BIA-660 Final Report

Coursera Course Recommendation System Project

Team 2

Hantao Xie

Shunchao Chen

Priyanka Thakur

Bhavana Meena

Instructor: Prof. Rong Liu

1. Motivation

As a student, we always try to enhance our knowledge from all the resources available in our environment. As resources are abundant, and we cannot observe everything at the same time. We should have a plan to achieve our goal. At school, we always get lessons from Teachers and Professors about what we should next and being a student, we try to follow those steps. But when we try to learn on ourselves like online courses, we do not have option of being guided by mentors and we face lot of trouble in deciding what should be done next. For example: while taking courses in Coursera, one of my team members was confused which course to take as she had already taken few beginner courses from another online sources. It would be helpful if she had provision to know as per her current level of knowledge. Absence of recommendation about next level courses in current Coursera recommendation system led us to think about updated Coursera recommendation system.

2. Introduction

a. The objective

Building up a comprehensive course recommender which can consider both the course learner has taken and the current level of the learner.

b. Novelty

As the online education boomed since the year 2010, many people began to study various courses on the Internet and this uprising trend is still continuous in recent years. Because of the popularity of online education, online learners demand more sophisticated course recommenders which can fit with their customized demand. For example, quite a few online learners require recommender to return the courses which are close to the previous courses they took, yet still, show some differences. A learner who has already taken ‘Python for dummies’ does not want to see ‘Python for everyone’ in recommendation result, but most course recommenders in the market do something exactly like this. Our team attempts to design a more advanced recommender.

3. Methodology

a. Algorithms

In this project, we mainly use six different algorithms: for the exploratory data analysis (EDA) part, we initially used two different clustering algorithms, Latent Dirichlet Allocation (LDA) and K-means, to cluster 'Mixed Level' labels to beginners, intermediate, and advanced levels, but these two clustering algorithms do not work well. Later, we implemented classification algorithms and the results have been improved. However, both classification and clustering do not provide a very satisfying result, it probably because the features data we collected do not have a very strong correlation to the dependent variable 'difficulty level'.

When it comes to the part of the recommendation system design, the project implemented cosine similarity and word vector similarity to generate the result. We collected 3 users record to evaluate the result and found that the recommender performed good, but word vector method is better than the cosine similarity method.

b. The input and output of algorithms

The input of the recommender is from the user. The user shall type in a course name which he/she has already taken and his/her current level stage (Beginner, intermediate or advanced). Correspondingly, the system will generate 15 courses: 5 for current difficulty level, 5 for upgraded difficulty level and 5 for mixed level (with difficulty level for reference) as output.

4. Data Scraping and Preprocessing

We used python programming language and Jupyter notebook to finish our project.

Data were scraped from Coursera website "coursera.org "using beautifulsoup. As we know that "Coursera" is giant website and it is tough to scrap all the categories of courses and finish the whole project in small span of time so, we narrow down the scope of categories to four which are Data Science, Computer Science, Social Science and Business. Totally, four features were scrapped: Course name, Description, Difficulty level, Rating and Course Language. Around 2500 data were scraped from four categories and finally we combined all four categories to create one datafile. All

non-English language and without rating were removed. Data with no level were marked as mixed level after cross checking from website. In our final datafile, we had 1521 data rows with all information.

Original data

course	description	level	rate	language
measuring causal effects in the social sciences	How can we know if the differences in wages between men and women are caused by discrin	Approx. 7 hou	4.2	English
social science approaches to the study of chinese society part 2	This course is intended as a first step for learners who seek to become producers of social sc	Beginner Leve	4.3	English
social science approaches to the study of chinese society part 1	This course seeks to turn learners into informed consumers of social science research. It intro	Beginner Leve	4.4	English
methods and statistics in social science - final research project	The Final Research Project consists of a research study that you will perform in collaboration v	Beginner Leve	3.9	English
classical sociological theory	This Massive Open Online Course (MOOC) will offer the participants an introduction into the m	no level	4.8	English
statistics with sas®	This introductory course is for SAS software users who perform statistical analyses using SAS	Intermediate	no rate	English
questionnaire design for social surveys	This course will cover the basic elements of designing and evaluating questionnaires. We will	no level	4.4	English
sampling people, networks and records	Good data collection is built on good samples. But the samples can be chosen in many ways	Beginner Leve	4.5	English
data collection: online, telephone and face-to-face	This course presents research conducted to increase our understanding of how data collectio	Beginner Leve	4.6	English
исследование статистических взаимосвязей	Курс рассматривает способы и инструменты исследования статистических взаимосвяз	Intermediate	4.7	Russian

There were some challenges we faced while scraping data,

1)The web page URL updated, and old script did not work. This issue was fixed with different loops.

2)Page 1 randomly occurs even we change the page number of the loop. This issue was solved by creating a branch sentence

3)Some courses have 'mixed' label in difficulty levels. Clustering and classification were used to solve this issue.

Data after cleaning

course	description	level	rate	language
sustainable business enterpris	This course will explore current challenges and	Mixed Level	4.5	English
excel skills for business: essent	In this first course of the specialization Excel Skills for	Mixed Level	4.9	English
business strategy	In this course you will learn how organizations create,	Mixed Level	4.8	English
Exploring and Producing Data fo	This course provides an analytical framework to help yo	Beginner Level	4.8	English
Brand Management: Aligning B	Professor Nader Tavassoli of London Business School co	Beginner Level	4.9	English
excel skills for business: interm	Spreadsheet software remains one of the most	Intermediate Level	4.9	English
digital business models	Digital business models are disrupting 50-year old	Mixed Level	4.5	English
english for business and entrep	Welcome to English for Business and Entrepreneurship,	Beginner Level	4.8	English
foundations of business strateg	Develop your ability to think strategically, analyze the c	Mixed Level	4.8	English
global impact: business ethics	Global business ethics is the study and analysis of how	Intermediate Level	4.8	English
effective business presentatio	This course is all about presenting the story of the data,	Beginner Level	4.5	English
blockchain and business: applic	Blockchain will bring about profound changes to busine	Mixed Level	4.9	English
excel skills for business: advan	Spreadsheet software remains one of the most	Intermediate Level	4.7	English
business model canvas: a tool f	What you'll achieve:	Beginner Level	4.6	English

5. The exploratory data analysis (EDA)

In EDA part, we tried to predict original difficulty level of "mixed level, so that we can assign them to proper levels which are Beginner, Intermediate and Advanced Level. First, clustering was used to decide the levels, but the output was not satisfied as Advanced level outcome was zero though F1 score was 65%. The reason for this could be the advanced level count was very less in our datafile. Later, we used classification to assign the proper level to mixed level. In classification, we used two models, one is Multinomial Naive Bayes and Support Vector Machine. In Multinomial Naive Bayes, the outcome was like clustering with F1 score value around 61%. In SVM model, we

found that the output was much better than Clustering and Naïve Bayes, it even calculated advanced level and F1 score was more than 70%. Finally, we used SVM to predict the original level of mixed level.

Outcome of Clustering

	precision	recall	f1-score	support
Advanced Level	0.00	0.00	0.00	14
Beginner Level	0.76	0.71	0.74	151
Intermediate Level	0.53	0.68	0.60	90
micro avg	0.66	0.66	0.66	255
macro avg	0.43	0.46	0.44	255
weighted avg	0.64	0.66	0.65	255

Outcome of NB

Advanced Level	0.00	0.00	0.00	14
Beginner Level	0.66	0.97	0.79	151
Intermediate Level	0.71	0.28	0.40	90
micro avg	0.67	0.67	0.67	255
macro avg	0.46	0.41	0.40	255
weighted avg	0.65	0.67	0.61	255

Outcome of SVM

	precision	recall	f1-score	support
Advanced Level	1.00	0.33	0.50	12
Beginner Level	0.77	0.86	0.81	126
Intermediate Level	0.64	0.58	0.61	74
micro avg	0.73	0.73	0.73	212
macro avg	0.80	0.59	0.64	212
weighted avg	0.74	0.73	0.72	212

<u>Final Datafile</u>

	course	description	level	rate	language
0	sustainable business enterprises	This course will explore current challenges an...	Mixed Level (Beginner Level)	4.5	English
1	excel skills for business: essentials	In this first course of the specialization Exc...	Mixed Level (Intermediate Level)	4.9	English
2	business strategy	In this course you will learn how organization...	Mixed Level (Beginner Level)	4.8	English
3	Exploring and Producing Data for Business Deci...	This course provides an analytical framework t...	Beginner Level	4.8	English
4	Brand Management: Aligning Business, Brand and...	Professor Nader Tavassoli of London Business S...	Beginner Level	4.9	English
5	excel skills for business: intermediate i	Spreadsheet software remains one of the most u...	Intermediate Level	4.9	English
6	digital business models	Digital business models are disrupting 50-year...	Mixed Level (Beginner Level)	4.5	English
7	english for business and entrepreneurship	Welcome to English for Business and Entreprene...	Beginner Level	4.8	English
8	foundations of business strategy	Develop your ability to think strategically, a...	Mixed Level (Beginner Level)	4.8	English
9	global impact: business ethics	Global business ethics is the study and analys...	Intermediate Level	4.8	English
10	effective business presentations with powerpoint	This course is all about presenting the story ...	Beginner Level	4.5	English
11	blockchain and business: applications and impl...	Blockchain will bring about profound changes t...	Mixed Level (Beginner Level)	4.9	English
12	excel skills for business: advanced	Spreadsheet software remains one of the most u...	Intermediate Level	4.7	English
13	business model canvas: a tool for entrepreneur...	What you'll achieve:\nIn this project-centered...	Beginner Level	4.5	English
14	(re)-invent your business model with the odyss...	This course gives you access to Odyssey 3.14 -...	Mixed Level (Beginner Level)	4.7	English

6. Algorithm

As mentioned above we used two algorithms to find the recommendation system for Coursera, we used “Cosine similarity” and “Word Vector & Document Vector”

Model 1: Cosine Similarity

Below steps were followed to find the recommended courses as per the input from people, the courses and their level.

Step 1: Import the poster-processing texts, tokenize texts, and
remove stop-words and lemmatization for normalizing
the description and title.

Step 2: Generate a TF-IDF matrix according to the result from
the last step.

Step 3: Calculate the cosine similarity using the TF-IDF matrix

The output will be 15 courses: 5 for current difficulty level, 5 for upgraded difficulty level and 5 for mixed level (with difficulty level for reference)

Model 2: Word Vector & Document Vector

Following steps were performed for result

Step 1: Import the poster-processing texts, tokenize texts, and
remove stop-words and lemmatization for normalizing
the description

Step 2: Label all descriptions with a unique tag

Step 3: Use ‘Doc2vec’ to generate word vectors and document
vectors for every description

Step 4: Extract similarity score for each course

7. The performance comparison

Both the models were providing their own recommendation for courses entered but there was similarity in both recommendation of courses. Both models were providing list of courses from equal level, intermediate level and advanced level as per the input information provided. Input information were course name and difficulty level. Even, we tried to provide recommendations for multiple courses as input.

Example1: For course “Python Data Structure”

course	level	rate	course	level	rate
python data representations	Beginner Level	4.7	using python to access web data	Mixed Level (Beginner Level)	4.8
python programming essentials	Beginner Level	4.8	introduction to data science in python	Intermediate Level	4.5
using python to access web data	Mixed Level (Beginner Level)	4.8	algorithms, part ii	Intermediate Level	5
python for data science	Beginner Level	4.6	python basics	Beginner Level	4.8
python functions, files, and dictionaries	Beginner Level	4.8	algorithms, part i	Intermediate Level	4.9
programming for everybody (getting started with python)	Mixed Level (Beginner Level)	4.8	algorithms for dna sequencing	Mixed Level (Intermediate Level)	4.8
python data analysis	Beginner Level	4.7	python functions, files, and dictionaries	Beginner Level	4.8
data collection and processing with python	Intermediate Level	4.7	pointers, arrays, and recursion	Beginner Level	4.5
introduction to data science in python	Intermediate Level	4.5	calculus: single variable part 1 - functions	Mixed Level (Intermediate Level)	4.8
python classes and inheritance	Intermediate Level	4.7	calculus: single variable part 4 - applications	Mixed Level (Intermediate Level)	4.9
using databases with python	Mixed Level (Intermediate Level)	4.8	java programming: arrays, lists, and structured data	Beginner Level	4.7
data structures	Mixed Level (Intermediate Level)	4.7	calculus: single variable part 3 - integration	Mixed Level (Intermediate Level)	4.9
data visualization with python	Intermediate Level	4.6	learn to program: the fundamentals	Beginner Level	4.7
applied plotting, charting & data representation in python	Intermediate Level	4.5	distributed database systems	Intermediate Level	4
python and statistics for financial analysis	Mixed Level (Beginner Level)	4.5	building database applications in php	Intermediate Level	4.9

Example2: For course “Python Basis”

course	level	rate	course	level	rate
python functions, files, and dictionaries	Beginner Level	4.8	python data analysis	Beginner Level	4.7
python programming essentials	Beginner Level	4.8	using python to access web data	Mixed Level (Beginner Level)	4.8
python data representations	Beginner Level	4.7	python functions, files, and dictionaries	Beginner Level	4.8
python for data science	Beginner Level	4.6	python data representations	Beginner Level	4.7
python classes and inheritance	Intermediate Level	4.7	python data structures	Mixed Level (Beginner Level)	4.9
python data analysis	Beginner Level	4.7	introduction to data science in python	Intermediate Level	4.5
data collection and processing with python	Intermediate Level	4.7	algorithms, part ii	Intermediate Level	5
programming for everybody (getting started with python)	Mixed Level (Beginner Level)	4.8	learn to program: crafting quality code	Mixed Level (Beginner Level)	4.6
using python to access web data	Mixed Level (Beginner Level)	4.8	programming for everybody (getting started with python)	Mixed Level (Beginner Level)	4.8
introduction to data science in python	Intermediate Level	4.5	algorithms, part i	Intermediate Level	4.9
python data structures	Mixed Level (Beginner Level)	4.9	javascript, jquery, and json	Intermediate Level	4.7
build a modern computer from first principles: nand to tetris part ii (project-centered course)	Mixed Level (Intermediate Level)	5	using databases with python	Mixed Level (Intermediate Level)	4.8
object oriented programming in java	Intermediate Level	4.7	parallel programming	Intermediate Level	4.5
be persuasive: write a convincing position paper or policy advice (project-centered course)	Mixed Level (Beginner Level)	4.5	learn to program: the fundamentals	Beginner Level	4.7
principles of computing (part 1)	Intermediate Level	4.8	java programming: arrays, lists, and structured data	Beginner Level	4.7

Multi-Input Recommendation:

course	level	rate	course	level	rate
machine learning foundations: a case study approach	Mixed Level (Intermediate Level)	4.6	matrix factorization and advanced techniques	Mixed Level (Intermediate Level)	4.2
machine learning with python	Mixed Level (Beginner Level)	4.7	big data applications: machine learning at scale	Advanced Level	3.8
how to win a data science competition: learn from top kagglers	Advanced Level	4.7	introduction to machine learning	Intermediate Level	4.7
launching into machine learning	Intermediate Level	4.6	convolutional neural networks	Intermediate Level	4.9
how google does machine learning	Intermediate Level	4.6	probabilistic graphical models 1: representation	Advanced Level	4.7
machine learning for data analysis	Mixed Level (Intermediate Level)	4.2	deep learning in computer vision	Advanced Level	3.9
internet giants: the law and economics of media platforms	Mixed Level (Beginner Level)	4.8	practical reinforcement learning	Advanced Level	4.1
be persuasive: write a convincing position paper or policy advice (project-centered course)	Mixed Level (Beginner Level)	4.5	probabilistic graphical models 3: learning	Advanced Level	4.6
big data applications: machine learning at scale	Advanced Level	3.8	statistical mechanics: algorithms and computations	Mixed Level (Beginner Level)	4.8
advanced machine learning and signal processing	Advanced Level	4.6	sequence models	Intermediate Level	4.8
fundamentals of machine learning in finance	Intermediate Level	3.6	practical machine learning on h2o	Intermediate Level	4.5
guided tour of machine learning in finance	Intermediate Level	3.7	algorithms on graphs	Intermediate Level	4.7
feature engineering	Intermediate Level	4.4	bayesian methods for machine learning	Mixed Level (Intermediate Level)	4.6
probabilistic graphical models 3: learning	Advanced Level	4.6	visual perception for self-driving cars	Mixed Level (Advanced Level)	4.4
image understanding with tensorflow on gcp	Advanced Level	4.6	computers, waves, simulations: a practical introduction to numerical methods using python	Mixed Level (Beginner Level)	4.7
end-to-end machine learning with tensorflow on gcp	Advanced Level	4.5	leveraging unstructured data with cloud dataproc on google cloud platform	Intermediate Level	4.5
serverless data analysis with google bigquery and cloud dataflow	Intermediate Level	4.5	production machine learning systems	Advanced Level	4.5
art and science of machine learning	Intermediate Level	4.6	how google does machine learning	Intermediate Level	4.6
production machine learning systems	Advanced Level	4.5	art and science of machine learning	Intermediate Level	4.6
applying machine learning to your data with gcp	Intermediate Level	4.6	recommendation systems with tensorflow on gcp	Advanced Level	4.4
sequence models for time series and natural language processing	Advanced Level	4.5	image understanding with tensorflow on gcp	Advanced Level	4.6
recommendation systems with tensorflow on gcp	Advanced Level	4.4	launching into machine learning	Intermediate Level	4.6
build a modern computer from first principles: nand to tetris part ii (project-centered course)	Mixed Level (Intermediate Level)	5	end-to-end machine learning with tensorflow on gcp	Advanced Level	4.5
communicating data science results	Mixed Level (Beginner Level)	3.6	getting started with application development	Intermediate Level	4.5
e-learning ecologies: innovative approaches to teaching and learning for the digital age	Mixed Level (Beginner Level)	4.5	genomic data science capstone	Mixed Level (Beginner Level)	4.6
			ibm cloud: deploying microservices with kubernetes	Mixed Level (Intermediate Level)	4.6
			sequence models for time series and natural language processing	Advanced Level	4.5
			genomic data science with galaxy	Mixed Level (Beginner Level)	3.7
			statistics for genomic data science	Mixed Level (Beginner Level)	4.1
			grow to greatness: smart growth for private businesses, part i	Mixed Level (Beginner Level)	4.7

As both models were providing recommended courses as per input, it was tough to evaluate which one is better so, we decided to take existing user history in order to compare the recommended outcome of both models.

We took few courses from the existing user which they have already completed and input in our models to compare the output. We observed that many courses which user has already completed, present in our both models’ outcome. For example, User 1 has completed two courses 1) Python data structures and 2) Python functions, files and dictionaries, we wanted to cross check whether if we input one course from their

history, another course is recommended by our models or not.

Our observation was most of the cases, both models recommended the courses but, in few cases, “Word Vector & Document Vector” recommended the course from user history but cosine similarity was unable to recommend.

Though, it is very tough to say that one recommendation system is better than others because there is no proper scale to evaluate the recommendation system. But from our two models, we can see that in one case, “Word Vector & Document Vector” was able to recommend the courses from user history but cosine similarity could not. **We can say probably integration of both systems would be better**

User1

User1		
Actual	**Cosine Predict**	**Word Predict**
python data structure	python data structure	python data structure
python functions, files, and dictionaries	python data representations	using python to access web data
	python programming essentials	introduction to data science in python
	using python to access web data	algorithms, part ii
	python for data science	python basics
	python functions, files, and dictionaries	algorithms, part i
	programming for everybody (getting started with python)	algorithms for dna sequencing
	python data analysis	python functions, files, and dictionaries
	data collection and processing with python	pointers, arrays, and recursion
	introduction to data science in python	calculus: single variable part 1 - functions
	python classes and inheritance	calculus: single variable part 4 - applications
	using databases with python	java programming: arrays, lists, and structured data
	data structures	calculus: single variable part 3 - integration
	data visualization with python	learn to program: the fundamentals
	applied plotting, charting & data representation in python	distributed database systems
	python and statistics for financial analysis	building database applications in php

User2

User2		
Actual	**Cosine Predict**	**Word Predict**
Python Basics	Python Basics	Python Basics
Using python to access web data	python data representations	python functions, files, and dictionaries
	python programming essentials	python programming essentials
	using python to access web data	python data representations
	python for data science	python for data science
	python functions, files, and dictionaries	python classes and inheritance
	programming for everybody (getting started with python)	python data analysis
	python data analysis	data collection and processing with python
	data collection and processing with python	programming for everybody (getting started with python)
	introduction to data science in python	using python to access web data
	python classes and inheritance	introduction to data science in python
	using databases with python	python data structures
	data structures	build a modern computer from first principles: nand to tetris part ii (project-centered course)
	data visualization with python	object oriented programming in java
	applied plotting, charting & data representation in python	be persuasive: write a convincing position paper or policy advice (project-centered course)
	python and statistics for financial analysis	principles of computing (part 1)

User3

User3		
Actual	**Cosine Predict**	**Word Predict**
Machine learning, feature engineering	Machine learning, feature engineering	Machine learning, feature engineering
bayesian methods for machine learning	machine learning foundations: a case study approach	matrix factorization and advanced techniques
	machine learning with python	big data applications: machine learning at scale
	how to win a data science competition: learn from top kagglers	introduction to machine learning
	launching into machine learning	convolutional neural networks
	how google does machine learning	probabilistic graphical models 1: representation
	machine learning for data analysis	deep learning in computer vision
	internet giants: the law and economics of media platforms	practical reinforcement learning
	be persuasive: write a convincing position paper or policy advice (project-centered course)	probabilistic graphical models 3: learning
	big data applications: machine learning at scale	statistical mechanics: algorithms and computations
	advanced machine learning and signal processing	sequence models
	fundamentals of machine learning in finance	practical machine learning on h2o
	guided tour of machine learning in finance	algorithms on graphs
	feature engineering	bayesian methods for machine learning
	probabilistic graphical models 3: learning	visual perception for self-driving cars
	image understanding with tensorflow on gcp	computers, waves, simulations: a practical introduction to numerical methods using python
	end-to-end machine learning with tensorflow on gcp	leveraging unstructured data with cloud dataproc on google cloud platform
	serverless data analysis with google bigquery and cloud dataflow	production machine learning systems
	art and science of machine learning	how google does machine learning
	production machine learning systems	art and science of machine learning
	applying machine learning to your data with gcp	recommendation systems with tensorflow on gcp
	sequence models for time series and natural language processing	image understanding with tensorflow on gcp
	recommendation systems with tensorflow on gcp	launching into machine learning
	build a modern computer from first principles: nand to tetris part ii (project-centered course)	end-to-end machine learning with tensorflow on gcp
	communicating data science results	getting started with application development
	e-learning ecologies: innovative approaches to teaching and learning for the digital age	genomic data science capstone
		ibm cloud: deploying microservices with kubernetes
		sequence models for time series and natural language processing
		genomic data science with galaxy
		statistics for genomic data science
		grow to greatness: smart growth for private businesses, part i

8. Analysis of Experiment results

When we used clustering model for getting the original level of mixed level, it did not work. Even In classification, multinomial naïve Bayes didn't work. It is because of less amount of data we had for training the model. IF we have, large amount of data, our models could be trained in better way and there would not have any bias in our model. We think that the integration of both models would give better output compare to single. In future, we will think of integrating both model for better outcome.

9. Business insight

Most course recommenders provide courses which are close to the user search, while our recommender provides courses not only based on similarity but also the courses difficulty and course diversification. It can attract users who are not satisfied with the current course recommenders.

Following the same data scraping process and algorithms, we can build a recommender beyond recommenders, which provide the course search services not only to Coursera but Udemy, Mooc, Udacity, etc. Users do not need to go the official websites to search courses any longer.

We pay attention to the niche market of online education and implement differentiation focus strategy to the course recommender.

10. Further work

In the future, we intend to improve the recommender from 4 aspects:

1) Scrape more high-quality features, especially user behavior related features.
2) Scrape more data to for training a more accurate recommender.
3) Combine courses from different online education websites together.
4) Integrate two-course machine learning models together.

References:

1. https://www.coursera.org/