order to be viewed, which makes the file browsing and scanning process far more difficult.

File Downloading

- Not all links lead to documents, and certain pages containing more than 50 documents do not have a navigation scheme to make files available via a URL. This requires more complicated Javascript navigation in order to access and download all of the documents.

Suggestions

- Standardization of input data:
    - Providing a form for intervenors for each stage of the submission process would help standardize the format of the data collected. It would also provide explicit metadata items that would not only make document analysis and interpretation

easier for the Canadian public and researchers, but also for the CRTC. Yes/no answers or numerical values for speed definitions would have made quantifying and aggregating questions, such as whether or not the internet should be a basic service, much easier and more accurate.

- File downloads
    - Access to the data could be improved by providing links to each of the submitted documents, and links for bulk downloading of all or multiple documents from a single public consultation process.
- Data de-identification
    - While documents and information submitted to the CRTC are on public record, it would be worth considering whether all personally identifiable information submitted, in particular by individual Canadians, should be made available or not. For example, a significant proportion of individual Canadians submitted their full names, along with phone numbers, mailing addresses, and six-character postal codes.
    - Note that explicit collection of metadata, as mentioned above, would help simplify the identification of fields potentially containing personally identifiable information.

Conclusions

Cybera's goal in this project was to develop methodology for gaining insights into materials submitted to the CRTC's public consultation processes. To do so, Cybera used CRTC consultation 2015-134, which conducted a review of basic telecommunications services in Canada, in order to develop a framework and case-study on how documents submitted to the CRTC can be aggregated, presented, and analyzed in a streamlined fashion. We demonstrated how to download, reorganize, and analyze a vast amount of complex, unstructured data. This provides a way to browse and interrogate the materials in a more intuitive and straightforward fashion.

Our subsequent analyses highlight differences in how groups of organizations, as well as individual Canadians, discuss the concepts and importance of affordable internet, and whether internet should be considered a basic service. While these analyses do not present a complete, self-contained picture, they often present the foundation for additional investigations. That is, while we were not able to fully automate the analysis, we hope to present a method for facilitating human-in-the-loop analyses, in which the automated component helps guide and target additional manual analysis of the data, and thereby speeds up the discovery of additional insights.

A key piece of our work is the Policy Browser platform, which allows for simple browsing of documents submitted to the CRTC, and organizes all documents, independent of submission format, according to the organization or time of document submission - something that is currently not possible on the CRTC's website. The Policy Browser also enables further analysis of the data, as it is tied in with Solr, a full-text search engine that allows for "fuzzy searching", and has the ability to tie data together that is perceived to be relevant to specific questions.

The browser also provides administrator functionality, allowing authorized users to log in to query and modify the underlying database, as well as customize additional aspects of the browser's analysis functionality. The Policy Browser is openly available for anyone interested in exploring documents submitted to the CRTC as part of consultation 2015-134 at policy-browser.data.cybera.ca. We invite users interested in using the admin functionality to contact us at datascience@cybera.ca.

In addition to making the browser openly accessible, Cybera has open sourced all of its analysis scripts, as well as code for running the Policy Browser (see github.com/cybera/policy-browser). This will allow anyone to easily build or iterate on our analysis. It also allows anyone to deploy their own policy browser, not only to explore documents related to CRTC 2015-134, but to download, process, and aggregate data related to any other consultation conducted by the CRTC. We hope this will be of use to Canadians and academic researchers interested in gaining a better understanding of the public consultation processes conducted by the CRTC.

Acknowledgements

Cybera would like to thank the Canadian Internet Registration Authority for funding received through its Community Investment Program to facilitate the work presented herein. Cybera would also like to thank Dr. Catherine Middleton and Hamzah Kobari for providing the dataset on intervenor lines of questioning. We also thank Dr. Middleton and Dr. Reza Rajabiun for discussing their research and advising us throughout our project.