IT492: Recommendation Systems

Lab Assignment - 02

[Deadline: 06th March 2022, Sunday 10 PM]

This assignment involves implementing content-based and hybrid recommendation techniques and comparison of their performance. You can use any text processing library of your choice. Some popular choices are Spacy, NLTK, Stanford CoreNLP. Transformer-based model embeddings can be used from Huggingface.

Marking scheme and requirements

Full marks will be given for (1) working, readable, reasonably efficient, documented code that achieves the assignment goals and (2) for providing appropriate answers to the questions in your Google colab file (`name format: LA02_rollnumber`) submitted via Google Classroom on the **assigned dataset only**.

Please refer to the Dataset Allocation

Sr No	Student Id	Student Name	Student email	Registration Type	For LA-01, 02
1	202118004	ABHISHEK SINGH	202118004@daiict.ac.in	AUDIT	last.FM (Hetrec 2011)
2	202018004	P SARAN PANDIAN	202018004@daiict.ac.in	AUDIT	
3	202018026	AAKANKSHA SHAH	202018026@daiict.ac.in	AUDIT	
4	202111002	SHARMA HARSH DHARMENDRAKUMAR	202111002@daiict.ac.in	AUDIT	
5	202111029	GORASIYA RAGHAV NARESH	202111029@daiict.ac.in	AUDIT	
6	202018042	ABHIJEET KUMAR	202018042@daiict.ac.in	REGULARADD	Movielens 20M (Grouplens 2016)
7	202111010	KEVIN JITENDRABHAI JADIYA	202111010@daiict.ac.in	REGULARADD	
8	202111035	VANSH RAHUL BHANJIBHAI	202111035@daiict.ac.in	REGULARADD	
9	202111048	MANSURI PINJARA MOHAMMED JUNED HANIFBHAI	202111048@daiict.ac.in	REGULARADD	
10	202112030	ARPITHA SREENIVASAN	202112030@daiict.ac.in	REGULARADD	
11	201801466	PARMAR SIDDHRAJ YOGESHBHAI	201801466@daiict.ac.in	REGULARADD	
12	202121004	SANDHYA KUMARI	202121004@daiict.ac.in	REGULARADD	
13	202116003	AMBUJ MISHRA	202116003@daiict.ac.in	REGULAR	Food Reviews (Kaggle 2019)
14	202116004	ARPITA NEMA	202116004@daiict.ac.in	REGULAR	
15	202116008	RAHUL KUMAR	202116008@daiict.ac.in	REGULAR	
16	202116009	RAHUL THAKUR	202116009@daiict.ac.in	REGULAR	
17	202116011	ROHAN BAGHEL	202116011@daiict.ac.in	REGULAR	
18	202116001	ABHISHEK YADAV	202116001@daiict.ac.in	REGULAR	
19	202116002	AKSHAY KAUSHIK	202116002@daiict.ac.in	REGULAR	

Links to download the Datasets

- *last.FM (Hetrec 2011)*
- *Movielens 20M (Grouplens 2016)*
- *Food Reviews (Kaggle 2019)*

IT492: Recommendation Systems

Please adhere to the lab policy on the course website

- Cite resources and give credit where it's due. If you happen to discuss the questions with your peers, please mention your collaborators in your report/assignments.
- Acts of plagiarism will not be tolerated and will result in a straight ZERO for that assignment.
- Students who don't submit their assignment by 08th March 2022, Tuesday 10 PM will simply get ZERO.

Main Assignment (10 Marks in Total)

Dataset Analysis (2 Mark)

1. Explore the dataset and summarize the item descriptor fields in the dataset assigned to you. Discuss which fields can be used as features for a content-based recommendation system. Explore PCA or sklearn's feature selection for dimensionality reduction of the item-feature matrix.

Content-based (Memory-based) Recommendation (4 Marks)

2. Construct the item feature matrix using (i) 1-hot keywords, (ii) tf-idf weighting scheme, and (iii) neural embedding such as word2Vec or BERT. Use the features identified in the previous question. You can use sklearn's TF-IDF vectorizer. Using each of the above three item feature matrix, compute top-N recommendations for each user and show the overall relevance score (i.e. precision) for N = {5, 10, 15, 20, 25}.

Content-based (Model-based) Recommendation (4 Marks)

3. Use the most common content-based models: (i) Naive Bayes, (ii) Logistic Regression, and (iii) k-NN to compute top-N recommendations for each user and show the overall relevance score (i.e. precision) for N = {5, 10, 15, 20, 25}.