

# **CMPE 256 Summer 2019**

Professor: Shih-Yu Chang

## **Individual Project Report**

Education Recommendation System  
Online Course (Coursera) Recommendation System

Name: Hongfei Xu

ID: 011833978

## Contents

<b>Abstract</b>	<b>1</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background	1
1.2 Motivation and Problem Discussion	1
1.3 Plan of Approach	1
<b>2. Data Preparation</b>	<b>2</b>
2.1 Data Collection	2
2.2 Data Pre-processing	2
2.3 Data Analysis	2
2.4 Feature Extraction	5
<b>3. Solution Implementation and Results</b>	<b>5</b>
3.1 System Design	5
3.2 Implementation	5
3.2.1 Knowledge-based Recommender	5
3.2.2 TF-IDF and Content-based Recommender	6
3.2.3 Hybrid System	8
<b>4. User Interface</b>	<b>9</b>
4.1 UI Design and Implementation	9
4.2 Test Results on UI	10
<b>5. Evaluation and Impact Discussion</b>	<b>10</b>
5.1 Evaluation	10
5.2 Impact Discussion	10
<b>Acknowledgment</b>	<b>11</b>

## **Abstract**

This project aims to build an online education recommendation system for public users. Specifically, online course recommendation is adopted as my topic. Using the information from large-scale MOOC website, a knowledge-based recommender, a content-based recommender and a hybrid system have been established.

(App URL - <https://xu256indiv.herokuapp.com/> It may take a few minutes to load if you open my app for the first time.)

## **1 Introduction**

### **1.1 Background**

With the development of the network, online teaching has become more and more convenient. In today's society, people at all levels are eager to upgrade themselves from different aspects. Therefore, MOOC (Massive Open Online Course) has become an unmissable choice, and how to choose a course that suits themselves and meets the requirements of self-improvement has turned to a problem that needs to be solved.

The most well-known websites that offer online courses are *Coursera*, *Udacity* and *edX*.

In my Project, the design purpose is to analyze data from Coursera and create a recommendation for its courses, since Coursera has more than 35 million users around the world and collaborates with 27 countries, moreover, its rich content and features are providing useful information for data analysis purposes.

### **1.2 Motivation and Problem Discussion**

Coursera offers a large number of courses to users, but according to my personal and peer experience, its recommendation function is not perfect - users can only filter the courses according to certain specified characteristics, but can not be completely customized own needs (such as the inability to choose according to the overall ratings and class enrolled number), and Coursera only periodically recommends the popular courses in each category instead of customized recommendation based on the course contents.

The purpose of this individual project is to build a recommendation model for certain course attributes that are not supported by the Coursera official website recommendation function, and to design a hybrid system in combination with the course contents.

### **1.3 Plan of Approach**

In order to achieve the above objectives, the project will be divided into several stages. The first stage is the collection and preprocessing of the data, then the analysis and extraction of features, the second stage is the system design and modeling for the dataset, in view of the problems mentioned above, the recommended system will be mainly based on the

knowledge-based method, supplemented by a content-based model. Finally, the UI will be built on top of the above and the system will be evaluated based on selected user feedback.

## 2. Data Preparation

### 2.1 Data Collection

The raw data was collected from popular courses under each major of *Coursera.org* using web crawling techniques. The attributes included in the raw data are course name, course description (or syllabus), provider information, enrolled students number, overall ratings by students, subtitles, target skills, Coursera url and category (some of them may have the attribute of sub-category). Figure 1 shows an overview of head part of the raw data. (Crawling code is included in “dataset” folder.)

Since my computer performance is relatively mediocre and can't handle too much data at the same time, I have sliced the data and used the most popular courses in each category as a sample to create this project.

	name	category	sub-category	provider	overall	enrolled	about	subtitles	skills	url
0	Introduction to Philosophy	Philosophy	Arts and Humanities	The University of Edinburgh	4.6	285712	111,820 This course will introduce you to som...	English, Hebrew	NaN	<a href="https://www.coursera.org/learn/philosophy">https://www.coursera.org/learn/philosophy</a>
1	Think Again I: How to Understand Arguments	Philosophy	Personal Development	Duke University	4.6	139699	65,149 How to Understand Arguments\n\nThink A...	Arabic, Ukrainian, Chinese (Simplified), Itali...	Evaluation Interpretation Language Linguistics	<a href="https://www.coursera.org/learn/understanding-a...">https://www.coursera.org/learn/understanding-a...</a>
2	De-Mystifying Mindfulness	Philosophy	Personal Development	Universiteit Leiden	4.8	109064	80,498 Interest in meditation, mindfulness, a...	English	Philosophy Gratitude Mindfulness Meditation	<a href="https://www.coursera.org/learn/mindfulness">https://www.coursera.org/learn/mindfulness</a>
3	Greek and Roman Mythology	Philosophy	Arts and Humanities	University of Pennsylvania	4.8	92630	59,950 Myths are traditional stories that hav...	English, Romanian, Chinese (Simplified)	Art History Greek Mythology History Mythology	<a href="https://www.coursera.org/learn/mythology">https://www.coursera.org/learn/mythology</a>

Figure 1. Overview of Raw Data

### 2.2 Data Pre-processing

After obtaining the raw data, the information about each attribute has been checked. From the tools from Pandas Dataframe in Python, it shows that there are some missing values in course description, target skills and providers (detailed process has been shown in file “#1.popular\_courses\_data.ipynb”). Since all of them are string features and cannot be verified due to the incomplete information from the official website, I’ve filled all the missing values with string “unknown”. Meanwhile, the duplicated data has also been removed from the dataset.

### 2.3 Data Analysis

By checking the category attribute, it is found that there are more than forty kinds of disciplines (detailed process has been shown in file “#1.popular\_courses\_data.ipynb”), some of which are very complicated and trivial, which is not convenient for beginners and ordinary



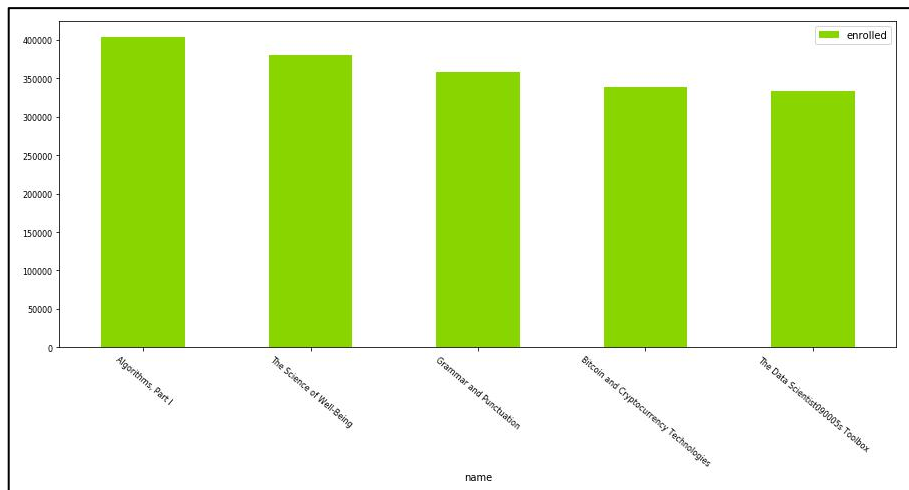


Figure 3. Top 10 Courses with Largest Number of Enrolled Students

Based on the results above, the courses in all 43 disciplines can be divided into six areas that are easy to understand for the general public (this classification is also commonly used in the division of higher education institutions). The specific division is shown in Table 1.

After completing the classification, a new column named “topic” has been added to the dataframe, the value is the area name based on Table 1.

Area	Disciplines
Nature Science	Basic Science, Chemistry, Environmental Science and Sustainability, Physics and Astronomy, Research, Research Methods
Computer Science and Math	Algorithms, Cloud Computing, Computer Security and Networks, Data Analysis, Data Management, Machine Learning, Mobile and Web Development, Networking, Probability and Statistics, Software Development
Engineering	Electrical Engineering, Mechanical Engineering, Security
Business	Business Essentials, Business Strategy, Economics, Entrepreneurship, Finance, Leadership and Management, Marketing, Support and Operations
Health	Animal Health, Health Informatics, Healthcare Management, Nutrition, Patient Care, Public Health
Language, Arts and Social	Design and Product, Education, Governance and Society, History, Law, Learning English, Music and Art, Other Languages, Philosophy, Psychology

Table 1. Area classification of all Disciplines in this Project

Another analysis of the data is the statistic of two numeric attributes: overall ratings of a course from users and the enrolled number. These two attributes shows the quality of the course, and how popular among the users, the statistics of which would be used in model building part.

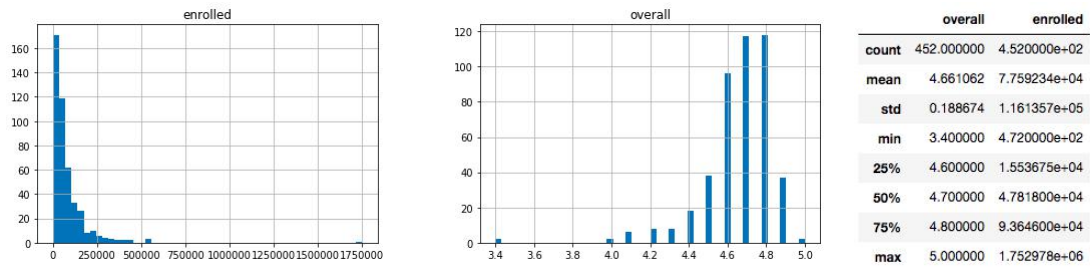


Figure 4. Enrolled Number Histogram (left), Overall Ratings Histogram (Middle), and Statistics (Right)

Through the plotting and statistics above, it is found that the two attributes is nearly normal distribution, and this would be applied to knowledge-model as reference.

## 2.4 Feature Extraction

From the steps above, the new feature “topic” has been extracted from the attributes category and sub-category. Since the two numeric attributes “enrolled number” and “overall ratings” also in normal distribution, these two features were kept in further use. Other attributes included in datasets are “course name”, “course description (about)”, “provider”, “skills you will gain”, “subtitles” and url, which are basic information that may help users to filter their choices.

## 3. Solution Implementation and Results

### 3.1 System Design

As stated in section 1.3, the aim of this project is to build a recommendation system by which the users may customize their selections from features that wasn’t supported by Coursera official filter function. In this case, three features mentioned above, “topic”, “enrolled number” and “overall ratings” would be combined with other basic attributes of courses to establish a knowledge-based recommender. Furthermore, to recommend other similar courses that the users may have interests in, a content-based recommendation strategy needs to be constructed together to form a hybridization system.

### 3.2 Implementation

#### 3.2.1 Knowledge-based Recommender

To solve the problem of recommending items based on user’s needs, constraint-based recommendation has been adopted in my project. Since it’s a constraint satisfaction problem, a knowledge model between user’s preferences and course features should be created.

The users may have interests in some areas, but not very familiar with the academic vocabulary, therefore the 43 disciplines were divided into six main areas that users are easy to understand in part 2. This step created a connection between users and items. Similarly, the



feature “enrolled number” has been divided into five levels according to the statistics, which are 2000, 5000, 10000, 20000 and 50000 - which will show an idea and range to users, in order to help them make a decision.

To deal with the context features (course name, course description, target skills, subtitles, topic), my python code will execute several Boolean functions correspondingly to find the rows in dataframe that contains the words. While for the numeric features (enrolled number and overall ratings), the code will make a logical judgment. The total number of features as inputs in knowledge-based (constraint-based) model are seven, and the recommendation system will find the intersection of seven features to fulfill the requests from users. Figure 5 shows a sample result of using system, and the detailed code is included in “#3.kb\_recommender.ipynb”.

inter_set('duke', None, 'data', 'english', 4, 20000, 'Business')									
	name	provider	overall	enrolled	about	subtitles	skills	url	topic
35	Mastering Data Analysis in Excel	Duke University	4.2	260590	128,508 Important: The focus of this course i...	English	Binary Classification Data Analysis Microsoft...	https://www.coursera.org/learn/analytics-excel	Business
82	Business Metrics for Data-Driven Companies	Duke University	4.6	129497	111,811 In this course, you will learn best p...	English, Arabic, Chinese (Simplified)	Data Analysis Business Analysis Business Anal...	https://www.coursera.org/learn/analytics-busin...	Business

Figure 5. Implementation Result of Knowledge-based Recommender

### 3.2.2 TF-IDF and Content-based Recommender

If the user get recommendation in my knowledge-based system and hope to acquire more information about the similar courses, the content-based strategy may contribute to user’s preferences.

The three features that may contain useful information about course contents and user’s requirements are “course name”, “course description (about)” and “skills”.

As we’ve learned from the lecture, the TF-IDF combined with cosine similarity is an efficient approach to process the contextual features.

$$w_{x,y} = \text{tf}_{x,y} \times \log\left(\frac{N}{\text{df}_x}\right)$$

**TF-IDF**

Term  $x$  within document  $y$

$\text{tf}_{x,y}$  = frequency of  $x$  in  $y$   
 $\text{df}_x$  = number of documents containing  $x$   
 $N$  = total number of documents

TF-IDF values show the information about how important a word in a set of documents, and by treating words as a vector in matrix, we can find the context similarity between any pair of items.

To obtain the similarity between courses, a TF-IDF should be calculated in advance. Before this procedure, the package of *NLTK* has been used to remove the English stop-words in contextual features; other pre-processing of the data includes word stemming and punctuation



removing. After completing context pre-processing, the noises in the data have been significantly reduced.

The tool applied to calculate TF-IDF value for each word is *TfidfVectorizer* from python package “sklearn.feature\_extraction.text”. In this phase, the features for calculation are “course name”, “course description (about)” and “skills”, with the weights of 60%, 20% and 20%, respectively, for that the name is the key summary of a course. Figure 6 depicts the head part of TF-IDF matrix of course description.

dftx.head()											
	_nsx_itautomation_reg	aa	aaflyg	aaron	aas	ab	abarca	abbreviation	abbreviations	abc	...
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
5 rows x 8726 columns											

Figure 6. TF-IDF Vectors of Course Description

The similarity between any two courses has been calculated based on Cosine Similarity, which measures the distance of vectors in vector spaces. The cosine similarity of course name, description and skills was obtained separately. Figure 7 shows the overview of similarity matrix from course description.

```
array([[1.          , 0.05181216, 0.09935946, ..., 0.02666174, 0.
0.          ],
[0.05181216, 1.          , 0.03383005, ..., 0.0837839 , 0.
0.          ],
[0.09935946, 0.03383005, 1.          , ..., 0.02372969, 0.
0.          ],
...,
[0.02666174, 0.0837839 , 0.02372969, ..., 1.          , 0.
0.          ],
[0.          , 0.          , 0.          , ..., 0.          , 1.
0.16792511],
[0.          , 0.          , 0.          , ..., 0.          , 0.16792511,
1.          ]])
```

Figure 7. Cosine Similarity Matrix of Course Description

As mentioned above, the weights of three features - name, description and skills are 60%, 20% and 20%, therefore, when recommending other similar courses for a specific course, the system will use the sum of three weighted similarity scores, which means all three attributes would contribute to the results. Figure 8 shows the result of recommending similar courses for course “English for Business and Entrepreneurship” (index number: 437) by my content-based system.

tfidf_sim(437)										
	name	provider	overall	enrolled	about	subtitles	skills	url	topic	index
438	English for Journalism	University of Pennsylvania	4.8	98141	72,665 Welcome to English for Journalism, a c...	English, Bengali	, English, Grammar, News, Reporting, English, ...	https://www.coursera.org/learn/journalism	Language, Arts and Social	438
322	English for Science, Technology, Engineering, ...	University of Pennsylvania	4.8	75578	75,941 Welcome to English for Science, Techno...	English	, English, Grammar, Nanotechnology, English, L...	https://www.coursera.org/learn/stem	Nature Science	322
54	Essentials of Entrepreneurship: Thinking & Action	University of California, Irvine	4.4	61887	12,011 Success in business can be greatly enh...	Arabic, Vietnamese, German, English, Spanish	, Strategic, Management, Management, Marketing...	https://www.coursera.org/learn/entrepreneurial...	Business	54
67	Entrepreneurship 1: Developing the Opportunity	University of Pennsylvania	4.8	64281	43,519 How does a good idea become a viable b...	English, Vietnamese, Arabic	, Discovery-Driven, Planning, Elevator, Pitch...	https://www.coursera.org/learn/wharton-entrepr...	Business	67
436	English Composition I	Duke University	4.6	152603	91,956 You will gain a foundation for college...	English, Spanish, French	, Essay, Writing, English, Language, Academic...	https://www.coursera.org/learn/english-composi...	Language, Arts and Social	436

Figure 8. Implementation Result of Knowledge-based Recommender

### 3.2.3 Hybrid System

As explained earlier, a single knowledge-based system does not handle user requirements very well, because it cannot determine the importance of words for content, so in order to avoid this shortcoming, this project uses the tf-idf value as a quantitative basis for text attributes. It is synthesized into the original knowledge-based system to form a complete monolithic hybridization system. This system will not only select courses that clearly meet the basic requirements of users, but also sort the list according to whether the search contains keywords and the influence degree of the keywords.

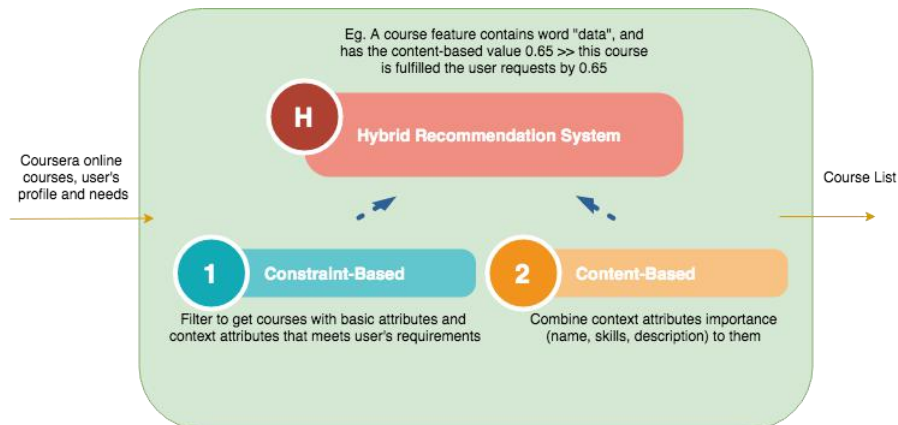


Figure 9. Hybrid System Flow Chart and Explanation for my Project

In this system, the code will go through the knowledge-based judgment function, and return a dataframe with all the courses meets the user's criteria first, then get a score of how they are similar to user's needs by content-based strategy. Figure 10 shows an example test of using my hybrid system to recommend courses (detailed information can be found in # 4.hybrid\_system.ipynb).

```
kb_combined(None, 'python programming', None, 'English', 4.5, None, None).head()
```

	name	provider	overall	enrolled	about	subtitles	skills	url	topic	index	Total
137	R Programming	Johns Hopkins University	4.6	414076.0	605,206 In this course you will learn how to ...	Arabic, French, Chinese (Simplified), Portuguese...	Data Analysis Debugging R Programming Rstudio	https://www.coursera.org/learn/r-programming	Math and Computer Science	137.0	0.862507
115	An Introduction to Interactive Programming In ...	Rice University	4.8	138441.0	148,137 This two-part course is designed to h...	Chinese (Simplified), Italian, Portuguese (Bra...	Programming Principles Python Syntax And Sema...	https://www.coursera.org/learn/interactive-pyt...	Math and Computer Science	115.0	0.739585
89	Python Data Structures	University of Michigan	4.9	288393.0	712,240 This course will introduce the core d...	English, Korean, Arabic	Python Syntax And Semantics Data Structure Tu...	https://www.coursera.org/learn/python-data	Math and Computer Science	89.0	0.612024
92	Introduction to Programming with MATLAB	Vanderbilt University	4.8	174803.0	244,562 This course teaches computer programm...	English, Greek	Computer Programming Problem Solving Matlab P...	https://www.coursera.org/learn/matlab	Math and Computer Science	92.0	0.501775
74	Applied Machine Learning In Python	University of Michigan	4.7	114702.0	261,385 This course will introduce the learne...	English, Korean	Python Programming Machine Learning (ML) Algo...	https://www.coursera.org/learn/python-machine-...	Business	74.0	0.455280

Figure 10. Implementation Result of Monolithic Hybrid Recommender


So far, all system design and implementation has been completed. Three simple experimental results also prove that my recommender systems can optimize the user's course recommendation experience on Coursera to a certain extent, and enable user to customize their own recommendation requirements.

## 4. User Interface

### 4.1 UI Design and Implementation

In order to create a simple UI that is suitable for all users, Flask Web has been applied as a framework of my project. Flask Web enables the python code be visualized on web-page, and the techniques adopted in this part includes HTML and CSS (some templates are common elements in my group project). Detailed UI design code has been uploaded to the folder “project\_web\_application”. Figure 11 shows the interface on web application.

**COURSE RECOMMENDATION SYSTEM**



This is a simple system to recommend courses from Coursera for you.  
Try customized recommendation now by entering your preferences!

Select Field of Interest:  User ratings are higher than...:  Enrolled students are more than...:

Course Keywords:  Provider Name, eg. Duke University:  Skills you want to gain, eg. writing:  Subtitles, eg. English, Chinese...:

Instruction: If you choose constraint-based, system will recommend you with precision results for your keywords and other requirements; if you choose hybrid recommendation, system will calculate similarity from your keywords and get a hybrid result. You can try them to compare!

Figure 11. Index Page of my Project UI

## 4.2 Test Results on UI

Since both hybrid system and knowledge-based system are implemented on web application. Users may choose the one they prefer to get recommendation list. Figure 12 shows the result comparison of the two systems by searching same keywords “data analytics”. They may also get the similar courses recommending on result page.

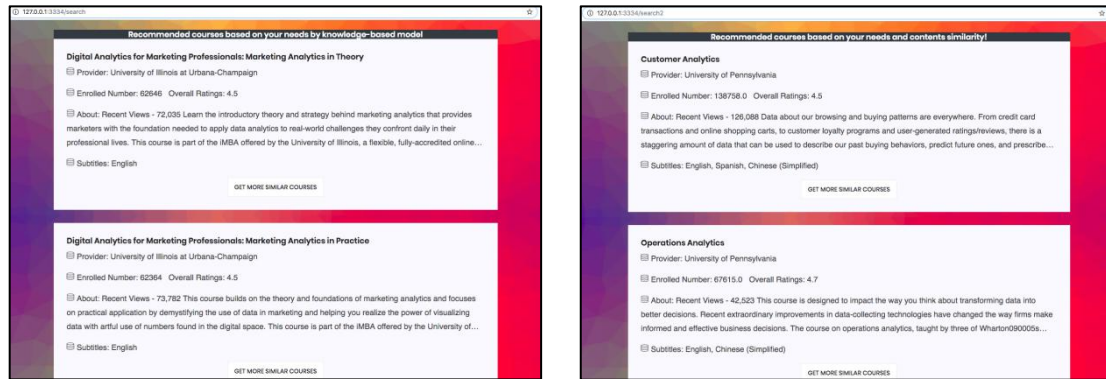


Figure 12. Results Comparison of Different Recommendation Systems on UI

## 5. Evaluation and Impact Discussion

### 5.1 Evaluation

In order to obtain experimental data, I collected feedback from 14 friends around me, 71% of whom said they were satisfied, and 29% thought that the app still has room for improvement. In addition, more experimental participants expressed a preference for using hybrid systems rather than constraint-based system.

In response to these problems, I have summarized the following problems to be solved: the sample size is relatively small due to the insufficient computer performance, which reduces the accuracy; since the website offers some courses information in multiple languages, there may be errors and omissions when dealing with text attributes in languages other than English; solution to solve the understanding and interpretation issue for long paragraphs in description attribute requires more complicated work.

### 5.2 Impact Discussion

Through this program, users can select courses based on more detailed criteria to meet their educational needs, including search attributes not supported by Coursera official website. At the same time, the system will more intelligently recommend courses that users may like based on user preferences and requirements.

**Acknowledgment**

I would like to express my appreciation to Professor Shih-Yu Chang, who gave us excellence lectures in this summer, and teaching assistant Surya Sonti, who helped us a lot after class time.

Thanks to my group members, Xiaoting Jin and Juan Chen, who provided me advice in spare time.