NLP - Take Home Assignment

Project Title: Topic Discovery

Your Name: Srinivasulu Devanuru
Date: 26-07-2024

Introduction

Brief Overview of the Problem

The goal of this project is to discover and visualize topics from a collection of Netflix reviews. This helps in understanding the main themes and sentiments expressed in the reviews.

Problem Statement

- The submission should allow the user to run the code and obtain the topics deemed important.
- How are you presenting the results back to the user?
- How do you validate the quality of your topics?
- Can the user “accept” or “reject” certain topics? How do you handle a new topic run if the user wants to accept certain topics?
- Can you present a topic in a hierarchical way?

Objectives

- Preprocess the text data to prepare it for topic modeling.
- Use Latent Dirichlet Allocation (LDA) to discover topics in the text data.
- Allow user interaction to refine the discovered topics.
- Visualize the topics using various techniques.

Data Description

Source of the Data

The data is sourced from Kaggle: [Netflix Reviews - Playstore Daily Updated]

Key Features and Their Descriptions

- `content`: The main text of the reviews.

Data Preprocessing

Steps Taken to Clean and Preprocess the Data

- **Lowercased the text**: Converting all text to lowercase to ensure uniformity.
- **Removed punctuation and numbers**: Eliminating non-alphabetic characters to clean the text.
- **Removed stopwords**: Filtering out common stopwords to retain meaningful terms.
- **Lemmatized the words**: Reducing words to their base forms to normalize the text.
- **Created bigrams and trigrams**: Constructing multi-word expressions to capture common phrases.

Feature Engineering and Transformations Applied

- Creation of bigrams and trigrams to capture meaningful multi-word expressions which enhance the context of the text.

Methodology

Description of the Algorithms Used

- **Latent Dirichlet Allocation (LDA)**: A generative probabilistic model used to discover abstract topics within a collection of documents by classifying words into topics.

Justification for the Chosen Methods

LDA is an effective method for topic discovery in large text corpora due to its ability to produce interpretable and coherent topics, making it suitable for understanding underlying themes in text data.

Implementation

Tools and Libraries Used

- **pandas**: For data manipulation and analysis.
- **gensim**: For topic modeling using LDA.
- **nltk**: For text preprocessing tasks.
- **matplotlib** and **plotly**: For data visualization.

Key Steps in the Implementation Process

1. **Loading and Preprocessing Data**: Reading the data and applying preprocessing steps to clean and prepare the text for modeling.
2. **Training the LDA Model**: Utilizing the preprocessed text to train the LDA model and discover topics.
3. **User Interaction**: Enabling users to accept, reject, or modify topics based on their relevance and coherence.
4. **Re-training the Model**: Refining the model with the accepted topics to enhance the quality of the results.
5. **Visualization**: Using dendrograms, word clouds, and sunburst charts to visually represent the topics and their hierarchical relationships.

Code Snippets

For detailed code and step-by-step implementation, please refer to the accompanying Jupyter notebook.

Evaluation

Metrics Used to Evaluate the Models

Coherence Score: Evaluates the semantic similarity of words within a topic, with higher scores indicating more coherent topics.

Perplexity: Measures the model's ability to predict a sample, with lower scores indicating better performance.

Performance Results

- **Initial Coherence Score**: 0.49
- **Refined Coherence Score**: 0.49
- **Initial Perplexity**: -6.77
- **Refined Perplexity**: -6.70

Results and Discussion

Key Findings from the Analysis

The topics discovered provide valuable insights into common themes in Netflix reviews, such as user experience, content quality, and service-related issues.

Interpretation of Results

Refined topics, as determined through user interaction, demonstrate improved coherence and better capture the main themes present in the reviews.

Conclusion

Summary of What Was Accomplished

- Successfully discovered and visualized topics from Netflix reviews using LDA.
- Implemented user interaction to refine the topics, resulting in more coherent and relevant themes.
- Utilized various visualization techniques to present the topics comprehensively.

Main Takeaways

- LDA is a robust method for uncovering latent topics in text data.
- User interaction plays a crucial role in enhancing the quality of topic modeling.

Suggestions for Future Work

- **Advanced Topic Models**: Explore models like BERTopic or neural network-based methods for improved topic representation.
- **Sentiment Analysis**: Integrate sentiment analysis to understand the emotions behind each topic.
- **Time-Series Analysis**: Investigate how topics evolve over time to identify trends in user reviews.