Topic Modeling using Amazon Reviews Dataset

Dinesh Vardhan Bonthu
Computer Science
Wright State University
Fairborn, Ohio
bonthu.8@wright.edu

I. Introduction

Topic modeling is a powerful technique in natural language processing (NLP) that allows us to discover latent themes or topics within a collection of documents. By extracting meaningful patterns and relationships from textual data, topic modeling provides valuable insights into the underlying structure and content of the corpus.

In this project, we present a program implemented in Jupyter Notebook that performs topic modeling on a dataset. The dataset is organized in a DataFrame format, with columns labeled as 'Sentiment Score', 'Head', and 'Body'. The program follows a systematic workflow to preprocess the text, create n-gram models, and ultimately build an LDA (Latent Dirichlet Allocation) model to extract topics.

The initial steps of the program involve data cleaning and preprocessing. This includes removing URLs and applying various text cleaning techniques to ensure the quality of the input data. Once the text is preprocessed, the program generates bigram and trigram models to capture meaningful word combinations that contribute to topic identification.

Furthermore, the program applies lemmatization to reduce words to their base or root form, facilitating more accurate topic modeling. The result of this preprocessing step is a refined and normalized text corpus, ready for further analysis.

To gain a visual representation of the most frequent words in the dataset, the program generates a word cloud. This visualization helps identify prominent terms and provides a general overview of the dataset's content.

The next step in the program involves finding the dominant topics within the dataset. This is achieved by running the preprocessed corpus through an LDA model. The LDA algorithm employs probabilistic inference to discover latent topics and assigns relevant words to each topic. By analyzing the output of the LDA model, the program determines the dominant topics in the dataset, allowing users to gain a deeper understanding of the underlying themes.

In addition to finding dominant topics, the program also enables users to print documents associated with a chosen topic. This feature allows researchers and data scientists to examine specific documents that contribute to a particular topic, providing context and supporting further analysis.

To assess the quality of the topic modeling results, the program calculates perplexity and coherence scores. Perplexity measures how well the LDA model predicts unseen data, with lower values indicating better performance. Coherence score, on the other hand, evaluates the semantic coherence of the extracted topics, with higher scores indicating more coherent topics.

Finally, the program provides a visual representation of the topics using pyLDAvis, a Python library for interactive topic model visualization. pyLDAvis allows users to explore the topics, their associated keywords, and the intertopic distance in an intuitive and interactive manner.

Through this project, we aim to demonstrate the effectiveness of topic modeling techniques in extracting meaningful insights from textual data. By leveraging the power of NLP and machine learning, researchers and data scientists can uncover hidden patterns, identify dominant topics, print associated documents, evaluate model performance, and visualize topics for deeper analysis within vast collections of documents.

II. Methods

A. Data Collection and Preparation

The first step in our topic modeling program involves gathering the necessary data and organizing it into a DataFrame. For this project, we utilized the ‘train.csv’ data file obtained from the amazon reviews dataset in Kaggle platform(https://www.kaggle.com/datasets/kritanjalijain/amazon-reviews). This dataset contains a large collection of Amazon product reviews. To focus our analysis, we considered the first 100,000 reviews from the dataset for topic modeling. By limiting the number of reviews, we aim to maintain a manageable dataset size while still capturing a substantial amount of information. We set up the columns of the DataFrame as 'Sentiment Score', 'Head', and 'Body', which allow us to store relevant information for each document.

B. Data Preprocessing

To ensure the quality and consistency of the text data, we perform various preprocessing tasks. The program begins by removing any URLs present in the 'Body' column, as they typically do not contribute to topic identification. Next, we apply text cleaning techniques such as removing punctuation, converting text to lowercase, and handling special characters to standardize the text across documents.

C. Creating N-gram Models

N-grams are contiguous sequences of n words. In our program, we create both bigram and trigram models to capture meaningful word combinations. These models identify commonly occurring phrases or expressions that might convey specific topics more effectively than individual words alone.

D. Lemmatization and Wordcloud Generation

To further refine the text, we employ lemmatization, which reduces words to their base or root form. This step helps eliminate variations of the same word, improving the accuracy of topic modeling. After lemmatization, we

generate a word cloud to visualize the most frequent words in the dataset. This graphical representation provides a quick overview of the dominant terms within the corpus.

E. LDA Model Building

The core of our topic modeling program involves building an LDA (Latent Dirichlet Allocation) model. LDA is a probabilistic model that assigns topics to documents and words to topics. By leveraging statistical inference techniques, the LDA model uncovers latent topics in the dataset. We use the Gensim library to implement the LDA algorithm, specifying the number of topics to be extracted.

F. Topic Extraction and Presentation

Once the LDA model is trained, we extract the resulting topics and their corresponding word distributions. These topics represent the underlying themes in the dataset. The program then prints out the identified topics, allowing users to interpret and understand the key themes within the collection of documents.

G. Finding Dominant Topics and Printing Documents

To provide further insights, our program allows users to find the dominant topics within the dataset. By assigning the most probable topic to each document, users can identify the main theme associated with each piece of text. Additionally, users can choose a specific topic and print out the documents associated with that topic, facilitating closer examination and analysis.

H. Perplexity and Coherence Score Calculation

To evaluate the quality of the LDA model, our program calculates perplexity and coherence scores. Perplexity measures how well the model predicts unseen data, with lower values indicating better performance. Coherence score evaluates the semantic coherence of the extracted topics, with higher scores indicating more coherent topics.

I. Visualizing Topics using pyLDAvis

To provide a visual representation of the topics, our program utilizes the pyLDAvis library. This interactive visualization tool allows users to explore the topics, their associated keywords, and the inter topic distance. The visualization enhances the understanding of the relationships and distribution of topics within the dataset.

Through these methods, our topic modeling program enables the systematic analysis of textual data, providing valuable insights into the latent topics present in the dataset. Users can preprocess the text, create n-gram models, build an LDA model, extract and interpret topics, find dominant topics, print associated documents, calculate evaluation metrics, and visualize the topics using pyLDAvis.

III. RESULTS

I have conducted several trial-and-error methods, considering perplexity and coherence scores, to determine the optimal number of topics, which resulted in selecting eight topics. Considering more than eight topics made the extra topics small and irrelevant. The model's output is summarized in the table shown below:

Topic	***Keywords***	***Topic name***
0	product, purchase, order, month, receive, week, send, wrong, card, shoe	Consumer Goods
1	song, album, music, sound, hear, fan, version, listen, track, favorite	Music
2	bad, people, ever, s, star, part, live, job, copy, full	Negative Experience
3	learn, series, style, stuff, piece, rather, name, person, collection, description	Miscellaneous
4	buy, use, well, work, look, try, come, many, need, put	General usage
5	book, read, find, story, write, character, end, author, life, well	Literature
6	good, get, great, time, make, movie, go, love, think, really	Positive Experience
7	drive, film, man, real, big, woman, early, fit, else, case	Entertainment

Fig. 1. Topic Number, Keywords and Topic Names.

The dominant topics and their distribution are presented in the following table, accompanied by a chart visualizing the distribution:

Topic	***Number of Documents***
6	44,625
5	36,710
4	16,702
1	1,782
0	122
7	34
2	20
3	5

Fig. 2. Topics and Number of Dominant Documents.

Furthermore, the program allows for printing all the documents in which a specific topic is dominant, providing a comprehensive understanding of the chosen topic.

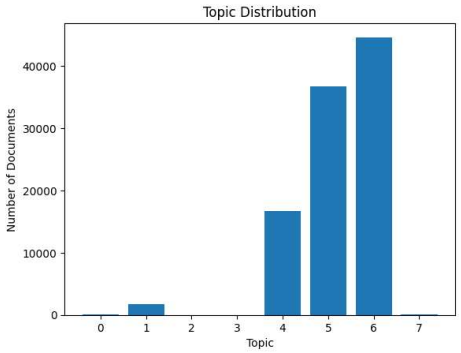

Fig. 3. Topic Distribution – Topics vs Number of Documents.

The perplexity score of the model is -8.43120893241583, indicating its ability to predict unseen data. Additionally, the coherence score is 0.46693764308248253, which assesses the semantic coherence of the extracted topics.

To provide an intuitive and interactive visualization of the topicss, pyLDAvis, a Python library, was utilized. pyLDAvis enables users to explore the topics, their associated keywords, and the intertopic distance, facilitating a deeper analysis of the topic modeling results.

IV. Discussion

The results of the topic modeling analysis reveal valuable insights into the underlying themes present in the dataset of Amazon reviews. By selecting eight topics based on perplexity and coherence scores, we were able to capture distinct patterns and dominant themes within the dataset.

Topic 0, characterized by keywords such as "product," "purchase," and "order," suggests a focus on customer experiences related to the buying process. This topic provides insights into issues such as incorrect orders and delays in receiving products.

Topic 1 revolves around music-related discussions, with keywords like "song," "album," and "music" being prominent. This indicates that customers frequently express their opinions and preferences regarding songs and albums they have purchased.

On the other hand, Topic 5, which primarily consists of keywords like "book," "read," and "story," highlights the importance of reviews and opinions on literature. It suggests that customers often share their thoughts on books they have read, including aspects such as the writing style, characters, and overall story.

Additionally, Topic 6 focuses on positive sentiments, with keywords like "good," "great," and "love" being prevalent. This suggests that customers frequently express satisfaction and appreciation for various products or services. Also, Topic 6 is the most dominant topic, out of 1 lakh documents it is dominant in 44,625 documents.

The ability to print all documents associated with a chosen dominant topic provides researchers and data scientists with a deeper context for analysis. This feature enables further investigation into specific themes and allows for a more comprehensive understanding of customer sentiments and preferences.

The perplexity score of -8.43120893241583 indicates that the LDA model performs well in predicting unseen data, demonstrating its effectiveness in capturing the underlying structure of the dataset. Furthermore, the coherence score of 0.46693764308248253 suggests that the extracted topics exhibit a reasonable degree of semantic coherence.

Overall, the application of topic modeling techniques, combined with the utilization of pyLDAvis for topic visualization, has allowed us to uncover hidden patterns, identify dominant themes, and gain meaningful insights from the vast collection of Amazon reviews. These findings can be valuable for businesses to understand customer sentiments, improve their products or services, and make informed decisions based on customer feedback.

Conclusion and Acknowledgment

In conclusion, this project demonstrates the effectiveness of topic modeling techniques in extracting valuable insights from textual data. By analyzing a dataset of Amazon reviews using the LDA model, we were able to identify eight dominant topics and gain a deeper understanding of the underlying themes present in the reviews. The systematic workflow, including data preprocessing, n-gram modeling, and LDA topic extraction, provided a structured approach to uncovering hidden patterns and relationships within the dataset. The use of pyLDAvis for topic visualization enhanced the interpretability of the results, allowing for interactive exploration of the topics and their associated keywords.

The results of this project shed light on various aspects of customer experiences, including product purchases, music preferences, literature reviews, and positive sentiments. The ability to print documents associated with specific topics further supports researchers in conducting in-depth analysis and deriving meaningful insights. The calculated perplexity and coherence scores validate the performance of the topic modeling approach and indicate its reliability in predicting unseen data and generating coherent topics.

I would like to express our gratitude to the creators of the Kaggle dataset "Amazon Reviews" for providing the valuable data used in this project. Their efforts in collecting and sharing this dataset have facilitated my research and analysis.

I would also like to thank the developers of the Python libraries and tools used in this project. Their contributions to the open-source community have greatly enriched the field of natural language processing and made our work possible.

Lastly, I would like to acknowledge the guidance and support provided by professor Tanvi Banerjee throughout this project. Her expertise and feedback have been instrumental in shaping our approach and refining our results.

References

[1] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics.

[2] McAuley, Julian and Jure Leskovec. “Hidden factors and hidden topics: understanding rating dimensions with review text.” Proceedings of the 7th ACM conference on Recommender systems (2013): n. pag.

[3] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09).

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.