Milestone - 1

Group Background:

1. Swapnil Pardeshi (swpard) : I have 3 years of experience in web development. I am well versed with python and have basic machine learning knowledge

2. Sumanth Gopalkrishna (sgopalk) : I have good prior experience with python, primarily with numpy, pandas, and scikit-learn. I have 2.5 years of work experience. I worked on .net projects for web development and also some projects in Robotic process automation.

3. Vijay Iyer (vsiyer): I have 4 years work experience. I worked on a machine learning project at undergrad and an NLP project recently as part of the LAIDEL course.

4. Ajinkya Pawale (ajpawale): I have 1.5 years of work experience in data engineering and visualization. I have worked with python and machine learning through various online/offline courses. Also, I worked on a project based on CNN & RNN for emotion recognition, during my final year at undergrad.

Proposed Project

Project Topic Description:

We are planning to work on a recommendation system using the clustering approach in machine learning. The recommendation system will be based for an e-commerce platform, for a specific set of products. Many eCommerce websites struggle to sell a large portion of their inventory. This is frequently attributed to a bad browsing experience for the user. Customers can waste hours browsing through hundreds, if not thousands, of things before finding anything they want. In order to create a better shopping experience, shoppers must be given recommendations based on their likes and needs, thus the system will attract more customers and in turn improve the sales as well. We are also planning to integrate a chat bot which will recommend the products based on some search inputs provided by the user.

Motivation for the Project:

Commonly used in streaming apps such as Netflix to help us choose which TV shows or movies to watch next. Many e-retailers, such as Amazon, have used recommender algorithms for improving their sales and a better user experience, but many smaller or newer sites are still in need of them. Recommendation systems provide a flexible way of personalising content for the users with a large number of products.

Technical Description:

To retrieve or recommend the most relevant results for a given search query on an eCommerce website, based on -

1. The content/ item being searched as well as

2. The information about user-item interactions of other users (collaborative filtering).

Recommendation systems are generally divided into two types:

1. Collaborative filtering- The main idea is that you're given a matrix of preferences by users for items, and these are used to predict missing preferences and recommend items with high predictions
2. Content based filtering- These algorithms are given user preferences for items and recommend similar items based on a domain-specific notion of item content. This approach also extends naturally to cases where item metadata is available.

We will use collaborative filtering method with the matrix factorization to relate users with the items. Yet to choose which approach to go with. Either Singular Value Decomposition ,Alternating Least Squares or Stochastic Gradient Descent.

After this using the clustering algorithms like KNN to group users with similar tastes to build the recommendation engine.

In content based filtering we can first use tf-idf encoding along with matrix factorization to relate similar products and then use one of the clustering approaches.

Lastly, we might try hybrid models to include both of them.

Description of the data that you will use, and how you will access it.

We are planning to use Amazon review dataset. This dataset was released by Amazon in 2018 and it is available publicly.The dataset includes reviews (ratings, text, helpfulness votes) and product metadata (descriptions, category information, price, brand, and image features)
The data is divided into different subcategories such as electronics / beauty products / software etc. We are planning to experiment with different subcategories and choose the best fit.

The link for the Amazon Reviews Datasets is -
https://nijianmo.github.io/amazon/index.html
http://deepyeti.ucsd.edu/jianmo/amazon/index.html

Some other datasets we may consider are -

1. https://archive.ics.uci.edu/ml/datasets/online+retail (This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail)
2. https://www.kaggle.com/retailrocket/ecommerce-dataset The dataset consists of three files: a file with behaviour data (events.csv), a file with item properties (item*properties.csv) and a file, which describes category tree (category*tree.csv). The data has been collected from a real-world ecommerce website. It is raw data, i.e. without any content transformations, however, all values are hashed due to confidential issues. The purpose of publishing is to motivate researches in the field of recommender systems with implicit feedback.

References:

- https://heartbeat.fritz.ai/recommender-systems-with-python-part-i-content-based-filtering-5df4940bd831
- https://medium.com/@ashutoshsingh93/recommendation-system-for-e-commerce-using-collaborative-filtering-fa04d6ab1fd8
- Recommendation System Series Part 1: An Executive Guide to Building Recommendation System
- http://cseweb.ucsd.edu/~jmcauley/pdfs/emnlp19a.pdf
- https://jessesw.com/Rec-System/

Potential challenges/hurdles:

1. One of the drawbacks of Collaborative filtering is that we encounter the cold start problem, where we do not have enough data for the new users to recommend items for them, we plan on using a content based recommender for this, which finds similarities with content based features. (using the metadata of the users and the items.)
2. Which performance metric to use for the evaluation of the unsupervised model.
3. Since we will be using unsupervised learning approaches, it might become difficult to improve the accuracy / performance of the model with the same data. For supervised learning methods, we might run a few more training epochs , change the validation sets to fine tune the model.