CS777 – Term Project Proposal

**Student: Prabal Mukherjee**

1. Data set description: Provide a detailed description of the public data set you have selected, including its source, format, and any relevant details about the data.

|  |
| --- |
| ***Dataset Description:*** *Amazon Review Dataset*  ***Source:***  *The Amazon Review Dataset is provided by*  *Jianmo Ni ,Phd(https://nijianmo.github.io/)*  *Julian McAuley, a professor at the University of California, San Diego (UCSD).*  ***Citation:***  *Justifying recommendations using distantly-labeled reviews and fined-grained aspects Jianmo Ni, Jiacheng Li, Julian McAuley, Empirical Methods in Natural Language Processing (EMNLP), 2019 [* [*http://cseweb.ucsd.edu/~jmcauley/pdfs/emnlp19a.pdf*](http://cseweb.ucsd.edu/~jmcauley/pdfs/emnlp19a.pdf) *]*  ***The dataset can be accessed through the following link:***  *[Amazon Review Dataset]([https://cseweb.ucsd.edu/~jmcauley/datasets/amazon\_v2/#sample-review](https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/" \l "sample-review) ).*  *Format:*  *The dataset is available in several JSON files, each corresponding to a specific category of products sold on Amazon. These categories include but are not limited to electronics, clothing, books, and home appliances. Each JSON file contains multiple entries, with each entry representing an individual review of a product.The data is organized in a structured format with key information such as reviewer ID, product ID, review text, rating, helpfulness votes, review timestamp, and product metadata.*  ***Sample Review:***  *{*  *"image": ["https://images-na.ssl-images-amazon.com/images/I/71eG75FTJJL.\_SY88.jpg"],*  *"overall": 5.0,*  *"vote": "2",*  *"verified": True,*  *"reviewTime": "01 1, 2018",*  *"reviewerID": "AUI6WTTT0QZYS",*  *"asin": "5120053084",*  *"style": {*  *"Size:": "Large",*  *"Color:": "Charcoal"*  *},*  *"reviewerName": "Abbey",*  *"reviewText": "I now have 4 of the 5 available colors of this shirt... ",*  *"summary": "Comfy, flattering, discreet--highly recommended!",*  *"unixReviewTime": 1514764800*  *}*  ***Data Details:***  *reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B*  *asin - ID of the product, e.g. 0000013714*  *reviewerName - name of the reviewer*  *vote - helpful votes of the review*  *style - a disctionary of the product metadata, e.g., "Format" is "Hardcover"*  *reviewText - text of the review*  *overall - rating of the product*  *summary - summary of the review*  *unixReviewTime - time of the review (unix time)*  *reviewTime - time of the review (raw)*  *image - images that users post after they have received the product*  ***Metadata***  *Metadata includes descriptions, price, sales-rank, brand info, and co-purchasing links:*  ***Sample Metadata:***  *{*  *"asin": "0000031852",*  *"title": "Girls Ballet Tutu Zebra Hot Pink",*  *"feature": ["Botiquecutie Trademark exclusive Brand",*  *"Hot Pink Layered Zebra Print Tutu",*  *"Fits girls up to a size 4T",*  *"Hand wash / Line Dry",*  *"Includes a Botiquecutie TM Exclusive hair flower bow"],*  *"description": "This tutu is great for dress up play for your little ballerina. Botiquecute Trade Mark exclusive brand. Hot Pink Zebra print tutu.",*  *"price": 3.17,*  *"imageURL": "http://ecx.images-amazon.com/images/I/51fAmVkTbyL.\_SY300\_.jpg",*  *"imageURLHighRes": "http://ecx.images-amazon.com/images/I/51fAmVkTbyL.jpg",*  *"also\_buy": ["B00JHONN1S", "B002BZX8Z6", ..."B00D103F8U", "B007R2RM8W"],*  *"also\_viewed": ["B002BZX8Z6", "B00JHONN1S", ... "B008F0SMUC", "B00BFXLZ8M"],*  *"salesRank": {"Toys & Games": 211836},*  *"brand": "Coxlures",*  *"categories": [["Sports & Outdoors", "Other Sports", "Dance"]]*  *}*  ***Relevance:***  *The Amazon Review Dataset is widely used for research purposes, particularly in the fields of natural language processing, sentiment analysis, recommender systems, and consumer behavior analysis. Researchers leverage this dataset to train machine learning models, extract insights into customer preferences and opinions, and develop algorithms for personalized recommendations and review summarization.* |

1. Research question: Clearly define your research question and explain why studying is important. What do you want to learn from the data?

|  |
| --- |
| ***Research Question:***  *How can natural language processing techniques be leveraged to develop an efficient search algorithm for product discovery on the Amazon Review Dataset? Additionally, how can the top reviews within a specific product category be utilized to create a recommender system, ultimately presenting users with a curated list of top 10 products for purchase?*  ***Key Components of the Research Question:***  ***1.******Natural Language Search Algorithm Development****: The research aims to explore the development of a search algorithm capable of interpreting natural language queries to identify relevant products within the Amazon Review Dataset. This algorithm will utilize techniques from natural language processing to extract meaningful information from user queries and match them with corresponding reviews in the dataset.*  ***2. Leveraging Top Reviews:*** *Once the search algorithm identifies products matching the user's query, the next step involves analyzing the top reviews within the selected product category. By considering the sentiments, ratings, and helpfulness of these reviews, the system aims to identify the most insightful and informative feedback from previous purchasers.*  ***3. Recommender System Creation:*** *Building upon the insights gathered from the top reviews, the research seeks to develop a recommender system. This system will utilize the top three reviews in the selected category to generate recommendations for the user. By employing techniques such as collaborative filtering or content-based filtering, the system will compile a list of the top 10 recommended products that align with the user's preferences and interests.*  ***Objective:***  *The primary objective of this research is to enhance the user experience in product discovery and decision-making on e-commerce platforms like Amazon. By integrating natural language processing and review analysis techniques, the proposed search algorithm and recommender system aim to provide users with personalized and relevant product recommendations based on their expressed preferences and the collective wisdom of previous consumers.* |

1. Machine Learning model: Specify the type of machine learning model you plan to use, such as classification or clustering, and explain why you have chosen this model.

|  |
| --- |
| ***1. TF-IDF for NLP Searching:***   * *TF-IDF (Term Frequency-Inverse Document Frequency) is a widely used technique in information retrieval and NLP for text analysis.* * *It quantifies the importance of a word in a document relative to a collection of documents.* * *TF-IDF is effective for searching similar text because it emphasizes words that are frequent in a document but rare across the entire corpus, thereby capturing the uniqueness of each document.* * *Given its proven effectiveness and simplicity, TF-IDF is a suitable choice for retrieving relevant products based on user queries in the context of NLP searching.*   ***2. ALS for Recommendation:***   * *ALS is a collaborative filtering technique commonly used in recommendation systems.* * *It leverages the patterns of users' preferences for items to make personalized recommendations.* * *ALS is well-suited for recommendation tasks because it can handle large-scale datasets efficiently and is robust to sparse and implicit feedback.* * *By modeling the latent factors of users and items, ALS captures the underlying structure of user-item interactions and generates recommendations based on similar user preferences.* * *Utilizing ALS for recommendation aligns with the objective of providing personalized and relevant product suggestions to users based on their preferences and historical interactions.*   ***3. Pyspark MLlib for Scalability:***   * *Leveraging Pyspark MLlib for both TF-IDF and ALS models is a pragmatic choice, particularly when dealing with large datasets.* * *Pyspark MLlib provides distributed implementations of machine learning algorithms, enabling scalability and efficient processing of large-scale data on distributed computing clusters.* * *Given the potentially vast amount of review data available in the Amazon Review Dataset, utilizing Pyspark MLlib ensures that the models can effectively handle the computational challenges posed by large-scale datasets.*   *In summary, the combination of TF-IDF for NLP searching and ALS for recommendation, implemented using Pyspark MLlib, offers a robust and scalable approach to address the research objectives of product discovery and recommendation on the Amazon Review Dataset.* |

1. Expected outcomes: What do you expect to achieve after implementing your learning model? What do you hope to learn or discover from your data analysis?

|  |
| --- |
| *After implementing the learning models, I expect to achieve several outcomes and gain valuable insights from the data analysis:*  ***1. Validation and Application of Learning Models:*** *By applying the TF-IDF model for NLP searching and the ALS model for recommendation on the Amazon review dataset, I aim to validate my understanding of these machine learning techniques in a real-world scenario. Through practical implementation, I anticipate solidifying my knowledge and gaining hands-on experience with these models.*  ***2. Enhanced Confidence in Machine Learning Skills:*** *Successfully implementing the learning models in a complex use case like the Amazon review dataset will bolster my confidence in my machine learning skills. This practical application will provide me with the assurance to tackle similar or more advanced projects in the future with confidence and competence.*  ***3. Insights into Product Discovery and Recommendation:*** *Through data analysis, I hope to gain insights into consumer behavior, product preferences, and review patterns within different product categories. Understanding these dynamics will enable me to better comprehend the factors influencing product discovery and recommendation on e-commerce platforms like Amazon.*  *Overall, I envision the implementation of these learning models as a transformative learning experience, equipping me with the skills, confidence, and insights necessary to excel in the field of machine learning and data analysis.* |

1. Evaluation plan: Explain how you plan to evaluate your project and assess the correctness of your model. What metrics or methods will you use to evaluate the effectiveness of your learning model? How well do you expect the model to work, and how will you measure its performance?

|  |
| --- |
| *Evaluating the performance of the ALS (Alternating Least Squares) model typically involves assessing its ability to accurately predict user preferences for items in the dataset. Here's a detailed explanation of how I can evaluate the ALS model and measure its performance:*  ***1. Train-Test Split:***   * *Divide dataset into training and testing sets. This is a standard practice in machine learning to evaluate model performance.* * *Typically, I would use a certain percentage (e.g., 70-80%) of the data for training and the remaining data for testing.*   ***2. Model Training:***   * *Train the ALS model on the training dataset using the Pyspark MLlib library.* * *The ALS model learns latent factors for users and items based on their interactions (in this case, user-item ratings/reviews).*   ***3. Model Evaluation:***   * *After training the model, I would use the testing dataset to evaluate its performance.* * *One common metric for evaluating recommendation systems is Root Mean Squared Error (RMSE).* * *RMSE measures the average difference between the actual ratings in the test set and the ratings predicted by the model.* * *Lower RMSE values indicate better model performance, as they suggest that the model's predictions are closer to the actual ratings.*   *By following these steps and evaluating the ALS model using appropriate metrics, I can assess its correctness and effectiveness in generating personalized recommendations for users based on their preferences and historical interactions with items in the dataset.* |