CS5344 - Amazon Reviews Big Data Analysis - Group 21

Our project endeavors to empower online sellers and brands operating in the dynamic realm of e-commerce, particularly within the Singaporean housing market. Leveraging cutting-edge big data techniques and analytics, our objective is to provide these stakeholders with enhanced insights into buyer sentiments. By simplifying the complexities inherent in the e-commerce landscape, we aim to furnish sellers with a comprehensive understanding, ultimately facilitating their success and resilience in navigating the intricacies of the Singaporean housing market

Table of Contents

Content

1. Install Dependencies

Ensure you have the required dependencies installed:

- Python 3.x
- Apache Spark 3.4.1 (Mac user, used brew to install apache spark) Reference: Mac Spark Installation guide
- Scala 2.12.17
- openjdk 20.0.1 2023-04-18
- OpenJDK Runtime Environment Homebrew (build 20.0.1)
- OpenJDK 64-Bit Server VM Homebrew (build 20.0.1, mixed mode, sharing)
- Spacy `pip install spacy`
- Spacy's `en_core_web_lg`
- NLTK `pip install nltk`
- NLTK's VADER `import nltk nltk.download('vader_lexicon')`
- Jupyter Notebook / Colab
- Tableau

2. Repository Structure

The repo conatins the following files:

- Dataset Folder
    - `Electronics.json` : This file contains the original Amazon reviews data.
    - `meta_Electronics.json` : This file contains the meta data
    - `Electronics_preprocessed` : This file contains preprocessed data after data cleaning.
    - `filtered_reviews` : This file contains reviews with minimum 5 upvotes.
    - `checkpoint folder` : contains check points (trained models from bert model training)
    - `Bert_Data folder` contains 3 subset data files from `Electronics_preprocessed` in the form train val test. It is used to train the bert model.
    - `extracted_aspects folder` : Contains the final extracted aspects. We have extracted top 5 aspects for each product and each aspect is associated with multiple opinions along with the sentiment of the opinion.
- src Folder - Contains all code files for different analysis performed on the data
    - Product KPI: contains code that is used to generate Product KPI's such as Total Average Rating and last year's rating.
    - Top 5 Products: contains code used to find Top 5 simiar products for each product clustered by Product Category.
    - dataprep: Data Preprocessing is carried out here
    - aspect_extraction: contains code to generate aspects
    - sentiment_analysis: contains code for Logistic Regression, Naive Bayes and Bert models for sentiment analysis

3. Code Execution

This code can be executed on either Jupyter Notebook or Google Colab or on local machine. Ensure to adjust the file paths according to your requirements before running the code.

- `DataPreprocessing.py` : `spark-submit DataPreprocessing.py`
- `AspectExtraction.py` : `spark-submit AspectExtraction.py`
- Notebook files can be run directly cell by cell