7-1-2023

Finding the Most Interpretable Topic Modeling Approach

Alex Algarra

Frank Lee

Finding the Most Interpretable Topic Modeling Approach

Alex Algarra
Georgia State University
aalgarra1@student.gsu.edu

Frank Lee
Georgia State University
flee@gsu.edu

ABSTRACT

As the number of unstructured data increases, many machine-learning techniques, including topic modeling and sentiment analysis, have replaced a traditional manual approach to analyzing textual data. Specifically, Latent Dirichlet Allocation (LDA), one of the topic modeling algorithms, is widely used to discover hidden semantic patterns in a large, relatively unstructured document corpus. The output of a topic model can only be as good as its input. Therefore, when building a topic model, it is essential to use the most informative features of the corpus. Some researchers suggested that building a topic model on a noun-only corpus may improve the model's performance. The suggestion of this noun-only approach is based on the fact that nouns are more informative of a document's content than other parts of speech, such as adjectives, adverbs, or verbs. While the noun-only approach may be informative in some areas, it may not always be instructive in many contexts and big data. This study aims to find and propose the most interpretable topic modeling approach by comparing different social media text versions: the Original text, the Lemmatized, and the Noun-only.

Keywords

Topic Modeling, Latent Dirichlet Allocation, Machine Learning, Text Analytics

EXTENDED ABSTRACT

A majority of data, over 80%, is unstructured information like text, video, and social media (Harbert, 2021), and they are waiting to be analyzed. These undiscovered data also can be a substantial resource with the potential to create a competitive advantage for many organizations. As the number of unstructured data increases, many machine-learning techniques, including topic modeling and sentiment analysis, have replaced a traditional manual approach to analyzing textual data. Specifically, Latent Dirichlet Allocation (LDA), one of the topic modeling algorithms, is widely used to discover hidden semantic patterns in a large, relatively unstructured document corpus. LDA discovers a topic that a group of words characterizes, and then the topics identified illustrate a document description. This model helps to express a document's semantic content, allowing a qualitative description of the document.

Text documents contain hidden semantic patterns called topics, and each of these topics is defined by a probability distribution over a fixed set of words (Blei et al., 2003). The output of a topic model can only be as good as its input. Therefore, when building a topic model, it is essential to use the most informative features of the corpus. Some researchers suggested that building a topic model on a noun-only corpus may improve the model's performance (Martin and Johnson, 2015). The suggestion of this noun-only approach is based on the fact that nouns are more informative of a document's content than other parts of speech, such as adjectives, adverbs, or verbs. While the noun-only approach may be informative in some areas, it may not always be instructive in many contexts and big data.

This study aims to find and propose the most interpretable topic modeling approach by comparing different social media text versions: the Original text, the Lemmatized, and the Noun-only. This study collected over 50,000 Twitter text data on employee attrition and analyzed the data using LDA. The first model was generated from the raw data, and the second was generated from the lemmatized version. Finally, the third model was generated from the lemmatized data reduced to nouns only. The process and result of three different versions of the LDA analysis were recorded and compared. The findings and implications of this study will be presented at the conference.

REFERENCES

1. Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," J. Mach. Learn. Res. (3), pp. 993–1022.
2. Harbert, T. 2021. Tapping the power of unstructured data. MIT Sloan. Feb 1, 3.
3. Martin, F. and Johnson, M. 2015. More Efficient Topic Modelling Through a Noun Only Approach. In Proceedings of the Australasian Language Technology Association Workshop 2015, pages 111–115, Parramatta, Australia.