CA06 – kNN based Recommender Engine

1. The Application

At scale, this would look like recommending products on Amazon, articles on Medium, movies on Netflix, or videos on YouTube. Although, we can be certain they all use more efficient means of making recommendations due to the enormous volume of data they process. However, we could replicate one of these recommender systems on a smaller scale using what we have learned here in this article. Let us build the core of a movies recommender system.

What question are we trying to answer?

Given a movies data set, what are the 5 most similar movies to a movie query?

2. Data Source and Contents

If we worked at Netflix, Hulu, or IMDb, we could grab the data from their data warehouse. Since we don't work at any of those companies, we have to get our data through some other means. We could use some movies data from the UCI Machine Learning Repository, IMDb's data set, or painstakingly create our own. In our case, we will use a small sub-set of the data extracted from the UCI's IMDB data set.

Data File Name: movies_recommendation_data.csv

File Location: https://github.com/ArinB/CA05-kNN/raw/master/movies_recommendation_data.csv

NOTE: Use the above EXACT URL in your code as data file location

The data contains thirty movies, including data for each movie across seven genres and their IMDB ratings. The labels column values are all zeroes because we aren't using this data set for classification or regression. You can ignore this column. The implementation assumes that all columns contain numerical data.

Additionally, there are relationships among the movies that will not be accounted for (e.g. actors, directors, and themes) when using the KNN algorithm simply because the data that captures those relationships are missing from the data set. Consequently, when we run the KNN algorithm on our data, similarity will be based solely on the included genres and the IMDB ratings of the movies.

A Snapshot of part of the Data set:

Search this file...

Movie ID	Movie Name	IMDB Rating	Biography	Drama	Thriller	Comedy	Crime	Mystery
58	The Imitation Game	8	1	1	1	0	0	0
8	Ex Machina	7.7	0	1	0	0	0	1
46	A Beautiful Mind	8.2	1	1	0	0	0	0
62	Good Will Hunting	8.3	0	1	0	0	0	0
97	Forrest Gump	8.8	0	1	0	0	0	0
98	21	6.8	0	1	0	0	1	0
31	Gifted	7.6	0	1	0	0	0	0
3	Travelling Salesman	5.9	0	1	0	0	0	1
51	Avatar	7.9	0	0	0	0	0	0
47	The Karate Kid	7.2	0	1	0	0	0	0
50	A Brilliant Young Mind	7.2	0	1	0	0	0	0
49	A Time To Kill	7.4	0	1	1	0	1	0
30	Interstellar	8.6	0	1	0	0	0	0
94	The Wolf of Wall Street	8.2	1	0	0	1	1	0
6	Black Panther	7.4	0	0	0	0	0	0
73	Inception	8.8	0	0	0	0	0	0
44	The Wind Rises	7.8	1	1	0	0	0	0
65	Spirited Away	8.6	0	0	0	0	0	0
48	Finding Forrester	7.3	0	1	0	0	0	0
27	The Fountain	7.3	0	0	0	0	0	0
57	The DaVinci Code	6.6	0	0	1	0	0	1
57	Stand and Deliver	7.3	0	1	0	0	0	0
14	The Terminator	8	0	0	0	0	0	0
69	21 Jump Street	7.2	0	0	0	1	1	0
98	The Avengers	8.1	0	0	0	0	0	0
17	Thor: Ragnarok	7.9	0	0	0	1	0	0
12	Spirit: Stallion of the Cimarron	7.1	0	0	0	0	0	0
1	Hacksaw Ridge	8.2	1	1	0	0	0	0
86	12 Years a Slave	8.1	1	1	0	0	0	0
46	Queen of Katwe	7.4	1	1	0	0	0	0

movies_recommendation_data.csv hosted with ❤ by GitHub view raw

3. Building your own Recommender System

You are building your own movie recommendation website which uses your Recommendation Engine at the back-end. You are going to build this back-end Recommendation Engine. Imagine a user is navigating your recommendation website, and he/she encounters a movie named “The Post”. The user is not sure if he/she wants to watch it, but its genres intrigue the user; he/she is curious about other similar movies. The user scrolls down to the “More Like This” section to see what recommendations your recommendation website will make, and the back-end algorithmic gears begin to turn.

Your website sends a request to its back-end for the 5 movies that are most similar to *The Post*. The back-end has a recommendation data set exactly like ours. It begins by creating the row representation (better known as a **feature vector**) for *The Post*, then it runs a program similar to the one below to search for the 5 movies that are most similar to *The Post*, and finally sends the results back to the user at your website.

Following is the genre information about the movie "The Post"

IMDB Rating = 7.2, Biography = Yes, Drama = Yes, Thriller = No, Comedy = No, Crime = No, Mystery = No, History = Yes

What recommendations he/she will see?

Implement this problem using Python scikit-learn and display the answer within the Notebook with proper narrative / comments.