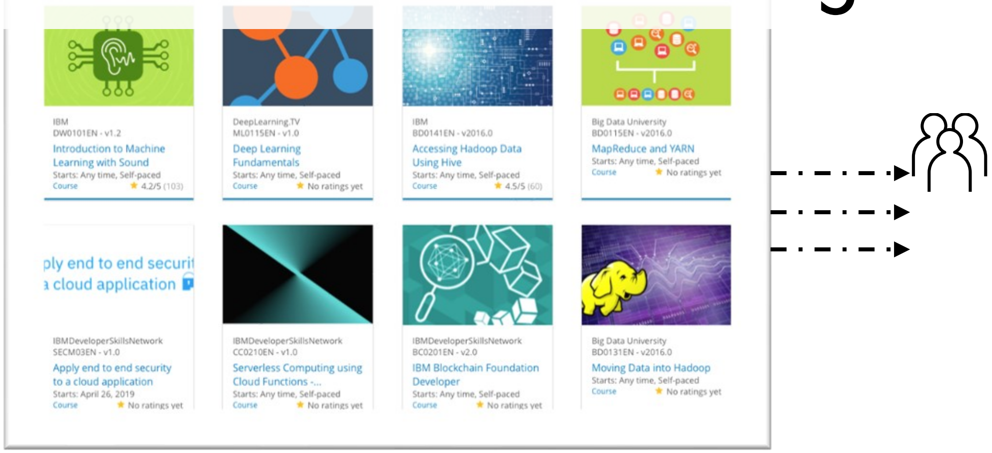
Build a Personalized Online Course Recommender System with Machine Learning

Tao Shan
2022-08-10



IBM
DW0101EN - v1.2
Introduction to Machine Learning with Sound
Starts: Any time, Self-paced
Course ★ 4.2/5 (103)

DeepLearning.TV
ML0115EN - v1.0
Deep Learning Fundamentals
Starts: Any time, Self-paced
Course ★ No ratings yet

IBM
BD0141EN - v2016.0
Accessing Hadoop Data Using Hive
Starts: Any time, Self-paced
Course ★ 4.5/5 (60)

Big Data University
BD0115EN - v2016.0
MapReduce and YARN
Starts: Any time, Self-paced
Course ★ No ratings yet

IBMDeveloperSkillsNetwork
SECM03EN - v1.0
Apply end to end security to a cloud application
Starts: April 26, 2019
Course ★ No ratings yet

IBMDeveloperSkillsNetwork
CC0210EN - v1.0
Serverless Computing using Cloud Functions -...
Starts: Any time, Self-paced
Course ★ No ratings yet

IBMDeveloperSkillsNetwork
BC0201EN - v2.0
IBM Blockchain Foundation Developer
Starts: Any time, Self-paced
Course ★ No ratings yet

Big Data University
BD0131EN - v2016.0
Moving Data into Hadoop
Starts: Any time, Self-paced
Course ★ No ratings yet

Outline

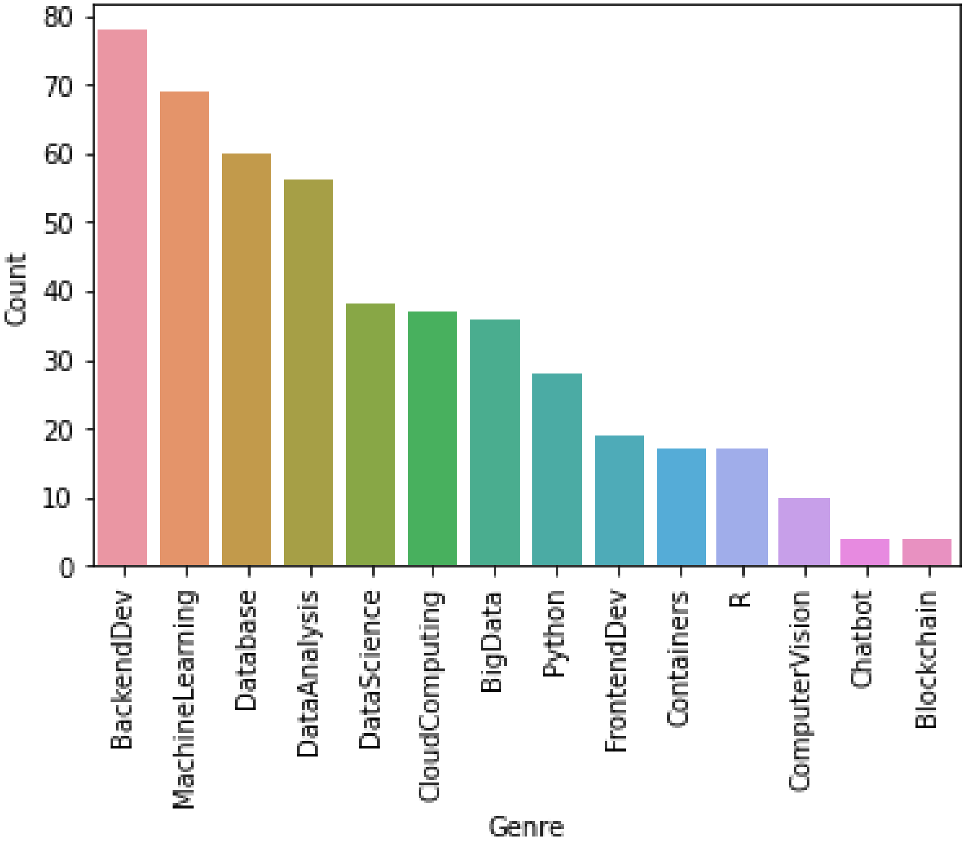
- Introduction and Background
- Exploratory Data Analysis
- Content-based Recommender System using Unsupervised Learning
- Collaborative-filtering based Recommender System using Supervised learning
- Conclusion
- Appendix

Introduction

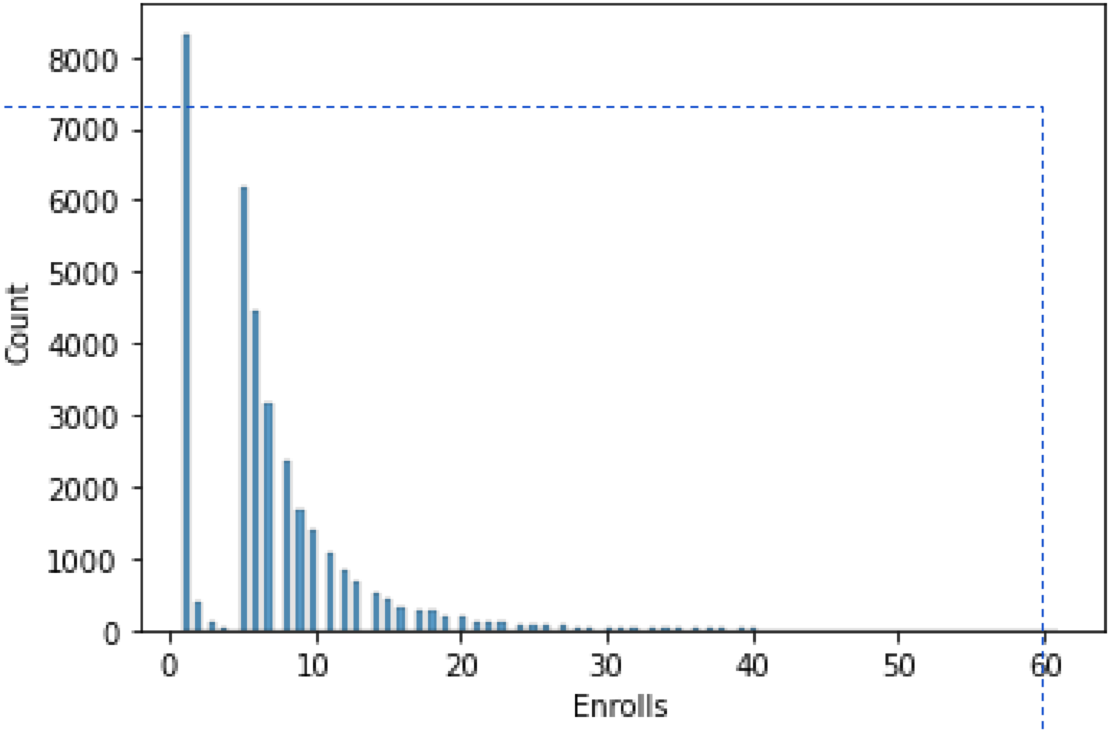
- Massive Open Online Courses (MOOCs) startup - AI Training Room
- Recommendation system for people select courses
- Assume we know people’s interest such as python, Cloud Computing

Exploratory Data Analysis

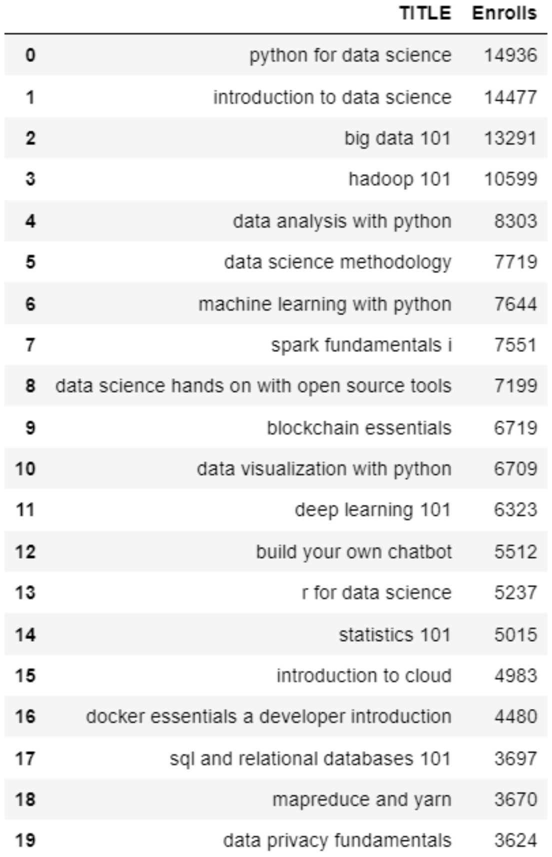
Course counts per genre



Course enrollment distribution

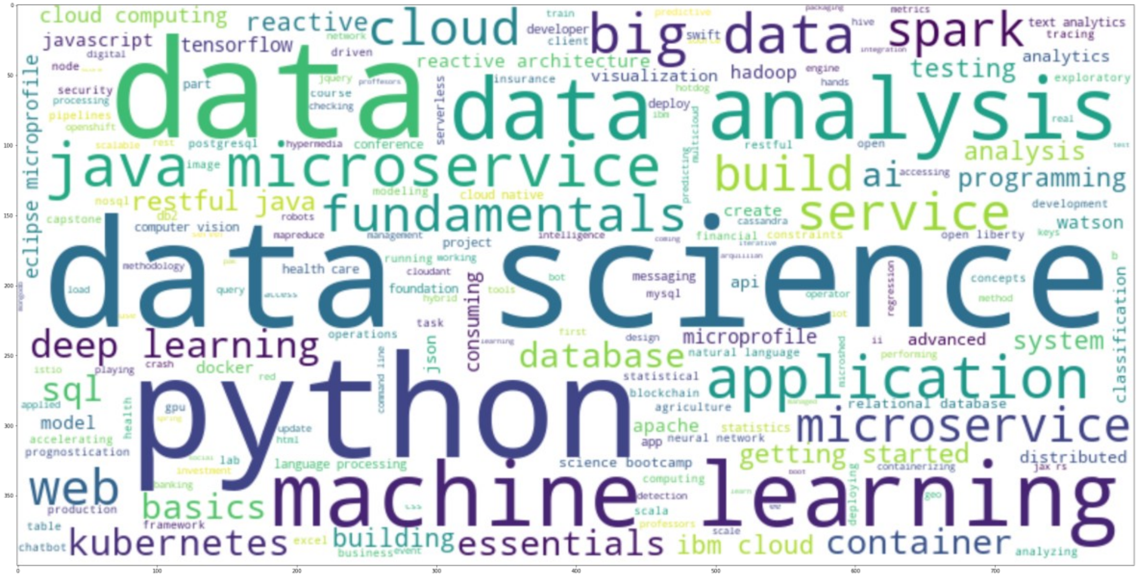


20 most popular courses

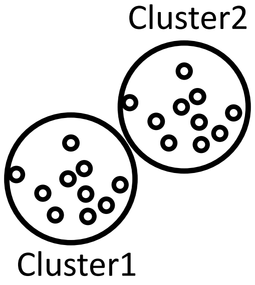


	TITLE	Enrolls
0	python for data science	14936
1	introduction to data science	14477
2	big data 101	13291
3	hadoop 101	10599
4	data analysis with python	8303
5	data science methodology	7719
6	machine learning with python	7644
7	spark fundamentals i	7551
8	data science hands on with open source tools	7199
9	blockchain essentials	6719
10	data visualization with python	6709
11	deep learning 101	6323
12	build your own chatbot	5512
13	r for data science	5237
14	statistics 101	5015
15	introduction to cloud	4983
16	docker essentials a developer introduction	4480
17	sql and relational databases 101	3697
18	mapreduce and yarn	3670
19	data privacy fundamentals	3624

Word cloud of course titles

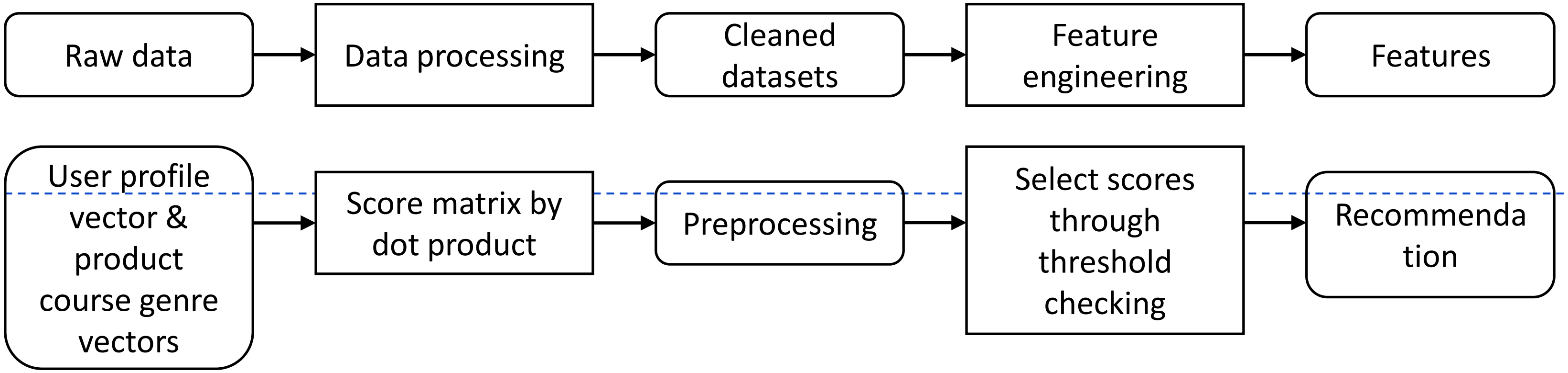


Content-based Recommender System using Unsupervised Learning



Flowchart of content-based recommender system using user profile and course genres

- Raw data: Course textual data and user profile data
- Data preprocessing: word tokenizer
- Cleaned datasets: stop word removal, upper case, stemming, lemmatization
- Feature engineering: Bow of words (doc2bow)
- Features: user profile dot product with course genre vector



Raw data

Data processing

Cleaned datasets

Feature engineering

Features

User profile vector & product course genre vectors

Score matrix by dot product

Preprocessing

Select scores through threshold checking

Recommenda tion

Evaluation results of user profile-based recommender system

Dot product value threshold for recommendation: I tried threshold value = 10,15,20,40 found 20 should be great.

Average number of recommendation

Threshold value = 10: 61.8 recommendations

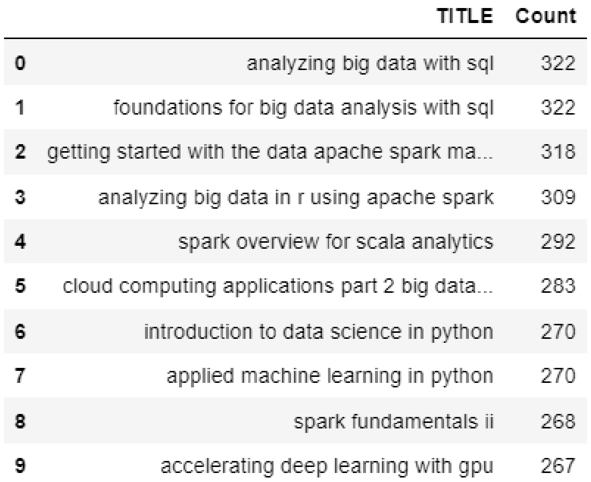
15: 28.3 recommendations

20: 19.1 recommendations

40: 2.0 recommendations

What are the most frequently recommended courses?

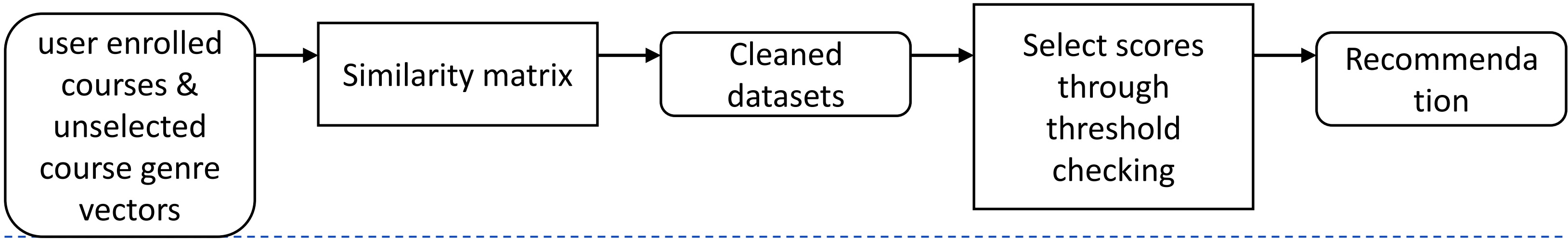
For Threshold value = 20:



	TITLE	Count
0	analyzing big data with sql	322
1	foundations for big data analysis with sql	322
2	getting started with the data apache spark ma...	318
3	analyzing big data in r using apache spark	309
4	spark overview for scala analytics	292
5	cloud computing applications part 2 big data...	283
6	introduction to data science in python	270
7	applied machine learning in python	270
8	spark fundamentals ii	268
9	accelerating deep learning with gpu	267

Flowchart of content-based recommender system using course similarity

- Raw data: user rated history, and which course user enrolled or not
- We need to find the cosine similarity (or using other distance matrics) between chosen course and unchosen course, to check if there is any unknown course can recommend to the person.
- Preprocessing method similar as before

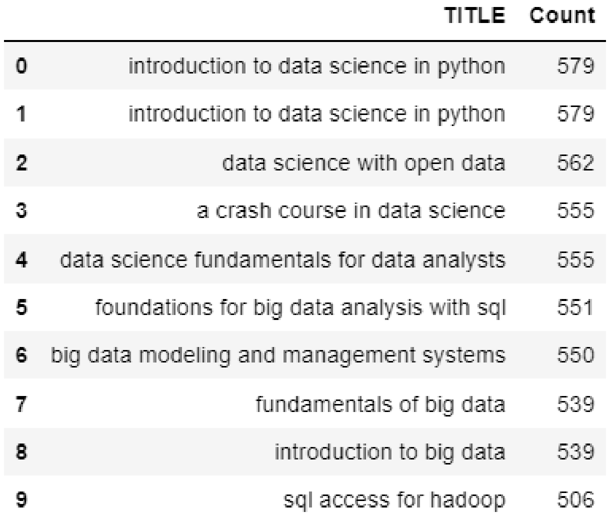

user enrolled courses & unselected course genre vectors
Similarity matrix
Cleaned datasets
Select scores through threshold checking
Recommenda tion

Evaluation results of course similarity based recommender system

similarity threshold: 0.6

On average, 11.4 courses have been recommended per user (in the test user dataset)

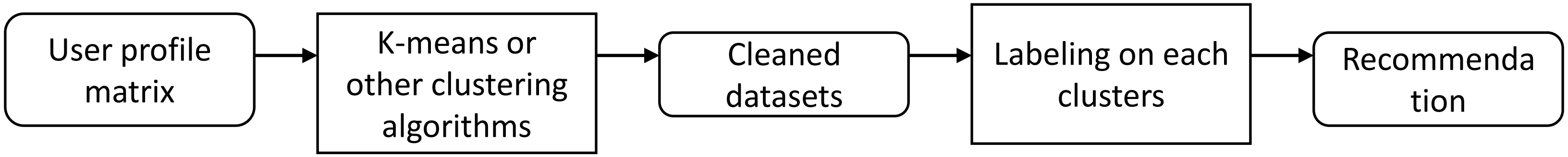
What are the most frequently recommended courses?



	TITLE	Count
0	introduction to data science in python	579
1	introduction to data science in python	579
2	data science with open data	562
3	a crash course in data science	555
4	data science fundamentals for data analysts	555
5	foundations for big data analysis with sql	551
6	big data modeling and management systems	550
7	fundamentals of big data	539
8	introduction to big data	539
9	sql access for hadoop	506

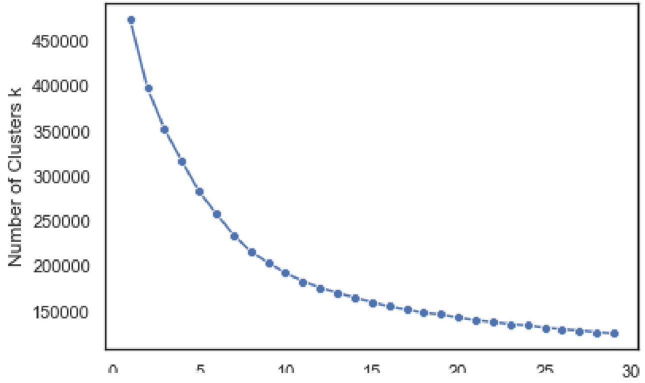
Flowchart of clustering-based recommender system

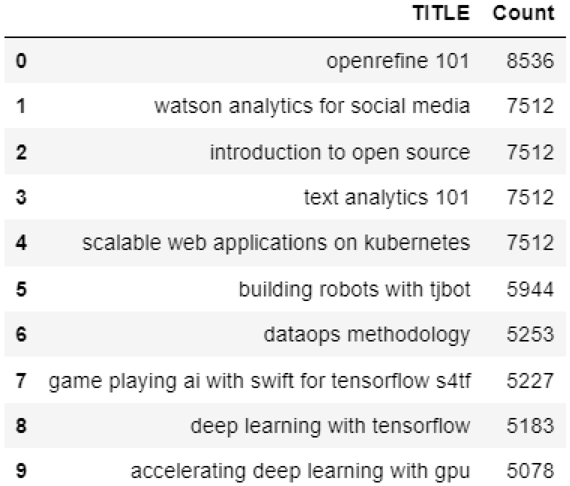
- Raw data: user rated history, and which course user enrolled or not
- We need to find the clusters among all the users, to find similar users to recommend new courses.
- Preprocessing method similar as before


User profile matrix
K-means or other clustering algorithms
Cleaned datasets
Labeling on each clusters
Recommenda tion

Evaluation results of clustering-based recommender system

K means hyper parameter: choose k = 30, since sum of square distance is the smallest.

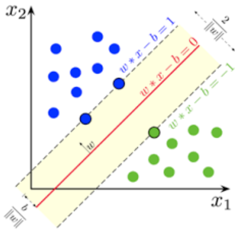




	TITLE	Count
0	openrefine 101	8536
1	watson analytics for social media	7512
2	introduction to open source	7512
3	text analytics 101	7512
4	scalable web applications on kubernetes	7512
5	building robots with tjbot	5944
6	dataops methodology	5253
7	game playing ai with swift for tensorflow s4tf	5227
8	deep learning with tensorflow	5183
9	accelerating deep learning with gpu	5078

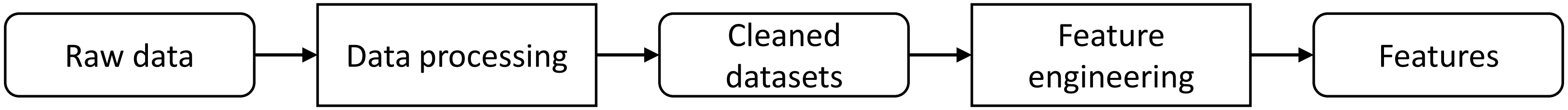
On average, 4 courses have been recommended per user (in the test user dataset)

Collaborative-filtering Recommender System using Supervised Learning



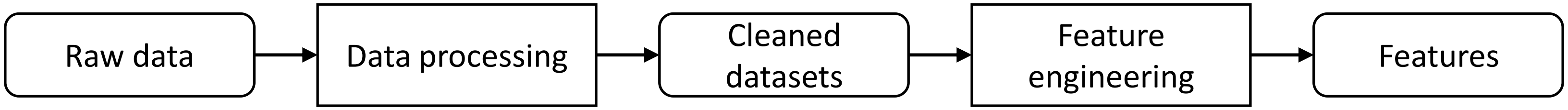
Flowchart of KNN based recommender system

- Raw data: user preference profile
- We are using KNN to find the nearest neighbors for users. For these nearest neighbors for users, We can find the similar user's choice for courses, then recommend to our users
- Preprocessing method similar as before



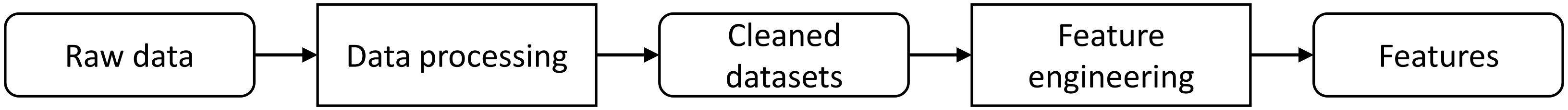
Flowchart of NMF based recommender system

- Raw data: user preference profile
- We are using NMF to find the user and item matrixes for users. The advantages is to reduce the high dimension for previous user matrix in KNN.
- Preprocessing method similar as before

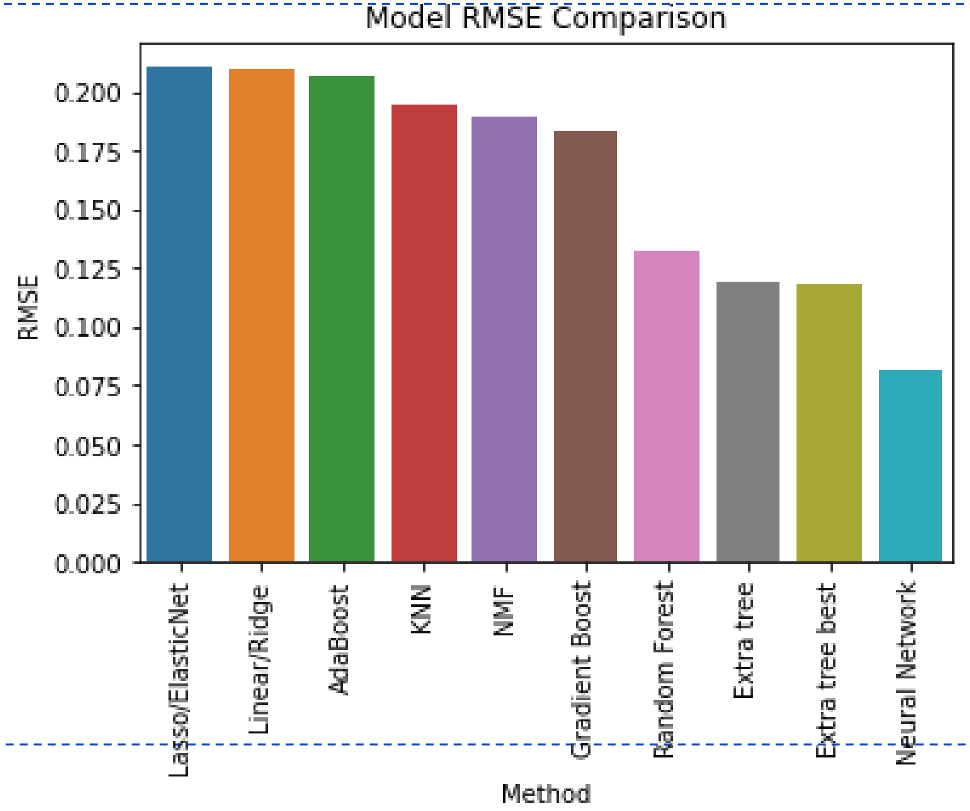


Flowchart of Neural Network Embedding based recommender system

- Raw data: user preference profile
- We are using neural networks to learning patterns from data and extract latent features.
- Preprocessing method similar as before



Compare the performance of collaborative-filtering models



Conclusions

- Popular Courses are all relates to data analytics
- content-based recommender system, user profile-based recommender system, course similarity based recommender system are all great recommendation systems. A essential number of courses are recommended to users
- Among KNN, NMF, Neural network, and other modeling strategies, Neural networks has the best performance, Extra Tree Regressor and Random forest also has great performance

Appendix

- Project Codes: https://github.com/davidshan0814/Recommendation-System