Topic Modeling on Short Text Data using Latent Dirichlet Allocation

Paper : Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis

Authors : Rania Albalawi, Tet Hin Yeap and Morad Benyoucef
Year : 2020
Journal : Frontiers in Artificial Intelligence

Nidhi Gowri Srinath
801199302

Problem Statement

- With more number of online applications that are used on a daily basis, our source of news/information is being limited to the sources on the internet.
- These apps are further reducing our attention spans at alarming rates. Applications that have textual information now need a ‘trending’ section which can summarize the most common topics of discussion at a given instant of time.
- Including these topics can help users locate the topics of interest/concern.
- To identify these common topics, I have chosen to with machine learning and natural language processing, specifically topic modeling, to find the most commonly used words in textual data.
- Once we know these common words, topic modeling can then be extended to applications to embed more functionality to interact with them.

Data Set (Tentative)

- For this project, I am considering one of the two dataset options
  - A Kaggle dataset that contains a million news headlines published over a period of 19 years by the Australian Broadcasting Corporation (ABC) which can be found here.
  - The twitter API ‘Tweepy’ which allows us to pull and refine the data related to users and their tweets on Twitter. This API requires authentication of the user trying to use it. Documentation for Tweepy can be found here.

As I personally do not use Twitter, I would have to work without much knowledge about how the API and the application work. I have therefore chosen two dataset options out of which one will be implemented in the final project.

Motivation

My primary reason for working on this project is to use Natural Language Processing Techniques to simplify tasks in everyday circumstances. For a long time, I have been consuming electronic news through apps. A news app that I use utilizes a brief description of everyday occurrences that users may swipe through to learn about. On some days, when I want to read a specific article, I have to sift through a number of headlines before I find what I'm looking for. An application that attracted my curiosity was one that used NLP methods with text to discover relevant information in a pile of data.

Survey on Related Work

- Paper 1 - Albalawi, Rania, et al., “Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis”, Frontier in Artificial Intelligence, 2020, DOI : 10.3389/frai.2020.00042

  Link here

- Paper 2 - Jelodar, Hamed, et al., “Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey”, arXiv, 2018, arXiv:1711.04305

  Link here

Motivation to choose the papers

- My primary reason to choose my first paper is that the authors have a similar motivation to explore topic modeling for short text data. They aspire to find ways to fetch important details from the abundance of information available on the internet. They also work with multiple techniques for topic modeling and evaluate them based on basic statistical evaluation metrics. They conclude upon trying 5 topic modeling methods that Latent Dirichlet Allocation provides the best results.
- The motivation to choose my second paper is to have a more deeper understanding of LDA as I have chosen that as my primary method to implement Topic Modeling. The authors discuss the importance of text mining and advantages of using LDA.

Paper Summary

“Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis”

- The authors intend to find application areas, methods and tools for topic modeling using machine learning and natural language processing techniques.
- They also compare five commonly used topic modeling techniques, namely; latent semantic analysis, latent Dirichlet allocation, non-negative matrix factorization, random projection, and principal component analysis.
- To evaluate these techniques, the authors consider statistical evaluation metrics such as f-score, precision, and topic coherence.
- As a result of the tests conducted by the authors, they found that Latent Dirichlet Allocation (LDA) had the most valuable results followed by Non-Negative Matrix Factorization ((NMF)

Paper Summary

“Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey”

- This paper focuses on the Latent Dirichlet Allocation technique for textual mining and topic modeling.
- The authors discuss a variety of scholarly articles between 2003 and 2016. These resources provide numerous factors to take into consideration while working with LDA.
- The paper discusses various applications and fields in which LDA and text mining could make a significant impact. Some of which include political science, social media applications, software engineering, geographic information, linguistic science, etc.

Project Timeline

- **Week 1** - Topic finalization and dataset research for Topic modeling.
- **Week 2** - Selecting and studying research papers to finalize the technology to be used and create a project proposal.
- **Week 3** - Research and install the required softwares and packages. Understand existing codes and the mathematics behind LDA.
- **Week 4** - Experiment and code for topic modeling to analyse and visualize the results.
- **Week 5** - Work on hyper parameter tuning and create visualizations to help with understanding the results obtained with LDA for text mining and topic modeling.

Expectations from the project

- To work on NLP techniques and understand concepts such as lemmatization, stemming, bag of words, topic modeling etc.
- Understanding how to integrate visualization techniques to help use the results obtained from topic modeling
- Understanding the areas of application for text mining.
- Understanding how to integrate the results of this project with real-world problems and applications
- Attempt further use of the topic modeling techniques on different domain data to achieve diverse results