***Introduction***

This paper study the general idea behind a movie recommendation system. Even though we will focus on movie recommendation system similar system could be build for any type of recommendation systems. Such engines are similar to the ones used by Netflix, Amazon, Hulu etc. The main point is to provide an engine that recommends to its users content that is relatable to them, and that they might be interested in. The technique used is content-based filter which refers to the content or attributes of the products you like. Content-based filtering recommends movies based on meta data such as gender, genre, producer etc. The basic idea is that if you like an item, then you will also like a “similar” item.

***Methods and tools:***

1. NLP

Natural Language Processing is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language.

The ultimate objective of NLP is to read, decipher, understand and make sense of the human languages in a manner that is valuable.Most NLP techniques rely on machine learning to derive meaning from human languages. Mostly used on the web & social media monitoring, Natural Language Processing is a great tool to comprehend and analyse the responses to the business messages published on social media platforms. It helps to analyse the attitude and emotional state of the writer (person commenting/engaging with posts). This application is also known as opinion mining. It is implemented through a combination of Natural Language Processing and statistics by assigning values to the text (positive, negative or neutral) and in turn making efforts to identify the underlying mood of the context (happy, sad, angry, annoyed, etc.). In the project NLP to process the dataset text and find similarities and achieve ranking of movies based on reviews

1. *TF-IDF to score movies based on content*

The *TF-IDF* algorithmis used to weigh a keyword in any document and assign the importance to that keyword based on the number of times it appears in the document. Put simply, the higher the TF\*IDF score (weight), the rarer and more important the term, and vice versa. Each word or term has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF\*IDF weight of that term.

The **TF (term frequency)** of a word is the number of times it appears in a document. You will be able to see the term too often or too infrequently .

The **IDF (inverse document frequency)** of a word is the measure of how significant that term is in the whole corpus.

1. *COSINE SIMILARITY to determine the similarity between movies*

Here we’ve calculated the cosine similarity of each item with every other item in the dataset, and then arranged them according to their similarity with item.

1. NAIVE BAYES Analyzer to score the reviews

Naive Bayes is a useful technique to apply in text classification problems.  Some applications are spam filtering, customer support ticket routing, authorship identification or topic prediction. Another case is sentiment analysis, whose objective is to determine the writer’s point of view about a particular topic, product, service, etc.

The Naive Bayes in the project return the positivity of the reviews and got the classifier of our reviews such as positive and negative. In the project we use the positivity of each reviews from 0 to 1.

1. NLTK to get the stem of word and get rid of from English stopwords.

The porter-stemming algorithm is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems.

1. VARIOUS DATASETS:

Movies dataset from Kaggle, and an automatically generated dataset for reviews

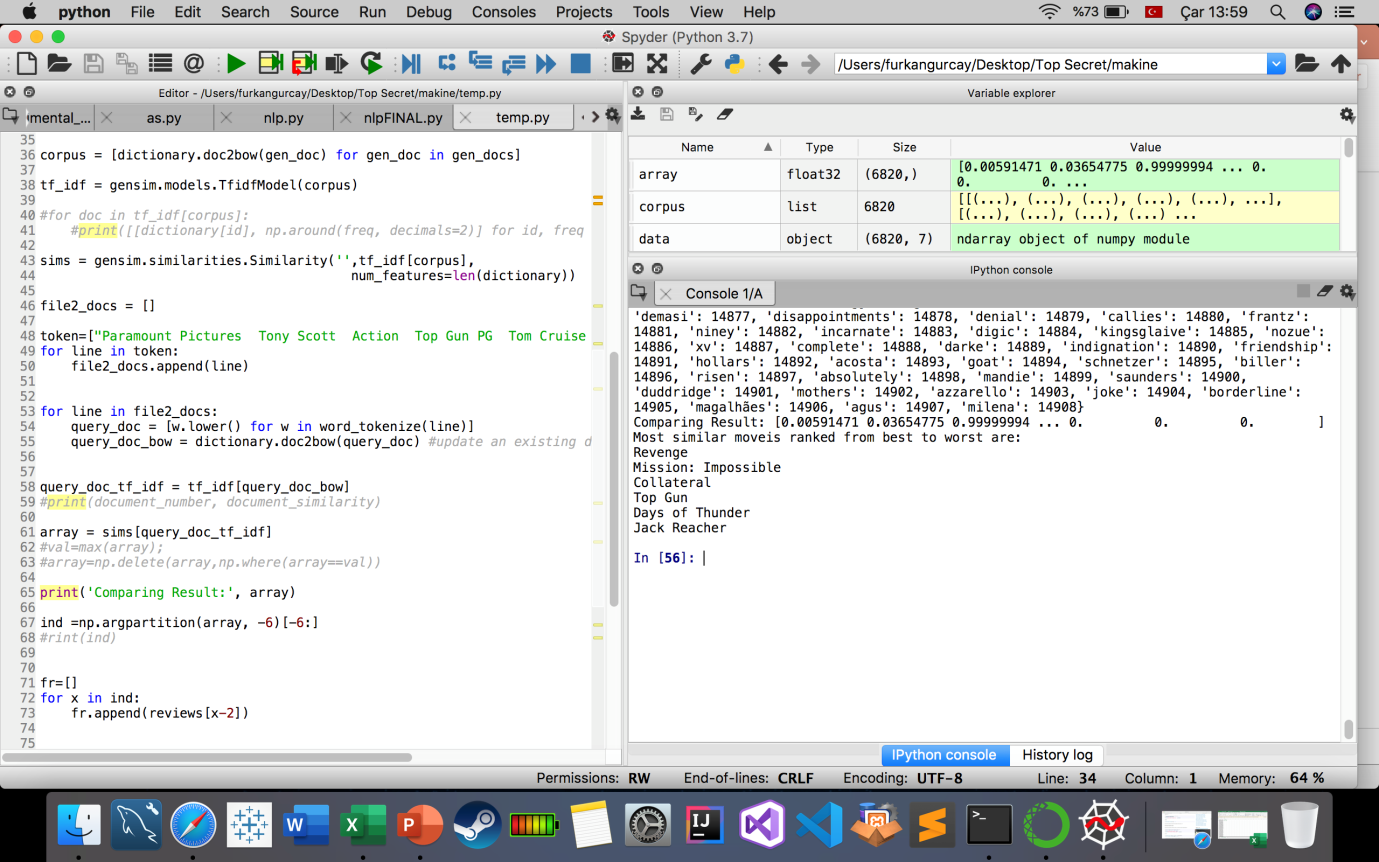
**Advantages of Content-Based Filtering**

* **User independence:** The content-based method only has to analyze the items and a single user’s profile for the recommendation, which makes the process less cumbersome. Content-based filtering would thus produce more reliable results with fewer users in the system.
* **Transparency:**Collaborative filtering gives recommendations based on other unknown users who have the same taste as a given user, but with content-based filtering, items are recommended on a feature-level basis.
* **No cold start:** As opposed to collaborative filtering, new items can be suggested before being rated by a substantial number of users.

**Disadvantages of Content-Based Filtering**

* **Limited content analysis**: If the content doesn’t contain enough information to discriminate the items precisely, the recommendation itself risks being imprecise.
* **Over-specialization:** Content-based filtering provides a limited degree of novelty since it has to match up the features of a user’s profile with available items. In the case of item-based filtering, only item profiles are created and users are suggested items similar to what they rate or search for, instead of their past history. A perfect content-based filtering system may suggest nothing unexpected or surprising.

***Code and result***

****

Worked: Kevin Albajrami 160201108, Suzana Marsela 160201100, Furkan Gurcay 150201018