Topic Modeling - Product Reviews

Project Documentation

https://www.linkedin.com/in/raviteja-padala/ https://github.com/raviteja-padala

1. Introduction

Welcome to the documentation for the "Topic Modeling - Product Reviews" project. This documentation provides a comprehensive overview of the project, its objectives, installation instructions, and how to use the web interface. Additionally, it explains the data preprocessing and topic modeling techniques employed, along with guidelines for interpreting the results.

2. Project overview and objectives

Project Name:

Topic Modeling - Product Reviews

Project Overview:

The "Topic Modeling - Product Reviews" project aims to provide a user-friendly web application for analyzing and understanding customer reviews of various products, focusing on health supplements. The project involves web scraping, data preprocessing, topic modeling, and topic labeling to uncover valuable insights from user-generated reviews.

Project Objectives:

1. Develop a web interface for users to select a product category and view insights from product reviews.

2. Extract reviews and related information from an e-commerce website (e.g., Flipkart) using web scraping techniques.

3. Preprocess the text data to clean and prepare it for analysis, including lowercase conversion, punctuation removal, stopword removal, tokenization, and lemmatization.

4. Implement Latent Dirichlet Allocation (LDA) topic modeling to identify latent topics within the reviews.

5. Assign meaningful labels to each identified topic based on significant keywords.

6. Present the results, including the product name, the total number of reviews extracted, assigned topic labels, identified topics with significant keywords, and topic weights (proportions).

3. Installation and setup instructions

Prerequisites: - Python 3.7 - Pip (Python package manager)

Steps:

1. Clone the project repository from GitHub:

```
git clone
<https://github.com/raviteja-padala/Topic_Modeling-Product_Reviews.git>
```

2. Navigate to the project directory:

```
cd Topic-Modeling-Product-Reviews
```

3. Install project dependencies:

```
pip install -r requirements.txt
```

4. Run the Flask application:

```
python app.py
```

5.Access the web interface by opening a web browser and navigating to `` `http://localhost:5000` ``

4. How to use the web interface

1. Upon accessing the web interface, you will see a product selection form. Choose a product from the predefined list
2. Click the "Submit" button to initiate the analysis process.

3. The application will extract reviews, preprocess the text data, perform topic modeling, and assign labels to topics. You will be presented with the following results:

- Product name

- Total number of reviews extracted

- Assigned topic labels for each topic

- Identified topics with significant keywords

- Topic weights (proportions) indicating the importance of each topic

4. To analyze reviews for a different product, simply return to the homepage and select a new product.

5. Project steps

This documentation outlines the step-by-step process involved in the "Topic Modeling - Product Reviews" project. This project aims to provide insights into user reviews for various products, focusing on health supplements. By utilizing web scraping, natural language processing (NLP), and Latent Dirichlet Allocation (LDA) topic modeling, this application extracts, analyzes, and labels latent topics within the reviews.

1. Product Selection

Objective: Create a user-friendly interface that allows users to select a product from a predefined list.

Description: In this initial step, users are presented with a user-friendly interface that lists predefined product categories. Users can select a product category from this list to proceed with the analysis.

2. Data Extraction

Objective: Implement web scraping to extract user reviews and related information for the selected product from an e-commerce website

Description: This step involves web scraping techniques to extract user-generated content, including product reviews, ratings, and comments, from the selected e-commerce website. The progress message is displayed to indicate that reviews are being extracted.

3. Data Preprocessing

Objective: Preprocess the extracted text data to clean and prepare it for analysis.

Description: Data preprocessing is crucial to ensure that the extracted text data is suitable for analysis. The following preprocessing steps are performed:

- Lowercasing the text : Convert all text to lowercase.
- Removing punctuation and special characters: Remove special characters , punctuation.
- Removing stopwords: Eliminate common stopwords from the text.
- Tokenizing the text: Split the text into individual words.
- Lemmatizing words to their base forms : Reduce words to their base or root form.

4. Topic Modeling (LDA)

Objective: Train an LDA (Latent Dirichlet Allocation) model on the preprocessed text data to identify latent topics within the reviews.

Description: Latent Dirichlet Allocation (LDA) is applied to the preprocessed text data to identify hidden topics within the reviews. The LDA model helps in understanding the underlying themes discussed in the reviews.

5. Topic Labeling

Objective: Define a set of rules to assign meaningful labels to each identified topic based on significant keywords. Automatically generate labels for each topic using the rules.

Description: In this step, a set of predefined rules is used to assign meaningful labels to the identified topics based on significant keywords extracted from the reviews. Labels are generated automatically based on the rules defined for each topic.

6. Results Presentation

Objective: Display the product name, show the total number of reviews extracted, display the assigned topic labels for each topic, present the identified topics along with their significant keywords, and show the topic weights (proportions) indicating the importance of each topic.

Description: The results are presented to the user in a clear and informative manner. Users can see the following information:

- The product name.

- The total number of reviews extracted.

- The assigned topic labels for each identified topic.

- A list of identified topics along with their significant keywords.

- Topic weights, indicating the importance of each topic in the reviews.

7. User Interaction

Objective: Allow users to repeat the process for different products if desired. Provide clear instructions and progress feedback throughout the analysis.

Description: The application allows users to analyze reviews for multiple products seamlessly. Users can return to the product selection step and choose a different product category for analysis. Clear instructions and progress feedback are provided to guide users throughout the analysis.

8. Error Handling

Objective: Implement error handling and provide informative error messages in case of issues during data extraction or analysis.

Description: Robust error handling mechanisms are in place to handle unexpected issues that may arise during data extraction or analysis. Informative error messages are displayed to guide users in case of errors.

9. Documentation

Objective: Document the project, including the code and the steps involved, for future reference.

Description: Comprehensive documentation is created to ensure that the project's code and procedures are well-documented. This documentation serves as a reference for developers and users.

10. Deployment

Objective: Deploy the application on a web server or hosting platform to make it accessible online.

Description: The application is deployed to a web server or hosting platform, making it accessible to users online. This deployment ensures that users can access the application without the need for local installation.

These documented project steps provide a comprehensive overview of the "Topic Modeling - Product Reviews" project, from data extraction to results presentation and user interaction. The project aims to simplify the process of extracting meaningful insights from product reviews, enhancing user understanding and decision-making.

6. Explanation of Topic modeling

Topic Modeling (LDA):

- **Latent Dirichlet Allocation (LDA):** In this project, we employ Latent Dirichlet Allocation (LDA), which is a powerful natural language processing technique. LDA helps us uncover hidden or latent topics within the product reviews. These latent topics represent underlying themes and subjects that users discuss when reviewing a product.

- **LDA Model Training:** To identify these latent topics, we train an LDA model using the preprocessed text data extracted from the product reviews. The training process involves analyzing the frequency and co-occurrence of words and phrases within the reviews to discover patterns that correspond to different topics.

- **Document-Topic Distribution:** Once the LDA model is trained, we calculate the document-topic distribution. This distribution tells us the degree to which each review is associated with various topics. In other words, it helps us understand which topics are prevalent in each review.

- **Topic Proportions:** We also calculate topic proportions, which indicate the importance or weight of each topic in the entire set of reviews. This allows us to identify the most dominant topics and their significance in the overall discussion.

- **Extraction of Significant Keywords:** To make the topics more interpretable, we extract significant keywords from the LDA model for each identified topic. These keywords represent the essential terms and phrases that define each topic's content. These keywords serve as valuable cues for understanding the topics' themes.

- **Topic Labeling Rules:** To further enhance the user experience, we define a set of rules for assigning meaningful labels to the topics based on these significant keywords. These rules help automate the process of summarizing the topics and provide users with clear, descriptive labels that convey the essence of each topic.

In summary, the LDA-based topic modeling approach employed in this project enables us to dissect the product reviews into coherent topics, each characterized by its unique set of keywords. These topics and their labels offer users a structured and insightful view of the reviews, making it easier to discern the main discussion points and sentiments expressed by customers.

This process of topic modeling and labeling significantly enhances the interpretability of the results and aids users in deriving actionable insights from the wealth of user-generated reviews.

7. Guidelines for interpreting results

- **Product Name:** The name of the selected product category.

- **Total Number of Reviews Extracted:** Indicates the quantity of user reviews obtained from the website.

- **Assigned Topic Labels:** Labels assigned to each identified topic based on significant keywords. These labels provide insights into the main themes discussed in the reviews.

- **Identified Topics with Significant Keywords:** Each topic is associated with keywords that represent its main content. These keywords help users understand what each topic is about.

- **Topic Weights (Proportions):** Indicates the proportion of reviews that belong to each topic. Higher weights suggest greater importance.

- **Interpreting Topics:**Users can gain insights into customer sentiments, preferences, and common discussion points related to the product. For example, topics might include "Product Quality," "User Experience," "Taste," and more.

- **Exploring Multiple Products:** Users can explore reviews for different products by returning to the homepage and selecting a new product.

By following these guidelines, users can effectively interpret and utilize the insights generated by the application for decision-making and understanding customer sentiments towards various products.

8. Conclusion

The "Topic Modeling - Product Reviews" project represents a valuable tool for extracting meaningful insights from user-generated product reviews. By employing advanced natural language processing techniques, such as Latent Dirichlet Allocation (LDA), we have successfully transformed raw text data into structured and interpretable topics. These topics provide a clear overview of the main themes discussed in the reviews and enable users to navigate through the data efficiently.

Thank you

https://www.linkedin.com/in/raviteja-padala/ https://github.com/raviteja-padala