Project Proposal

Title: Movie/ Web series recommendation system

Group Size: 5 (five).
Timeline: 5 Weeks (This would require the candidates to work on an average of 10-12 hours per week)

Planned Start Date: TBD

Planned End Date: TBD

Expected Member Profile:

- **Math:** Basic math including algebra, statistics and probability (permutation and combinations).
- **Programming**: Some exposure to programming (not necessarily python).
- Can commit the required time (10-12 hours per week) and don't miss any sessions/sections.
- Data science and Machine learning: Knowledge of clustering, tree- based methods.
- Worked on data: From csv/excel, plotting and summarizing using excel or other tools

Recommended Preparation and Study Material (from Packt Library):
Programming:

- The Python Workshop
  Sections 9, 10, 11
- The Data Science Workshop
  Sections 3, 4, 5, 6, 7, 8, 12, 13
- The Data Visualization Workshop
  Sections 2, 3, 4
- The Machine Learning Workshop
  Sections 2, 3, 4

Technology Stack:

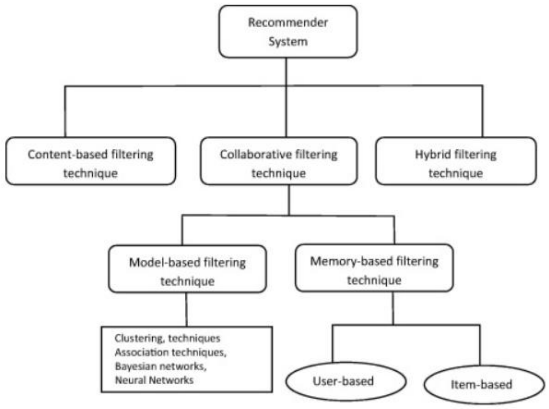
- Python 3x and Anaconda distribution (https://www.anaconda.com/products/individual)
- Jupyter IDE notebook (comes with Anaconda distribution)
- Plotting/visualization libraries: Matplotlib and Seaborn (again comes with Anaconda distribution)
- Machine learning libraries: Sklearn (again comes with Anaconda distribution)
- Flask for creating an API
- Spreadsheet: Excel or openoffice/libreoffice or Google Sheets

Topics:
Python Data Structures, Pandas, Numpy, Unsupervised Learning, Self-Supervised Learning, Clustering Models, Term Frequency-Inverse Document Frequency, Cosine-Similarity, Content filtering and

Collaborative filtering, Flask.

Datasets: https://developers.google.com/youtube/v3

Project Architecture:



GitHub Repo Link:
https://github.com/TeamEpicProjects/Recommendation-System

Recommended System Setup:

Hardware Requirements
Laptop/Desktop with at least 8GB RAM.

Software Requirements
Operating system: Linux, Mac or Windows. Python 3.5+, Anaconda version (supports Python3.x) – available for Mac, Windows and Linux (Debian/Ubuntu). Spreadsheet(excel/libreoffice).

https://docs.anaconda.com/anaconda/user-guide/faq/

Problem Statement
YouTube has undoubtedly emerged as a top platform for videos, vlogs, and so on. They have now planned to launch its own OTT platform “YouTube Watch” wherein they will be featuring movies, web series, documentaries and so on specifically. Netflix, Amazon prime video being its top competitors, YT now wants to build a movie/ series recommendation system based on the previous watch history, preferred genre, and other behavioral insights of the user.

Commented [ST1]: Can we add the data sources here?

Commented [je2R1]: We're using the google api for th data, you can add the link for YT Api here.

Commented [ST3]: Please add a simple flowchart or architecture diagram.

YT is part of Google and YT itself is huge in India and they have humungous amount of data of Indian audience. YT after 2016 particularly grew exponentially when internet got cheaper and became more mainstream in Tier 1 and Tier 2 cities.

Now, not only are the people using YT to entertain themselves, but they are also using it to create content and entertain others such as from Cooking Channels of Dadaji's made by their grandchildren to farmers.
YT is already filled with tones and tones of Indian content; your task is to use that data and create a recommendations model using Content-Based Filtering and collaborative filter.

Level 1: Analyze information and build different brackets: Analyze the search history, previous watched content, preferred genre, average time spent on YT videos of the user and build an analysis to show suitable recommendations under different brackets such as “Popular on YT Watch”, “You may also like”, “Trending in India”, “New releases”, “Critically Acclaimed”, “Since you watched Friends, you may like.”

Level 2: Analyze information of your connections: YT Watch should also be able to analyze the above data of the connections of the user and suggest movies/ series under “What your connections are watching”

Level 3: Send Notifications for New Releases: YT Watch should also be able to send app notifications to the user, with recommendations of newly released content.

Commented [ST4]: This is a part of the problem statement.

Expected Solution

Solution will focus on developing a decision system that is a combination of a machine learning based system, which specifically requires Collaborative and Content Filtering or a clustering model requiring understanding of viewers as well as the content developers and their behavior, to fine tune the underwriting algorithm to differentiate and find similarities between viewers. The solution should address the users' concern of getting recommend more content which they prefer consuming.

User Stories:

Commented [ST5]: The user stories need to be role specific. Please refer the sample project proposal shared.

- You are the head of YT in India, and you can see the growth of YT and Internet content India, Google wants to set up an Indian YT which hosts a lot of Indian Films, TV -Show and other content such as Stand-Up Comedy.
- As a Brand Manager, you are aware that there are already enough Indian specific content and now to increase the brand value of YT more, you want to create a branch of YT which is Indian content specific and can be used as Brand itself.
- As a Marketing Lead, I would try to understand the customer's behavior and try to understand what kind of content is used mostly and promote more of that content to get more eyes on the product.

- As a manager, you have a task to undertake those ideas and try to create a model where you can understand the consumer behavior.
- As a Data Scientist, you must create a model which is able to use the Indian content and create a recommendation system which uses that Data and gets more eyes on the brand and is able to get more new audience involved.

Project Timeline

Work Package

Well documented analysis with basic exploration, data mining for pattern recognition and actionable insight for the business.

The first task to be done is use Google YouTube API v3 to gather data,

https://developers.google.com/youtube/v3

1. Apply for the access keys.
2. Go through the procedure to gather data and access it using Python.

The objective is to find a solution to a business problem using data by collaborating with other team members where an individual's contribution to the overall solution is imperative and is evaluated throughout the course.

Commented [ST6]: How long will the access keys be active? Is this free for use?

Commented [je7R6]: They're permenant

Milestone 1: Understanding the dataset and preliminary observations
User Story - 1, 2 and 3
Week 1
Induction and project overview:

- Team induction and explanation of the project by the team mentor
- Detailed plan for the team to work in the week, including key objectives and various steps.

Individual task:

- Setup the software environment required, access the data set and combine/segment different sets appropriately.
- Explore and review the features available in the data.
- Summarize the initial data.
- Collaborate with others for a summary.

Group task:

- Group to discuss and consolidate the data summary - features, classification of features etc.
- Organize all the findings and summarize them in the Jupyter notebook.

Milestone 2: Data cleansing and Exploratory Data Analysis (EDA)
User story: 1
Week 2
Individual task:

- Missing values and imputation in each of the datasets
- Identify outliers in the data
- Imbalance analysis, categorical and numerical value analysis. And Temporal and correlation analysis.
- Multivariate analysis: Multiple cuts of data with more than one variable and its correlation with other variables and with the dependent variable.

Group task:

- Consolidate the data cleaning summary - missing data, imputation, univariate and multivariate analysis etc. Organize all the findings and summarize in the Jupyter notebook.

Milestone 3: Merged data analysis and Feature engineering
User story: 2 and 3
Week 3
Individual task:

- Merging different data sets into a single data-frame and analyzing the same
- Come with engineered features - at least 3x the available raw variables/features
- Check for the dependent variable correlation for the engineered features.
- Bucket features appropriately for modeling step

Group task:

- Encoding work for model can be completed by splitting among the group for various features
- Collaborate to consolidate all the analysis - the single consolidated notebook to have the same definition and feature

Milestone 4: Train and validate the ML model and identify key drivers
User story: 2 and 3
Week 4
Individual task:

- Develop model architect - entity, dataset, test/train/validation, and evaluation
- Define evaluation metrics and train the model
- Scoring: Clustering and Filtering
- Refining the score method and model tuning.

Group task:

- Discuss among the group to train different models and variations of the model.
- Summarize various models in a single notebook and key features.

Milestone 5: Tune model, summarize and present
User story: 1, 2 and 3

Week 5
Individual task:

- Tune the model for improved accuracy.
- Identify the content similarity using evaluation method.
- To highlight that create a Flask API to highlight the recommendation system's working.
- Suggestion based on the model summary to the business owners and other stakeholders.

Group task:

- Tune the model after discussion with others for different settings which can be split among others in the group.
- Summarize impact to the business with the model.
- Create a presentation of the implementation and the working of the model to highlight that to the stakeholders.
- Present the findings to the panel with business application and solution to the users
- List the caveats and other shortcomings in the note.

Project Deliverables:

- Jupyter notebooks and other relevant code files
- Simple UI to demonstrate the recommendation
- Project Report
- Presentation deck for final project demonstration

Notes:

- Focus on understanding the problem building. Understand the factors/ variable of the dataset.
- EDA is a way to understand and explain the data that you are working on to the stakeholders as

Commented [ST8]: I was looking to use Azure DevOps for project management and then use Azure Pipelines for the building , testing and deployment. Do you think we can include that scope here?

Commented [je9R8]: That's a good idea but I would suggest to extend the project duration for that.

Commented [ST10]: Does the Google API come into the picture at this stage?

Commented [je11R10]: It does

Commented [ST12]: Group Task for final week will also include work done for project demonstration.

Commented [ST13]: I have added this section. Please review and feel free to make changes if any.

well.

- Visualization is the key here. Creating a model is the final task and that is far easier when you understand the data.
- To create a model, use clustering method first by using association or k- means clustering.
- After clustering you are done with creating labels and after that you can apply a supervised classification/ clustering method that is what we call Self Supervised Machine Learning.

Assessment criteria for the solution

- Data Summary and Exploratory Data Analysis (EDA):
    - Organization of data summary
    - Quality and reusability of the code
    - Presentation of the exploratory data analysis and highlights of the findings.
    - Finding insights on the content and ability to label them
    - Running of statistical tests to understand data
- Creativity and its quality:
    - Creative usage of the API and finding unsolicited details of the tags used
    - Displaying interest in content and finding more insights into a particular type/ class of content.
    - Usage of self-supervised machine learning
- Model performance:
    - Overall efficacy of the analysis and model – How well the model segments the customers.
    - Usage of different matrices to find the efficiency and accuracy of the model.
    - And, apart from the mathematical matrices for efficiency, a trail by the Data Scientist himself if the model can recommend the content seen by the user
- Business solution:
    - Actionable insights for the content.
    - User problems addressed in the solution.
    - Bottom line impact with the model
- Future Improvements:
    - A list of improvements that can be made in the future, whether that is model based, or business based