Topic Modeling - Product Reviews

High-Level Design Document

Table of Contents:

1. Introduction

1.1. Project Overview

The "Topic Modeling - Product Reviews" project is a web-based application designed to extract meaningful insights from user-generated product reviews. It leverages natural language processing (NLP) techniques, particularly Latent Dirichlet Allocation (LDA), to uncover latent topics within the reviews, facilitating a deeper understanding of customer sentiments and feedback.

1.2. Objective

The primary objective of this project is to provide users with an intuitive interface to select a product, extract user reviews from an e-commerce website, preprocess the text data, perform topic modeling using LDA, and present the results in an accessible format. This application aims to empower businesses, researchers, and consumers to make data-driven decisions based on product reviews.

1.3. Scope

The scope of this project includes:

- Web scraping to extract product reviews.

- Data preprocessing (lowercasing, punctuation removal, stopword removal, tokenization, lemmatization).

- LDA-based topic modeling.

- Automated topic labeling based on keywords.

- User-friendly web interface for interaction.

1.4. Audience

The target audience for this project includes:

- Business analysts seeking insights from customer feedback.

- Researchers analyzing trends and sentiments in product reviews.

- Consumers interested in understanding product pros and cons.

2. System Architecture

2.1. Web Interface

The web interface allows users to select a product, initiate the analysis, and view the results. It's built using the Flask web framework for Python.

2.2. Data Extraction

Web scraping modules are responsible for extracting user reviews and related information from e-commerce websites. Requests and BeautifulSoup libraries are used for this purpose.

2.3. Data Preprocessing

Text preprocessing functions clean and prepare the extracted text data. This includes lowercasing, punctuation removal, stopword removal, tokenization, and lemmatization.

2.4. Topic Modeling (LDA)

Latent Dirichlet Allocation (LDA) is employed to identify latent topics within the reviews. LDA models are trained on preprocessed text data to generate document-topic distributions, topic proportions, and significant keywords.

2.5. Results Presentation

The application presents results, including the product name, total reviews analyzed, assigned topic labels, identified topics, significant keywords, and topic weights. This information is displayed in a user-friendly format.

3. Key Components

3.1. Flask Application

The Flask application serves as the core of the project, handling user requests, data processing, and result presentation.

3.2. Web Scraping Module

Web scraping modules extract reviews and information from e-commerce websites using HTTP requests and parsing with BeautifulSoup.

3.3. Text Preprocessing Functions

These functions perform text cleaning, tokenization, lemmatization, and other preprocessing tasks on the text data.

3.4. LDA Model

The LDA model is trained on preprocessed text data to identify latent topics within the reviews.

3.5. Topic Labeling Module

Rules and keywords are used to assign meaningful labels to topics, enhancing result interpretability.

4. Data Flow

4.1. User Interaction

Users select a product through the web interface, triggering the analysis process.

4.2. Data Extraction

Web scraping modules extract user reviews and related data from e-commerce websites.

4.3. Data Preprocessing

Text preprocessing functions clean and prepare the text data.

4.4. Topic Modeling

The LDA model identifies latent topics, and topic labeling assigns meaningful labels based on keywords.

4.5. Results Presentation

The results, including topics, keywords, and topic weights, are presented to the user.

5. Error Handling

Error handling mechanisms are in place to provide informative error messages in case of issues during data extraction, analysis, or user interaction.

6. Deployment

6.1. Hosting Environment

The application can be deployed on a web server or hosting platform for online accessibility.

6.2. Scalability

The system is designed to accommodate additional products and adapt to evolving requirements.

7. User Guide

A comprehensive user guide is available to help users navigate the application, understand the results, and troubleshoot common issues.

8. Conclusion

The "Topic Modeling - Product Reviews" project offers a versatile solution for analyzing user-generated product reviews, with potential applications across various domains. It empowers users to gain valuable insights from unstructured text data.

This high-level design document outlines the architecture, components, and flow of the "Topic Modeling - Product Reviews" project, providing a roadmap for its development and understanding its functionality.