Assessment and Visual Analysis of Trends using Article Reviews

Jaideep Misra Padmanabhan Rajendrakumar Jamshed Khan

Motivation

In the age of Big Data, there is a tremendous amount of textual information available through the internet. It has resulted in the following problems :

Information Overload

Difficulty in Organizing large amount of text

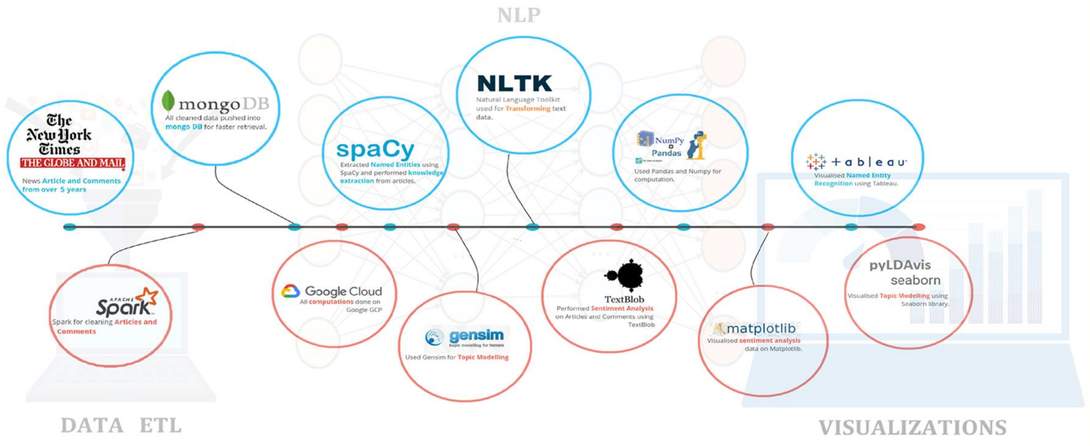
Mining topics from vast data

Understanding people's sentiments

Objectives

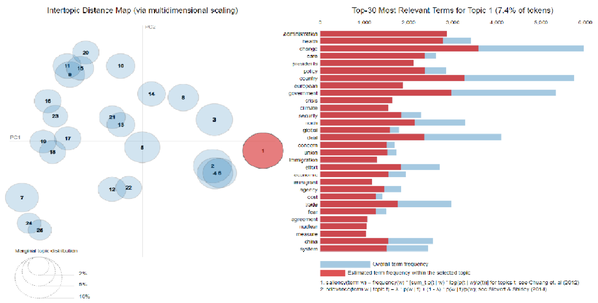
- Our primary aim is to analyse news topics and people's comments from leading news sites in order to understand the most prevalent and trending topics at a given time.
- To perform text-mining and discover the hidden semantic structures in a text body.
- Extraction of information and classification of named entities mentioned in this unstructured text into pre-defined categories such as person names, organizations etc.
- To examine people's feelings about current situations and happenings.(eg: politics, election, government, etc) from their comments and opinions.

Methodology



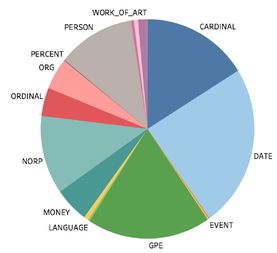
- Collection of news articles along with their comments and opinions from sites like "Globe and Mail" and "NY Times". Cleaning and transforming them with Apache Spark before storing them in MongoDB.
- Identification of the top 25 topics from text data by performing Topic Modelling with the 2 NLP methods of Latent Dirichlet Association(LDA) and Latent Semantic Indexing(LSI).
- Extraction of named entities such as people, organizations, countries, products etc. by using the pretrained model "en_core_web_lg" from SpaCy with an accuracy of 86.25%.
- Identification of people's reactions by performing sentiment analysis on the data. Using NLTK and TextBlob's sentiment libraries, identifying polarity and subjectivity of comments along with general trend of responses.
- Visualization of findings with Tableau, Seaborn, PyLDAvis and Matplotlib.

Visualizations



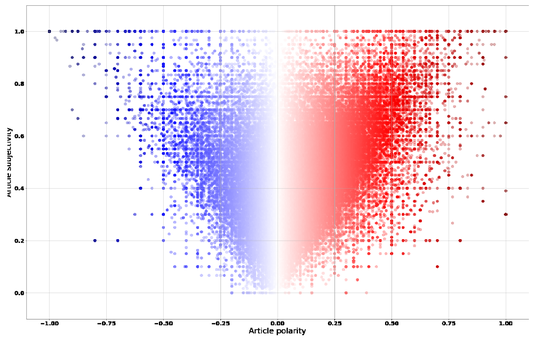
Results of the LDA model.

- The figure on the left depicts 25 topics and the distance between topics. The closer the bubbles get to each other, the higher the similarity between topics. The prevalence of topics is depicted by the size of the bubbles.
- The figure on the right shows the prevalence of top 30 words in topic 1. Similarly, we can analyse the prevalence of words for each topic.



Distribution of entities in text

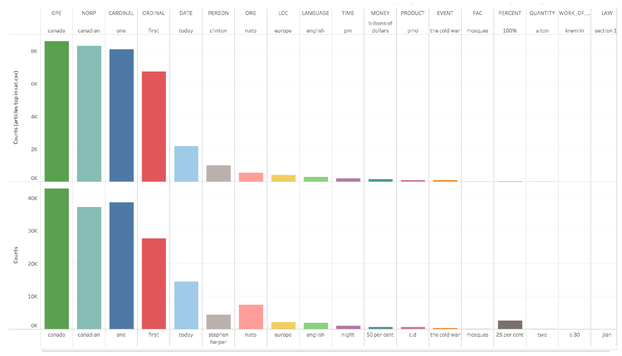
- This pie depicts the frequency of occurence of named entities in text.



- Polarity and Subjectivity of people's comments conveying emotion or sentiment.

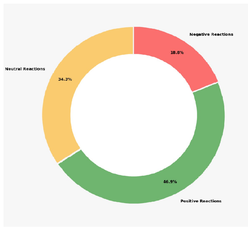
Word Cloud Visualization

- The most utilized words for the top 4 topics in news. Sizes of words denote frequency.



Named Entity Recognition on Articles and Comments

- The bar chart on top depicts the Top Entities present in each category of tags from news articles data.
- The one on the bottom denotes the same for comments data.



- Segregation of comments based on sentiment.

Learning

- Topics get closer to the point with the number of passes through the entire text data. In other words, the topic gets narrower when interating through the documents multiple times.
- Using NER and POS (Part of Speech) tags in the SpaCy model, we can extract more information other than just Named Entities. For instance extracting skills from resumes.
- LDA model is much slower than the LSI model and requires more computational resources but yields a higher accuracy and thus achieves better results overall.

Conclusion

Topic modelling is used when you have a huge number of documents and want to know what they are about without having to read them all. It has various real world applications in text classification, clustering, sentiment analysis etc.