Big Data Final Project - Proposal

Group members:

Jackson Qu - hq20xx

Runze Li - rl50xx

Problem Statement:

In the e-commerce industry, user reviews contain a lot of valuable information about user needs, preferences, and product quality. However, with the rapid growth of review data, how to efficiently extract useful information from massive reviews, accurately identify user emotions, and provide users with personalized recommendations has become a major challenge. At the same time, the industry is highly competitive. It is of great significance for companies to capture market trends in a timely manner and reasonably optimize the recommendation system to improve user experience, strengthen customer relationships, and optimize marketing strategies. Therefore, it is crucial to deeply explore the value of user reviews through big data technology and sentiment analysis for the business improvement and competitiveness enhancement of e-commerce platforms.

Why is this project a big data project:

The main reason is that we are considering a large amount of e-commerce data over the years, looking at a great volume of data. Every minute, a large number of users post comments on different platforms, and the number of these comments has reached TB or even PB levels. This surge in data volume poses a severe challenge to traditional data processing methods. In order to effectively store, manage and analyze data of such a scale, the project must adopt a distributed computing framework such as Hadoop or Spark. These frameworks are able to store data in a dispersed manner on multiple nodes and improve the speed and efficiency of data analysis through parallel processing. In addition, to ensure the scalability of the system, when the amount of data continues to grow, the system can seamlessly expand to meet the increasing storage and computing needs.

Goal:

The project aims to develop an e-commerce user review analysis system based on big data, which can extract user insights from massive reviews through modules such as data cleaning, sentiment analysis, personalized recommendations, and trend prediction. The system will use distributed computing, natural language processing, and recommendation algorithms to have good scalability in data scale and achieve efficient processing and analysis. At the same time, through clear visualization of analysis results, it can provide key decision support for corporate management, help companies optimize product recommendations, improve user experience, and provide a basis for market trend prediction.

Data Description:

https://snap.stanford.edu/data/web-Amazon.html

The web-Amazon dataset from Stanford's SNAP (Stanford Network Analysis Project) consists of data collected from the Amazon e-commerce platform, including product reviews and social networks. This dataset is valuable for various analyses in big data contexts, such as sentiment analysis, recommendation systems, and network analysis.

Objectives: There are multiple tools of this project.

- Data Preprocessing:
    - pandas, numpy: Used to handle missing values and duplicates.
    - re: Text cleaning for removing special characters and formatting.
- Data Exploration and Visualization:
    - matplotlib: To calculate and visualize mean, median, and create rating distribution and product review count distribution.
    - plotly: For plotting the trend of review volume over time.
- Rating and Ranking System Based on Product Reviews:
    - Collaborative Filtering: Use Surprise library's SVD (Singular Value Decomposition) algorithm for collaborative filtering recommendations.
    - Content-based Ranking: Use Word2Vec to vectorize text from product reviews and rank products based on content similarity.

Expectation:

The successful implementation of this project is expected to significantly improve the analysis capabilities of e-commerce user reviews and the level of corporate decision-making. Through efficient data processing and analysis, the system will be able to process millions of user comments in real time, quickly extract key sentiments and trend insights, and provide enterprises with timely market feedback. In addition, the implementation of the personalized recommendation module is expected to significantly increase user satisfaction and purchase conversion rates, thereby increasing customer retention rates. Through clear visual display, management will obtain intuitive data analysis results, support more scientific decisions, and optimize product strategies. Ultimately, the project will provide companies with reliable trend predictions in a highly competitive market, helping them adjust market strategies in a timely manner and enhance their market competitiveness.