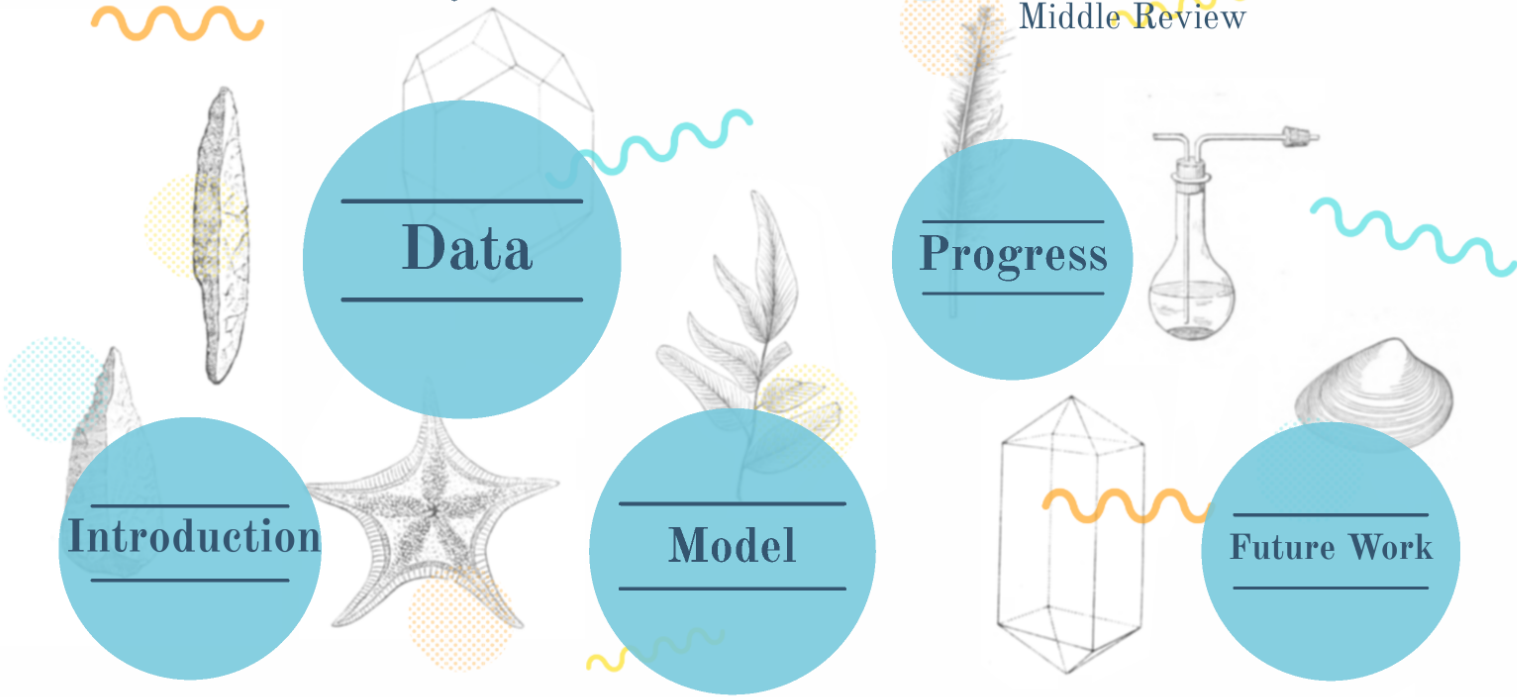
Online Education Recommendation System

Hongfei Xu (011833978)

Summer 2019

CMPE 256 Individual Project

Middle Review



Introduction

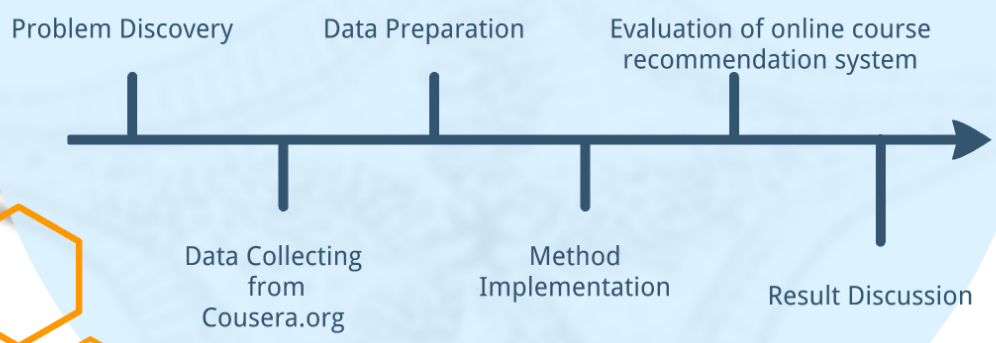
With the rapid development of information and network technology, online courses are increasingly favored by people all over the world.
This project aims to design a recommendation system that recommends courses to users based on historical data or mines potential users for specific courses.

Project Timeline

Approaches

Timeline

This project will be conducted in two months, including several phases as shown in timeline.



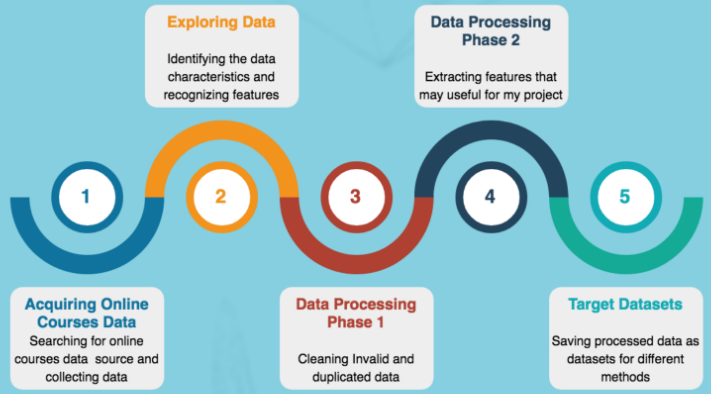
Approaches

To build a recommendation system for online course, several methods are applicable to the data.
Considering the characteristic of the data, content-based and knowledge-based methods seems to be appropriate - the information of the courses is sufficient, and the users shows their preference and requirements for filtering.

Collaborative filtering may also work for my problem, however, most of the users comment the courses privately, which causes it's hard to collect ratings information from each user.

Data

Data preparation is one of the important parts of my project. In order to get the target datasets, I went through several steps: data acquiring, data exploring and data processing.



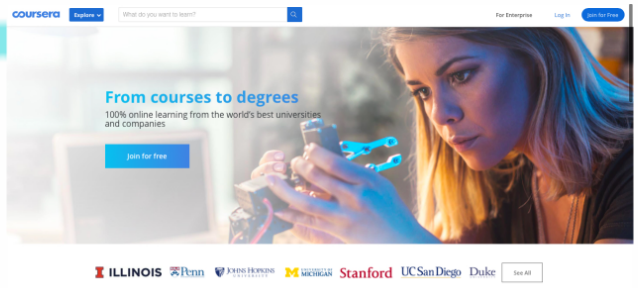
Data Acquiring

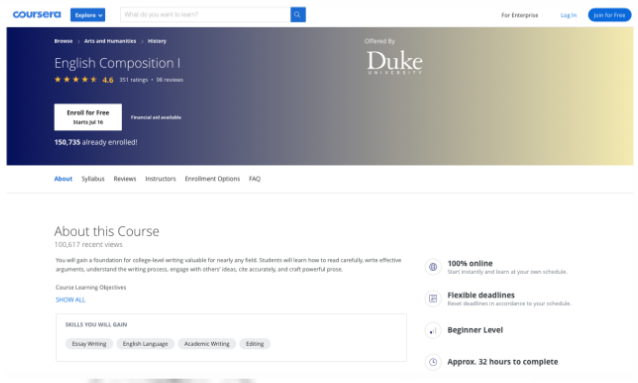
Data Exploring

Data Processing

Data Acquiring

The most well-known websites that offer online courses are *Coursera*, *Udacity* and *edX*.
In my Project, the information of online courses from *Coursera* are my target data. Why *Coursera*? Because *Coursera* has more than 35 million users around the world and collaborates with 27 countries. Its rich content and features are providing useful information for data analysis purpose.
As representation, I've applied API to fetch course information from the most popular courses under each category for my project.





Data Exploring

After collecting the raw data, the key step before feature extraction and data analysis is to understand the data characteristics. There are various attributes of each course's information. The details of the attributes are shown in the table.

Feature	Description
Course ID	Unique ID for recommendation system
Category	Category of course
Sub-category	Sub category under the main categories
Ratings	Overall ratings from all students for this course
About	Summary of the course
Provider	The university/organization providing the course
Enrolled Number	Total number of students that have taken this course
Level	Difficulty of the course for potential users
Skills	Skills that will gain after completing the course
Spending Time	The estimates for time spending on this course
Subtitle	Subtitle language(s) provided by the course

Data Processing

With understanding of the data, the process of dataset has started.

Since there are some missing values in the raw data, we have to drop them at first step.

Some features may be constructed by classification method, for knowledge-based recommendation use, I've considered to encode them by binary numbers.

```
df.head()
```

	category	sub_category	Provider	star	enrolled	About_this_Course	Approx	Subtitles	SKILL	URL
0	Health	Psychology	Princeton University	4.8	292360.0	About this Course 100,552 recent viewsThe Dala...	Approx. 18 hours to complete Suggested: 2-5 ho...	English, Spanish, Hungarian	Skills you will gain Philosophy Psychology Min...	https://www.coursera.org/learn/science-of-medi...
1	Arts and Humanities	History	Duke University	4.6	150555.0	About this Course 99,683 recent viewsYou will ...	Approx. 32 hours to complete	English, Spanish, French	Skills you will gain Essay Writing English Lan...	https://www.coursera.org/learn/english-composi...
2	Arts and Humanities	History	University of Virginia	4.5	73381.0	About this Course 22,776 recent viewsA unique ...	Approx. 10 hours to complete Suggested: 10 hou...	English	NaN	https://www.coursera.org/learn/historical-fiction
3	Physical Science and Engineering	Physics and Astronomy	Stanford University	4.9	68439.0	About this Course 85,283 recent viewsIn this c...	Approx. 40 hours to complete Suggested: 8 week...	English	NaN	https://www.coursera.org/learn/einstein-relati...

Overview of part of processed data

Model

Considering the data attributes and the information the online course data has provided, the two idea model to deal with this project are content-based and knowledge based recommendation system.

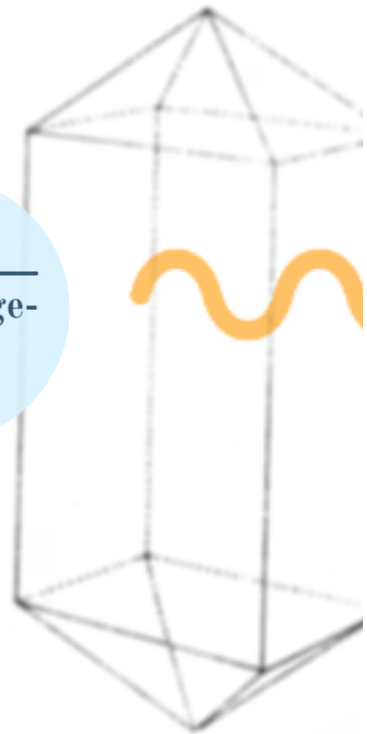
Feature with context such as course summary and skills can be used as input for content-based method. While other features with clear classification or numeric characteristic may be applied with knowledge-based model.

Collaborative filtering (CF) is also an approach to solve the problem, but the issue is that the data of each user's review is not sufficient.

Content-Based

Knowledge-Based

Collaborative Filtering



Content-Based

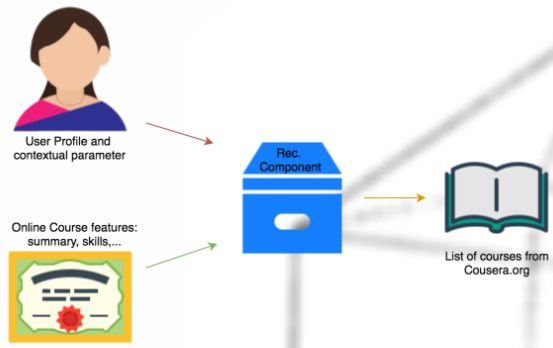
Content of the courses can be represented as text description by some features.
In this recommendation part, the measure I take is TF-IDF.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF
Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Input with features: course summary, skills will gain from course.



System Chart

Knowledge-Based

In this part, the measure applied to my online course recommendation system is "constraint-based". Since most of the users could explicitly define a set of requirements.

Therefore, the task of this knowledge-based model is to fulfill the user's constraints.

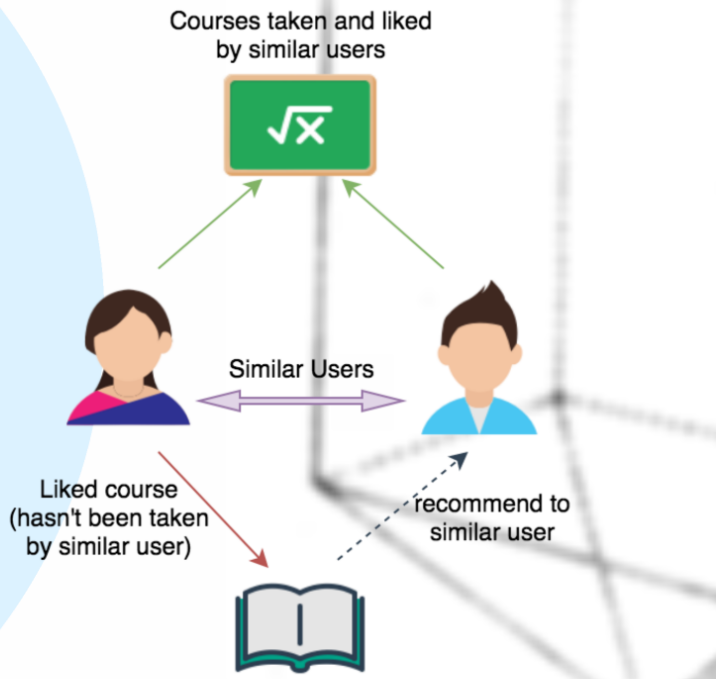
Input with features: course level, provider, subtitle, category, overall rating.

System Chart

Collaborative Filtering

Collaborative Filtering (CF) is a proper method to locate the users with same interest on courses, and mine the potential users to recommend online course.

However, in the analysis process, the problem came up to me that most of the users would like to comment anonymously, or keep their review in private status. Since that, I've tried to collect the data that commented under Coursera user forum and explored the approach to establish CF model on user-based method.



System Chart

Partial Results

- The implementation of content-based recommendation system is almost done. By using TF-IDF tools provided *Sklearn* library in Python, the system will recommend user courses by searching keywords.
- Implementation of knowledge-based recommendation system is in progress. The model of constraint-based filtering is constructing.
- As mentioned in previous part, dataset for CF method is still in adjustment. The similarity score of users will be calculated after the dataset adjustment.

	able	about	academic	accessories	accounting	accurately	active	actually	address	addressed	...	writing	written	wrote	year	years
0	0.000000	0.063726	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000	0.0
1	0.000000	0.035763	0.000000	0.0	0.0	0.102705	0.0	0.000000	0.000000	0.000000	...	0.513527	0.102705	0.0	0.000000	0.0
2	0.000000	0.052798	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000	0.0
3	0.000000	0.046681	0.044686	0.0	0.0	0.000000	0.0	0.044686	0.044686	0.000000	...	0.000000	0.000000	0.0	0.134059	0.0
4	0.041072	0.042905	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.041072	...	0.000000	0.000000	0.0	0.000000	0.0

TF-IDF word matrix

	category	sub_category	Provider	star	enrolled	About_this_Course	Approx	Subtitles	SKILL	URL	writi
1	Arts and Humanities	History	Duke University	4.6	150555.0	About this Course , recent viewsYou will gain ...	Approx. 32 hours to complete	English, Spanish, French	Skills you will gain Essay Writing English Lan...	https://www.coursera.org/learn/english-composi...	0.5135

Content-Based Recommendation testing:
Recommending course by keywords "writing"
(Last column shows the TF-IDF score)

Future Work

Work for next period of my project:

- Completing the knowledge-based model
- Implementing CF recommendation system (user-based)
- Evaluation on the three recommendation strategies and fine-tuning the parameters
- Testing and result discussion

Thank You