Topic Modeling and Visualization of SFU Linguistic Department Articles and Comments

Maria Babaeva, Andy Chen, Pushkar Sinha | Jiannan Wang, Steven Bergner | SFU CMPT 733, BigData, Spring 2018

BACKGROUND

In the age of Big Data, large amounts of data are generated each day. It is increasingly difficult to process, analyze, and extract useful information from the data. Topic modeling provides a set of methods to analyze textual data.

Some common uses of topic modeling include:

1. Finding latent topics
2. Extracting entities such as people and organizations
3. Measuring sentiments in the documents.

PROJECT GOALS

The project goal is to create a framework for topic modeling on unstructured text. In particular, the project aims to answer the following:

1. What are the top topics in both articles and comments corpuses.
2. What is the overlap of topics in those corpuses?
3. What are the entities in both articles and comments and the entities from articles mentioned most in comments?
4. Who are the similar types of authors based on the similarity of topics/entities in their text?
5. What is the trend of sentiments of comments over time? What are the distributions of sentiments for articles and comments?
6. What are the important authors and topics from articles in comments?

DATA SOURCE/TOOLS

The SFU Opinion and Comments Corpus (SOCC) is part of a project a that investigates the linguistic characteristics of online comments.

The data was obtained from SOCC and contains:

- 10,339 opinion Globe and Mail articles (editorials, columns, and op-eds)
- 663,173 comments in response to the articles
- 303,665 comment threads in response to the articles
- a subset of SOCC for constructiveness and toxicity.

DATA PIPELINE

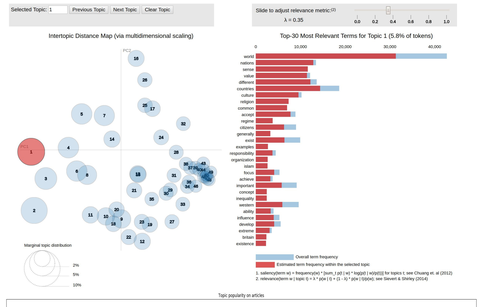
Corpus → Text Cleaning, Tokenizing, Removing Stop Words, Tagging → Modeling (LDA, NMF, Entity Extraction, Sentiment Analysis, Clustering) → Visualization

THE GLOBE AND MAIL | NLTK | Spark | NLTK | scikit learn | gensim | matplotlib | pyLDAvis

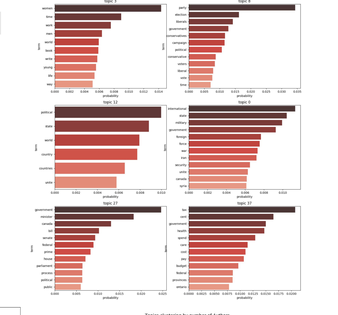
TEXT ANALYTICS

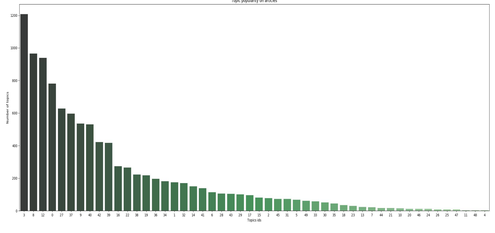
TOPIC EXTRACTION

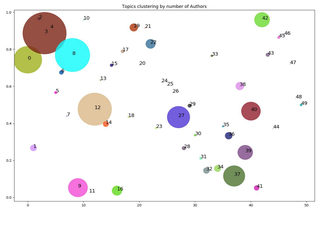
Topics clustering (Interactiv) by words(terms) frequency from 663,173 comments



The most popular 6 out of 50 topics from 10,339 articles







Example of the first 7 out of 50 topics from 10, 339 articles

Topic Ids	Topic words
0	**0.011*"international" + 0.010*"state" + 0.010*"military" + 0.009*"government" + 0.008*"foreign" + 0.008*"force" + 0.007*"war" + 0.007*"iran" + 0.007*"security" + 0.006*"unite" + 0.006*"canada" + 0.006*"syria"**
1	0.025*"government" + 0.018*"company" + 0.015*"public" + 0.012*"business" + 0.010*"service" + 0.010*"money" + 0.009*"pay" + 0.007*"board" + 0.007*"report" + 0.006*"bank" + 0.006*"private" + 0.006*"industry" + 0.0
2	0.007*"office" + 0.007*"man" + 0.007*"hes" + 0.006*"former" + 0.006*"show" + 0.005*"run" + 0.005*"staff" + 0.005*"media" + 0.004*"news" + 0.004*"guy" + 0.004*"week" + 0.004*"scandal" + 0.004*"way" + 0.004*"press
3	**0.014*"women" + 0.009*"time" + 0.008*"work" + 0.006*"men" + 0.006*"world" + 0.006*"book" + 0.006*"write" + 0.006*"young" + 0.005*"life" + 0.005*"way" + 0.004*"day" + 0.004*"feel" + 0.004*"love" + 0.004*"**
4	0.010*"tunisia" + 0.009*"isc" + 0.008*"burke" + 0.004*"readers" + 0.004*"insite" + 0.004*"pisa" + 0.003*"pluralism" + 0.003*"vivid" + 0.003*"transatlantic" + 0.003*"portland" + 0.003*"hoc" + 0.003*"accommodations" + 0
5	0.046*"obama" + 0.025*"president" + 0.021*"state" + 0.017*"american" + 0.016*"barack" + 0.015*"americans" + 0.014*"obamas" + 0.014*"republicans" + 0.013*"republican" + 0.012*"bush" + 0.012*"romney" + 0.009*"ho
6	0.020*"income" + 0.017*"children" + 0.012*"families" + 0.011*"work" + 0.010*"money" + 0.010*"cent" + 0.010*"poverty" + 0.009*"family" + 0.008*"child" + 0.008*"poor" + 0.007*"raise" + 0.007*"parent" + 0.007*"earn" + (

Analysis of the top article topics:
The most common topic “topic 3” (appeared in 1200 articles) talks about women, time , work, men etc. **expressing working women and feminism**. Next topic is “number 8” (appeared in 980 articles) which talks about party, election, liberals, government and etc. **expressing the election time and its important components**. Further “topic 12” (appeared in 960 articles) talks about world politics. followed by “topic 0” (appeared in 780 articles) which talks about **war, terrorism and defense**, “topic 27” (appeared in 610 articles) talking about **policy makers and their major projects** and “topic 37” (appeared in 600 articles) **mentioning the tax and economy**.

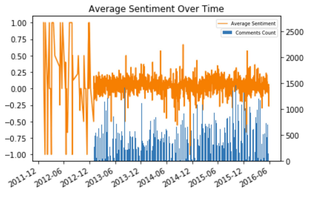
Analysis of the top comment topics: “topic 1” being the most common in comments express **canadian and world culture and values**. Next important topic(topic 2) is about **legal issues and its prosecution and government**. Followed by topic 3 about **jobs , salary and government**. followed by topic 4 about **governmental expenditures**. topic 5 tells us about the **canadian government and its economy.**

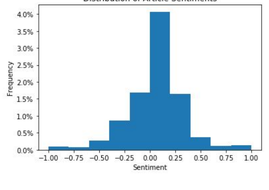
SENTIMENT ANALYSIS

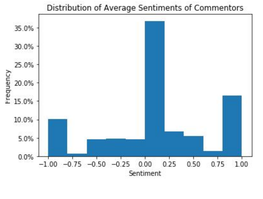
The sentiment scores vary widely before 2013 due to lack of data. Between 2013 and 2016, the trend of sentiment scores is rather stable (between -0.25 and 0.25). There are some sharp peaks and troughs reaching 0.5 and -0.5, and a time with highly negative sentiments in December 2014. These sharp peaks/troughs are likely due to major news events.

The distribution of sentiments across all articles is centered around 0 and resembles a normal distribution. This is expected as reporters are often expected to remain neutral when reporting.

The distribution of average sentiment of comments is again centered around 0. The majority of commenters made neutral comments overall. The distribution also shows a bipolar nature, with 10% of commenters consistently making negative comments and 15% consistently making positive ones.

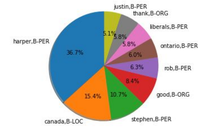




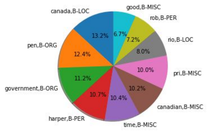


ENTITY EXTRACTION

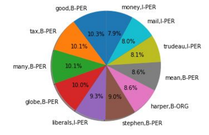
- These entities are extracted using the stanford-ner english corpus over nltk library as well as tagger project from github.
- Entities like Justin Harper (person), canada (location), good (organization), Trudeau (person) are the entities from articlea occurring most in the comments.
- Government (organization), rio (location), canada (location) are the maximum occurring entities in the articles.
- The entities in articles generally talk about the canadian economy , politics, as well as people, location and organization associated with them.
- The overlapping entities show us the common interest of the the commentators as well as the authors.



Entities from articles in comments



Entities in Articles



Entities in comments

VALUE DELIVERED

The project delivers a data product that details the insights from analyzing news articles with topic modeling methods. In particular, we present topics/entities/sentiments to find more about what the corpuses say, train a topic modeler and entity tagger, and visualize the data in various graphs and pie charts.

New skills learned during the project include data cleaning methods (tokenizing, stopword removal, stemming, and lemmatizing), new NLP algorithms (LDA, NMF, IOB tagging), python packages (NLTK, Sklearn, Spark-ML), and visualization packages (pyLDAvis, matplotlib).

Future extensions include using news API data to train the model for better entity tagging, scraping more data about the authors and commenters (demographics, location, occupation, etc) to gain more insights into the relationship between characteristics and personality of commenters the sentiments expressed in them.