IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024

DOI: 10.17148/IJARCCE.2024.13589

Topic Modeling With Latent Dirichlet Allocation(LDA) using Machine Learning

Karishma Borse[1], Pingale Divya Vijay[2], Mahajan Pornima Dattatraya[*3,] Patil Komal Vinod[4]

Department of Computer Engineering, R. C. Patel Institute of Technology, Shirpur[1-4]

Abstract: Topic modeling is a very efficient data mining technique for mining text, latent data identification, and establishing links between text documents and data. Many studies in this field have been published by researchers, and these findings have been implemented in linguistic science, software engineering, political science, and medicine, among other fields.

Keywords: Topic Modeling, Latent Dirichlet Allocation, Machine Learning, Applications.

I. INTRODUCTION

Natural language processing uses topic modeling approaches extensively for topic identification and semantic mining from unordered materials. They are powerful, intelligent algorithms. Broadly speaking, text mining, information retrieval, and natural language processing have all benefited from the use of topic modeling techniques based on LDA. For example, topic modeling based on social media analytics makes it easier to comprehend the responses and exchanges among members of online communities and to identify patterns that may be easily understood from such interactions.

A difficult area of computer science study is called natural language processing, or NLP, which gives computers the ability to understand meaning from human language processing in text texts. In computer science, topic models are crucial for natural language processing and text mining. Any disorganized text, including emails, blog posts, journal papers, book chapters, and diary entries. Subject models are unable to comprehend the terminology and meanings included in text documents. Conversely, they presume that any part of the content is put together by selecting terms from word baskets that are likely to be related to a certain issue.

II. TOPIC MODELING AN OVERVIEW

One text-mining technique that is widely utilized to find latent semantic patterns in a text body is topic modeling. It seems sense that certain words would appear more or less frequently in a document about a given topic. A topic modeling is machine learning approach uses text for analysis and automatically find cluster of words for a collection of texts.

It uses unsupervised machine learning techniques. Topic modeling is a simple approach to examine your data because it does not require any training. But you never be sure about the results you get they can be accurate, that is why a lot of companies first spend time on training a topic categorization model. It is captured by a topic model.

III. TECHNIQUES USED FOR TOPIC MODELING

A. Classification: The purpose is to adapt a well-known framework to fresh data. The selection process for multiple groups was based on a series of training data with findings for which the group is known.

B. Clustering: Analysing a cluster is a task of involving several connected components within a single community (referred to as a cluster). Without utilizing pre-existing data structures, the task is to find categories and hierarchies in the data that are "like" in any way.

C. Communication rule learning: One well-liked method for identifying correlations between variables in large datasets for focus is this technique. It searches for relationships among variables.

D. Detection of anomaly: Also referred to as Outliner detection. The work involves looking for unusual data documents, actions, or annotations that might be interesting or data mistakes and identifying them or exploring them further.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 5, May 2024
DOI: 10.17148/IJARCCE.2024.13589

E. Summarization: A machine software's auto-summary feature condenses a text document to provide a synopsis of the main points of the original content. It involves presenting and reporting together with a more condensed presentation of the data bundle. The growing amount of knowledge and better quality data has led to a growth in interest in automated synthesis.

IV. TOPIC MODELING APPLICATIONS

Large amounts of data are produced now a days in every industry, including organizational and personal data. These acquired data are crucial resources that can be analyzed and assessed to extract information for use in making decisions and cutting expenses. It is possible to divide the topic modeling technology into different categories.

A. Text classification:

The process of identifying several themes within a set of documents is known as topic detection. One method of topic detection involves assigning a subject to each document within the corpus. The term "topic" refers to any word or collection of words that expresses the focus of the work.

B. Sentiment Analysis:

As the World Wide Web has grown and gained traction, sentiment analysis has grown in popularity as a field of study for web data analysis and information retrieval. Sentiment analysis, which deals with the extraction of opinions and sentiments, has drawn academics from academia and business because to the massive amount of user-generated content on blogs, forums, social media, etc.

C. Summarization:

In order for businesses to enhance their goods and for governments to improve services, it is critical to summarize the recently uncovered viewpoints. Sentence-level topic detection is comparable to opinion detection because no questions are asked beforehand.

V. CONCLUSION

The use of topic models in text mining is significant in computer science. Word lists that appear in statistically significant ways are called topics in topic modeling. Emails, journal articles, blog posts, book chapters, and other unstructured texts can all be considered texts. It is assumed that any part of the information is assembled by finding words from likely word clusters, each of which is associated with a topic. It remains on the allocation of words into clusters that call themes, the tool repeatedly goes through this procedure.

REFERENCES

[1] Hindle, A., Campbell, J. C, Hindle, A., Stroulia, A. (2014). Latent Dirichlet Allocation: Extracting Topic. The Art and Science of Analyzing Software Data.
[2] Sun, X. (2014). Textual Document Clustering Using Topic Models. 10th International Conference on Semantics. Knowledge and Grids. 1-4.
[3] Feldman, R., and Sanger, J. (2007). The Text Mining Handbook. Cambridge: University Pers. New York.
[4] Hong, L., and Davison, B. (2010). Empirical Study of Topic Modeling in Twitter. 1st Workshop on Social Media Analytics (SOMA'10).
[5] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng (2017). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey.