



Continue Watching for clutter



Trending Now



Content-Based Recommender using
Natural Language Processing

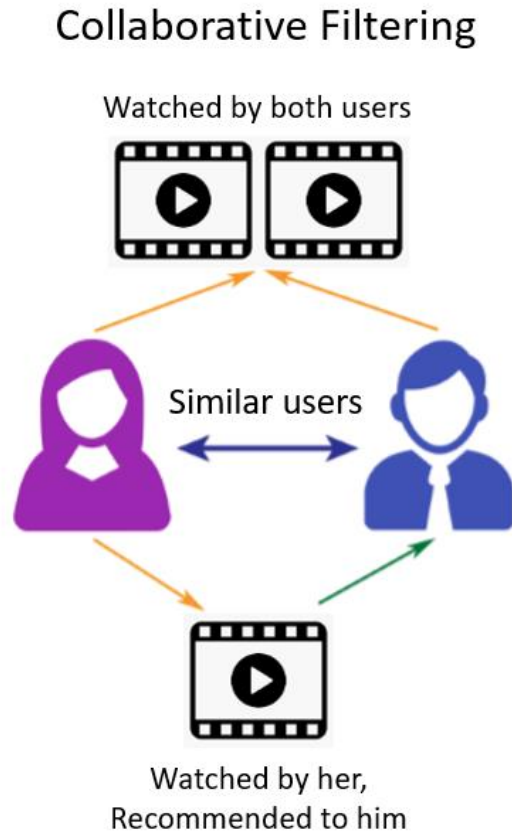
What are Recommender Systems?

A recommender system is a type of information filtering system that attempts to forecast a user's "rating" or "preference" for an item. Its purpose is to make relevant suggestions to users for things or products that they might be interested in. Real-world examples of the operation of industry-strength recommender systems include book recommendations on Amazon and movie recommendations on Netflix, etc. Recommendation engines are a subclass of machine learning which generally deal with ranking or rating products / users.

There are 2 types of Recommender Systems:

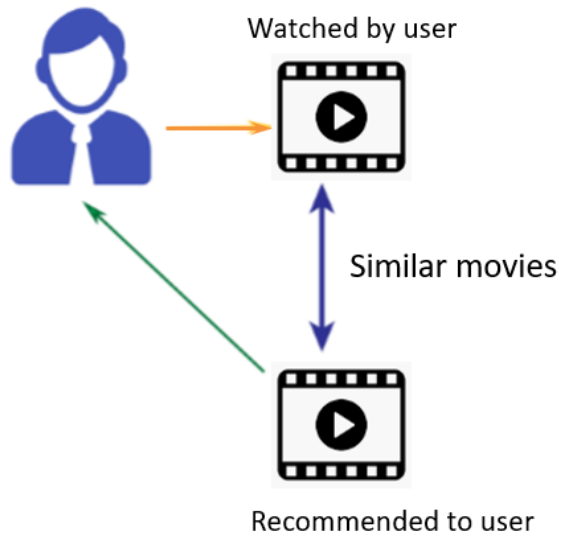
- ▣ Collaborative Recommender System
- ▣ Content-Based Recommender System

Collaborative Recommender Systems



Users are grouped together based on their ratings and purchases, and then products/services are recommended to them.

Content-Based Filtering



Content-Based Recommender Systems

Content-based recommenders utilize the data that the user has provided. The data is used to create a user profile, which is then used to make suggestions to the user. As the user provides additional inputs or actions in response to recommendations, more and more accurate results are generated by the engine. This is also referred to as metadata.

Motivations & Challenges


According to IMDb, the Indian film industry releases between 1500 and 2,000 films per year in 20 languages. This sparked my interest in creating a database on Indian movies.

Indian film industries are numerous, but obtaining the right dataset posed a challenge. Searching through many sources, I found a dataset that had all the components, but not enough data points. Therefore, I looked for other sources where I could find a dataset with more data points. But I was unsuccessful. So, I have decided to combine the two datasets into a single dataset.



Chosen Dataset

As mentioned, I have used two sources for this project. One dataset which I have used is from IMDB_TopIndianmovies downloaded from data.world and other from Indian Movies Dataset from kaggle which is scraped from IMDB.



RAKE: Rapid Automatic Keyword Extraction

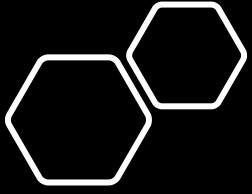
The extraction of keywords is the most fundamental and initial step in the NLP (Natural Language Processing). NLP offers a wide range of algorithms for extracting keywords from text, but for this project I have chosen "RAKE".

"RAKE is a domain independent keyword extraction algorithm which tries to determine key phrases in a body of text by analyzing the frequency of word appearance and its co-occurrence with other words in the text."



RAKE

- This technique can extract a wide range of keywords from individual documents, which can then be applied to a dynamic collection. Also, it is very successful in managing a variety of documents, especially those that follow certain grammatical rules.
- Using stop words and delimiters, it partitions the page into keywords; these keywords are usually the terms that help developers obtain the specific keywords from the document that they need.



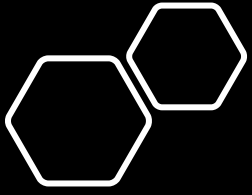
Data Pre-Processing

To begin, data must be pre-processed using natural language processing (NLP) to produce only one column containing all of the movie's attributes (in words). After that, vectorization is used to transform the data into numbers, and each word is given a score. The cosine similarity may then be calculated.

Similarity Matrix

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} = \frac{\sum_{i=1}^n u_i v_i}{\sqrt{\sum_{i=1}^n u_i^2} \sqrt{\sum_{i=1}^n v_i^2}}$$

$$\mathbf{u} \cdot \mathbf{v} = [u_1 \ u_2 \ \dots \ u_n] \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = u_1 v_1 + u_2 v_2 + \dots + u_n v_n = \sum_{i=1}^n u_i v_i$$



Bag_of_Words

Four columns, 'Genre, Title, Language, and KeyWords' are merged into a new column called 'Bag_of_words' after data pre-processing. Now, there are just two columns in the final DataFrame.

```
      Title      Bag_of_words
0      [anand]  D r a m a anand H i n d i anand
1  [drishyam]  C r i m e , D r a m a , T h r i l l e r drishy...
2  [nayakan]  C r i m e , D r a m a nayakan T a m i l H i n ...
3  [anbesivam]  A d v e n t u r e , C o m e d y , D r a m a an...
4  [dangal]    A c t i o n , B i o g r a p h y , D r a m a da...
...      ...      ...
2245  [suttakadhai]  C o m e d y , C r i m e      suttakadhai t a m ...
2246  [kireedamillatharajakkanmar]  C o m e d y      kireedamillatharajakkanmar m a...
2247  [aadinagalu]  C r i m e , D r a m a , R o m a n c e      aadi...
2248  [sitamgar]    A c t i o n      sitamgar h i n d i sitamgar
2249  [english:anautumninlondon]  D r a m a      english:anautumninlondon m a l a...

[2250 rows x 2 columns]

D r a m a anand H i n d i anand
```

Reflective Conclusion

While working on this project, I have used multiple sources to make a whole dataset, which resulted in some incomplete attributes. This mistake gave me few hiccups while executing the code.

Legal, Ethical and Social aspects

We can use recommender systems to take controlled sips from the information stream pointed our way every day, every week. Specifically, by highlighting a few items from a vast catalog that are especially relevant or valuable. While these items are incredibly valuable, they also have some serious ethical failures, which are often caused by the tendency for companies to build recommenders to reflect user feedback without considering the broader implications these systems have for society and human civilization.

Reference

[1] J. Ng, "Content-Based Recommender Using Natural Language Processing (NLP)," 16 Nov 2019. [Online]. Available: <https://jnyh.medium.com/content-based-recommender-using-natural-language-processing-nlp-159d0925a649>

[2] <https://pypi.org/project/rake-nltk/>

[3] K. Agarwal, "RAKE: Rapid Automatic Keyword Extraction Algorithm," 8 April 2020. [Online]. Available: <https://medium.datadriveninvestor.com/rake-rapid-automatic-keyword-extraction-algorithm-f4ec17b2886c>.

[4] N. BHAT, "50,000+ Indian Movies Data set," 2021. [Online]. Available: <https://www.kaggle.com/datasets/nareshbhat/indian-moviesimdb>.

[5] J. Harris, 27 January 2021. [Online]. Available: <https://towardsdatascience.com/ethical-problems-with-recommender-systems-398198b5a4d2>.

[6] Shuvayan, "Beginners Guide to learn about Content Based Recommender Engines," 11 August 2015. [Online]. Available: <https://www.analyticsvidhya.com/blog/2015/08/beginners-guide-learn-content-based-recommender-systems/>

Thank You

