COMPARISON OF CATEGORICAL PROPERTIES OFFERED BY MULTIPLE MOOC PLATFORMS

Using a Web Crawler in Python with Scrapy

Bachelor Thesis - Final Presentation

Louis Mbuyu
Aufgabensteller: **Prof. Dr. François Bry**
Betreuer: **Prof. Dr. François Bry, Yingding Wang**

12.04.18

AGENDA

1. Introduction / Goal
2. Defining MOOC model
3. Web Scraper / Results
4. Gold standard selection
5. Text categorisation approach
6. Gold standard evaluation
7. Evaluation of all platforms
8. Conclusion & Future work

1. Introduction / Goal

Motivation

- Irom - **I**ntelligent **R**ecommender **O**f **M**OOCs
- MOOC - **M**assive **O**pen **O**nline **C**ourse

The goal of Irom

- To improve the learning and studying at the university.
- To develop an intelligent MOOCs search engine

Goal of thesis

- Define unified set of categories across all MOOC platforms.

Motivation

Udemy

Categories | Search for Courses

- Development
- Business
- IT & Software
- Office Productivity
- Personal Development
- Design
- Marketing
- Lifestyle
- Photography
- Health & Fitness
- Teacher Training
- Music
- Academics
- Language
- Test Prep

coursera

Catalog | Search catalog

CATEGORIES

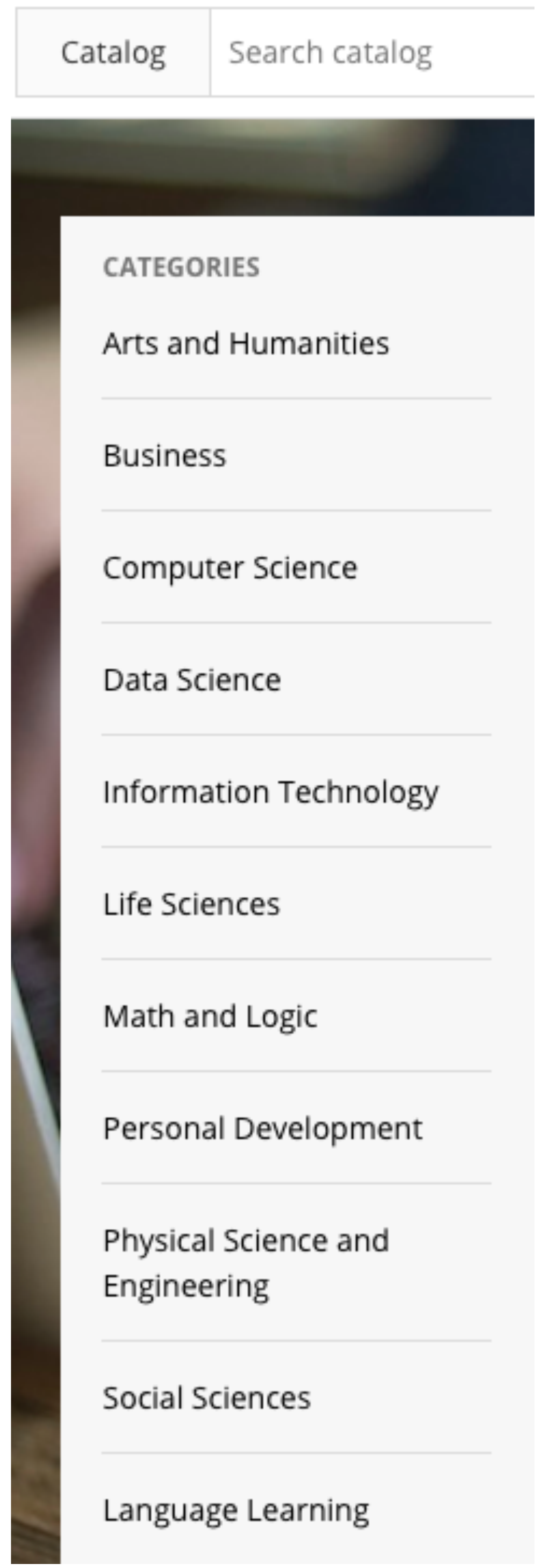
- Arts and Humanities
- Business
- Computer Science
- Data Science
- Information Technology
- Life Sciences
- Math and Logic
- Personal Development
- Physical Science and Engineering
- Social Sciences
- Language Learning

edX

Architecture	22
Art & Culture	116
Biology & Life Sciences	140
Business & Management	361
Chemistry	42
Communication	79
Computer Science	522
Data Analysis & Statistics	189
Design	47
Economics & Finance	180
Education & Teacher Training	94
Electronics	55
Energy & Earth Sciences	23
Engineering	273
Environmental Studies	82
Ethics	16
Food & Nutrition	22
Health & Safety	86
History	120
Humanities	231
Language	29
Law	45
Literature	75
Math	115
Medicine	79
Music	27
Philanthropy	1
Philosophy & Ethics	68
Physics	135
Science	180
Social Sciences	251

Modified Goal

coursera



Tasks

1. Define a MOOC model
2. Build a Web Scraper and extract data
3. Select a platform as the “Gold standard”
4. Text categorisation approach (TF-IDF & cos sim.)
5. Evaluate tf-idf and cosine similarity approach
6. Categorise courses from other platforms
7. Evaluate the results

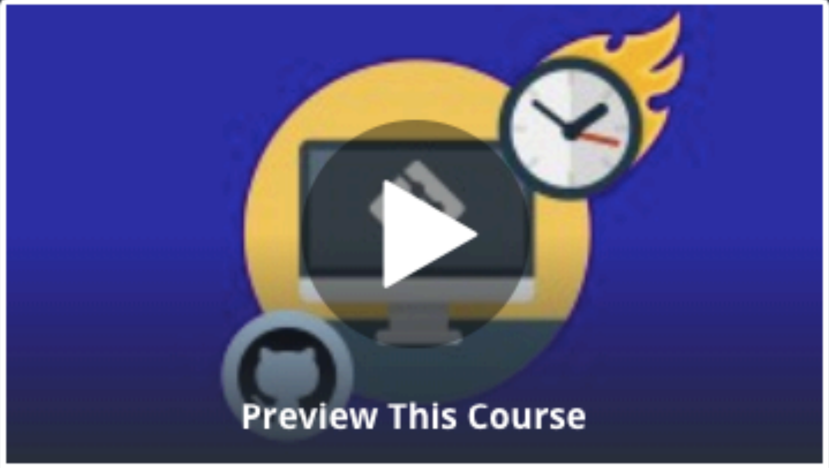
2. Defining MOOC model

Git Started with GitHub

Learn the basics of Git and GitHub, with clear, step-by-step instructions by a highly rated teacher.

4.4 (6,932 ratings) 82,309 students enrolled

Created by Jason Taylor Last updated 8/2017 English English



Free

Enroll Now

Includes:

- 1 hour on-demand video
- 1 Article
- 2 Supplemental Resources
- Full lifetime access
- Access on mobile and TV
- Certificate of Completion

What Will I Learn?

- Save project changes into source control using Git
- Create a new repository on GitHub
- Copy a repository from GitHub onto a local computer for using with a project
- Install and configure Git on Windows and Mac computer systems

Requirements

- Basic computer skills
- Ability to install software, which may require admin rights

Description

This course is designed to jump right into showing how Git and GitHub work together, focusing on the Git basic workflow. Students can expect to learn the minimum needed to start using Git in about 30 minutes.

Recent Updates:

- *October 16th*: Added *Updates and Errata* Section

Course Outline

UDACITY

Explore | Nanodegree | Catalog | For Business | Blog | My Classroom

FREE COURSE

Intro to Machine Learning

Pattern Recognition for Fun and Profit

START FREE COURSE

NANODEGREE PROGRAM

Machine Learning Engineer by kaggle

Make Predictive Models

★★★★★

Accelerate your career with the credential that fast-tracks you to job success. »

About this Course

Machine Learning is a first-class ticket to the most exciting careers in data analysis today. As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions.

Machine learning brings together computer science and statistics to harness that predictive power. It's a must-have skill for all aspiring data analysts and data scientists, or anyone else who wants to wrestle all that raw data into refined trends and

COURSE COST

Free

TIMELINE

Approx. 10 weeks

SKILL LEVEL

Intermediate

INCLUDED IN COURSE

Rich Learning Content

Interactive Quizzes

Motivation - MOOC platforms

coursera

FutureLearn

OPEN 2 STUDY

UDACITY

Udemy

edX

Unified MOOC Model (Table)

	Coursera	Udacity	Edx	FutureLearn	Open2Study	Udemy
Url	✓	✓	✓	✓	✓	✓
Title	✓	✓	✓	✓	✓	✓
Summary	✓	✓	✓	✓	✓	✓
Description	✓	✓	✓	✓	✓	✓
Subcategory	✓	✗	✗	✗	✓	✓
Category	✓	✓	✓	✓	✓	✓
WhyTakeThis Course	✓	✓	✓	✓	✓	✓
Provider	✓	✓	✓	✓	✓	✓
Level	✓	✓	✓	✗	✗	✓
ImageUrl	✓	✓	✓	✓	✓	✓
Price	✓	✓	✓	✓	✓	✓
Duration	✓	✓	✓	✓	✓	✓
RatingValue	✓	✗	✓	✓	✓	✓
RatingAmount	✓	✗	✓	✓	✓	✓
StartDate	✓	✗	✓	✓	✓	✗
EndDate	✗	✗	✗	✓	✓	✗

3. Web Scraper / Results

Scraped Data (Table)

	Coursera	Udacity	Edx	FutureLearn	Open2Study	Udemy	All
Total number of courses	3.032	232	1.098	193	49	40.003	44.607

4. Gold standard selection

Gold standard criteria

1. Number of categories

2. Number of courses

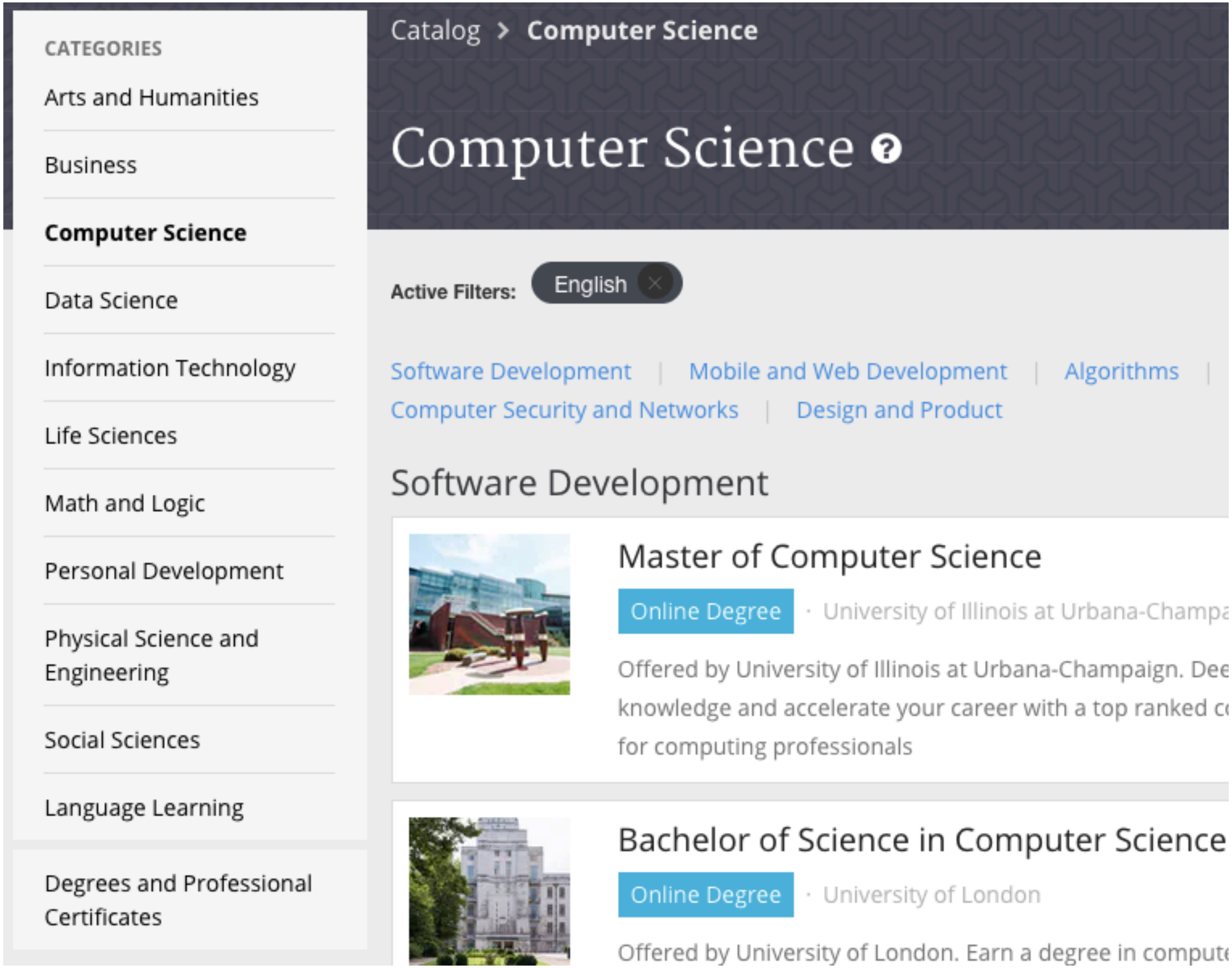
3. Diversity

4. Represent University Subjects

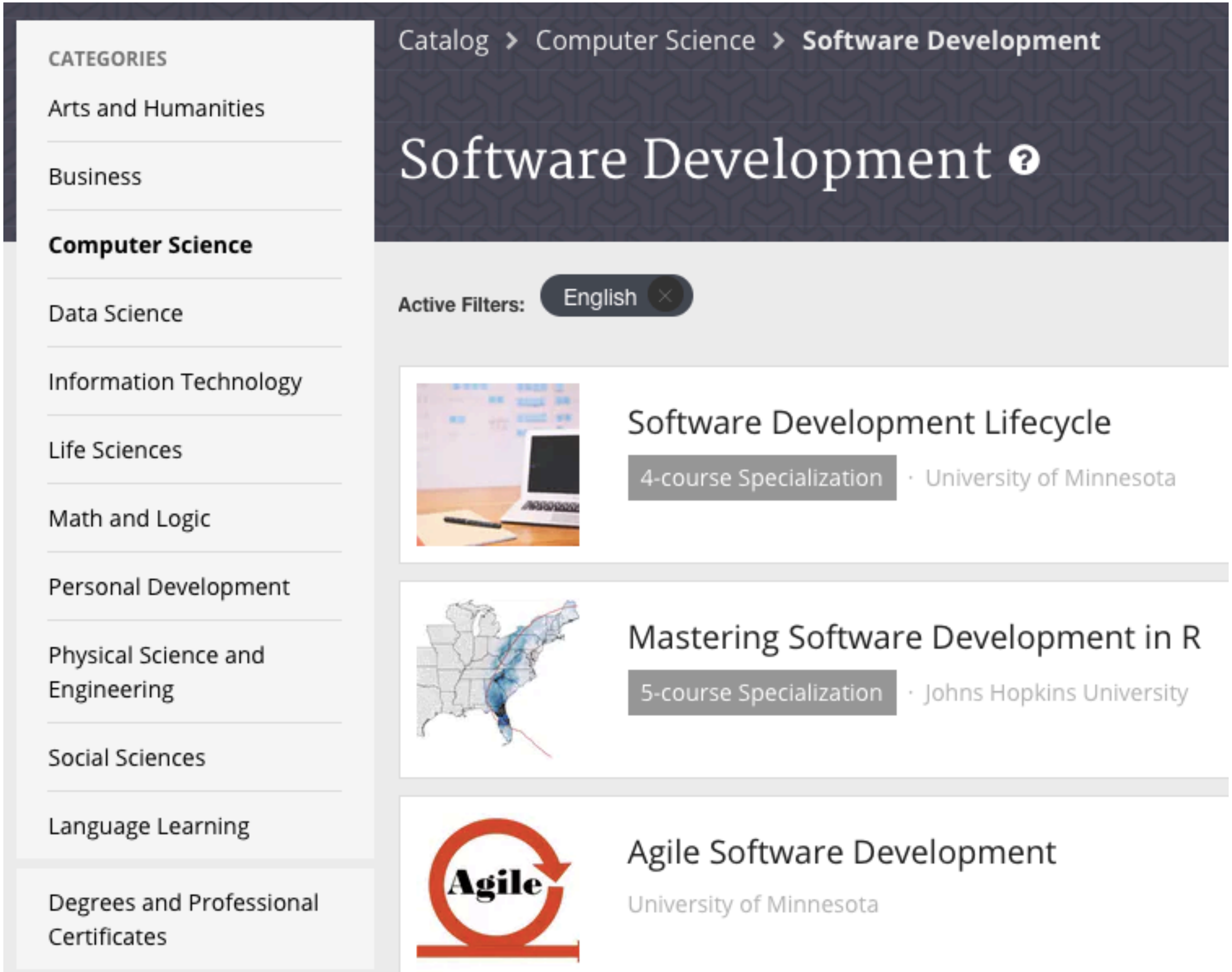
Gold standard elimination process

	Coursera	Udacity	Edx	FutureLearn	Open2Study	Udemy
No. of categories	✓	✓	✗	✓	✗	✓
No. of courses	✓	✗	✓	✗	✗	✓
Diversity	✓	✗	✓	✓	✓	✓
University rep.	✓	✗	✓	✓	✓	✗

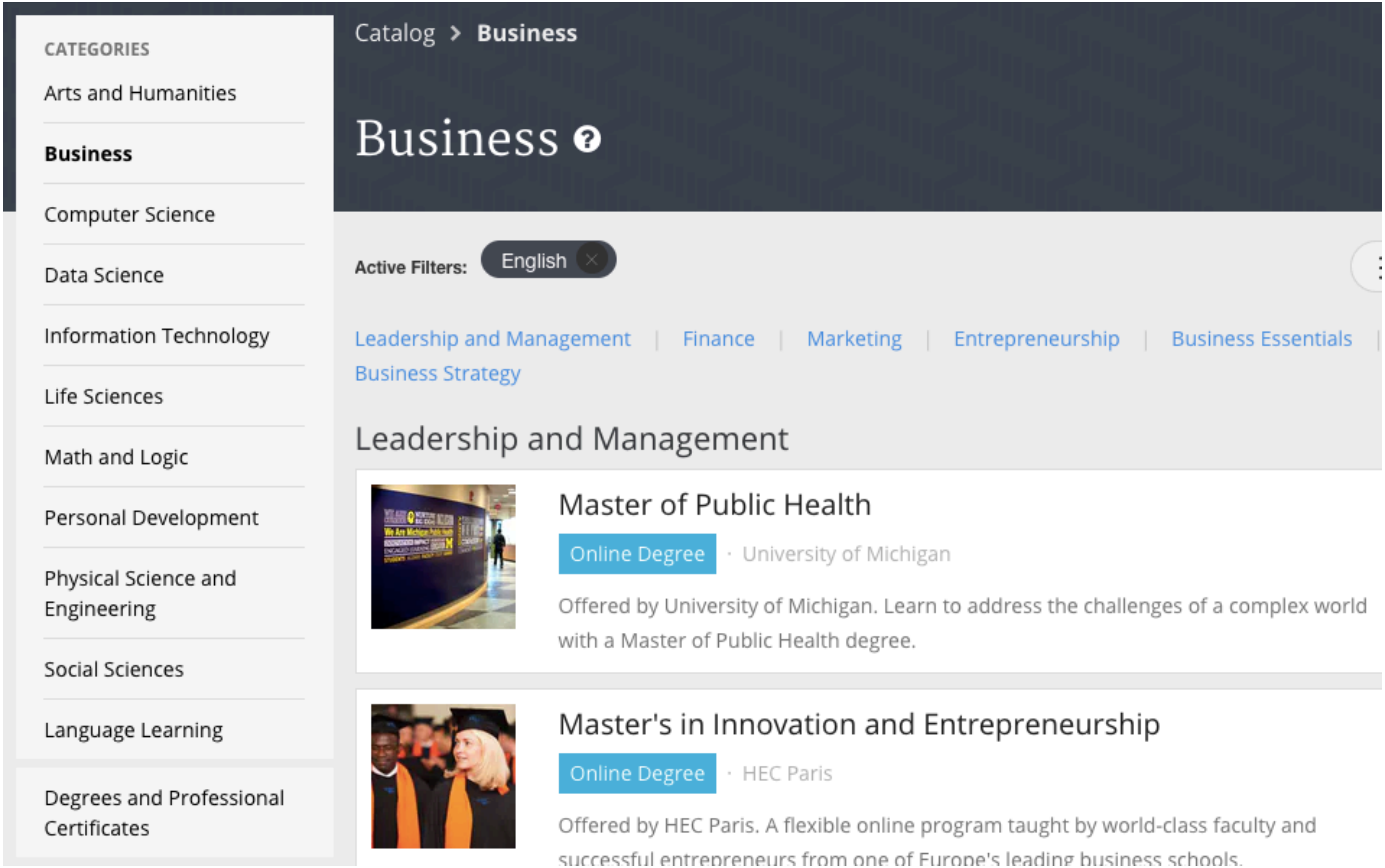
Gold standard structure



Gold standard structure



Gold standard structure



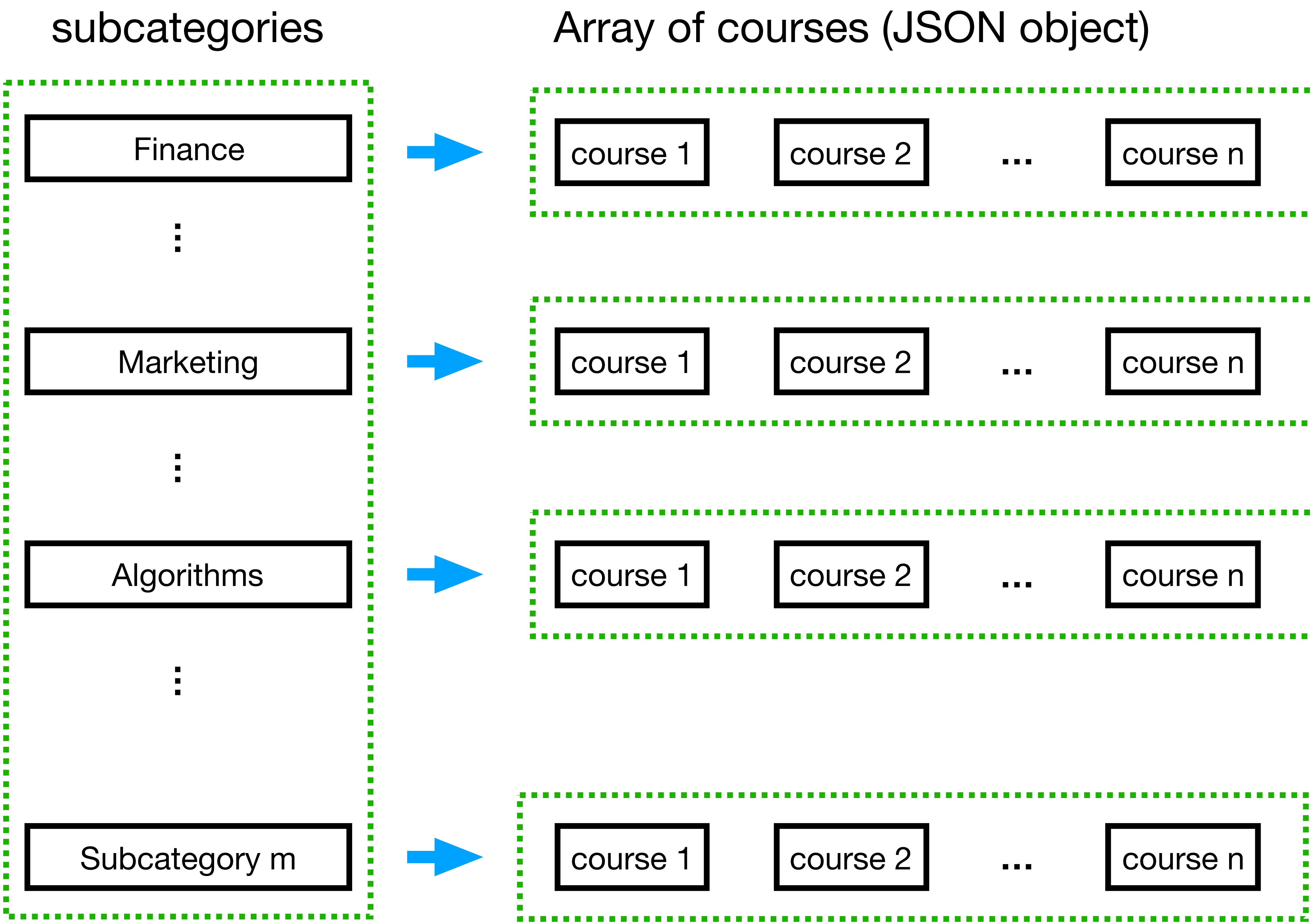
5. Text categorisation approach

Text categorisation approach (Step 1)

Query Database : ‘Platform’ = ‘Coursera’ AND GROUB BY ‘Subcategory’

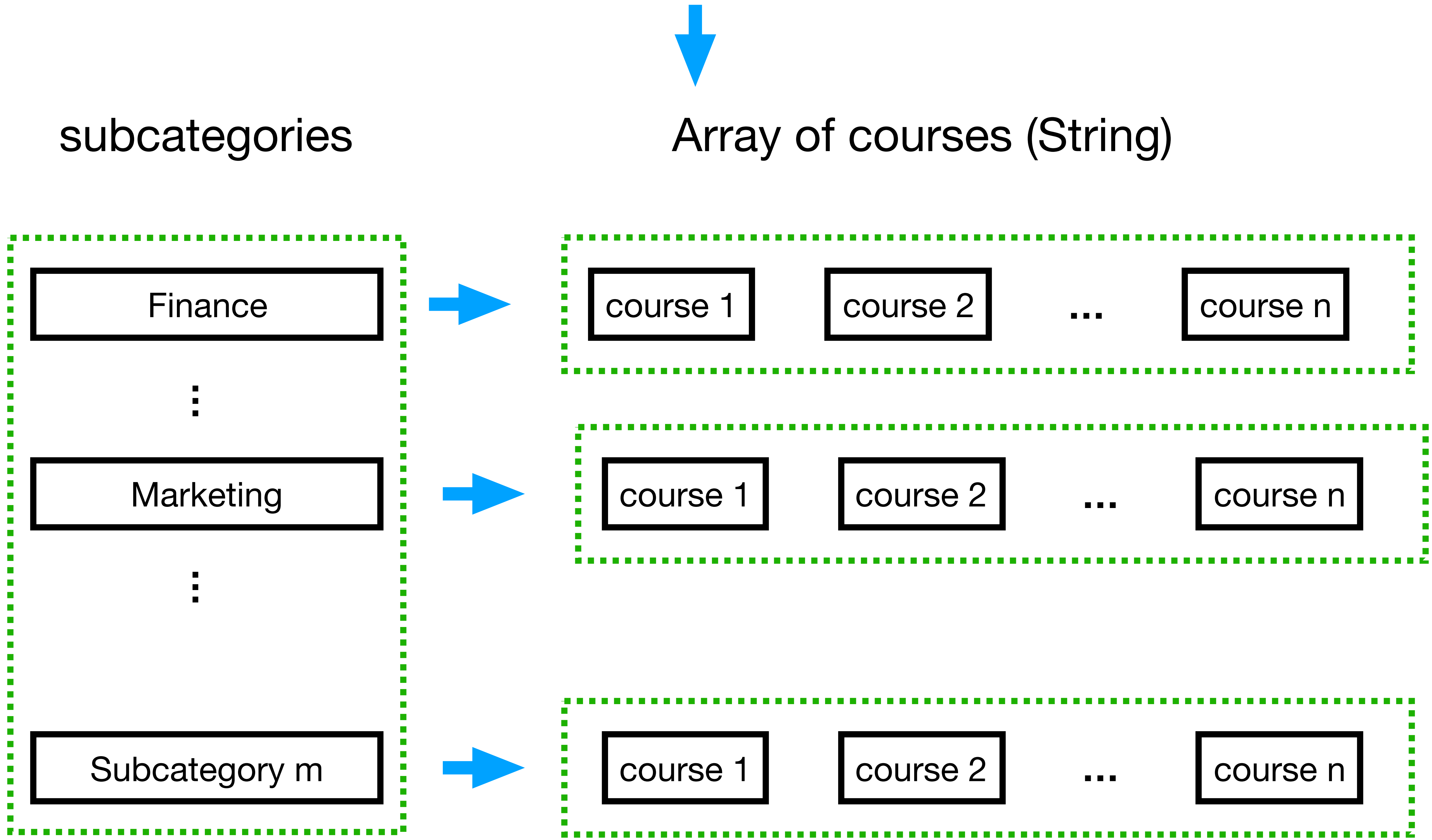
MongoDB

Text categorisation approach (Step 2)



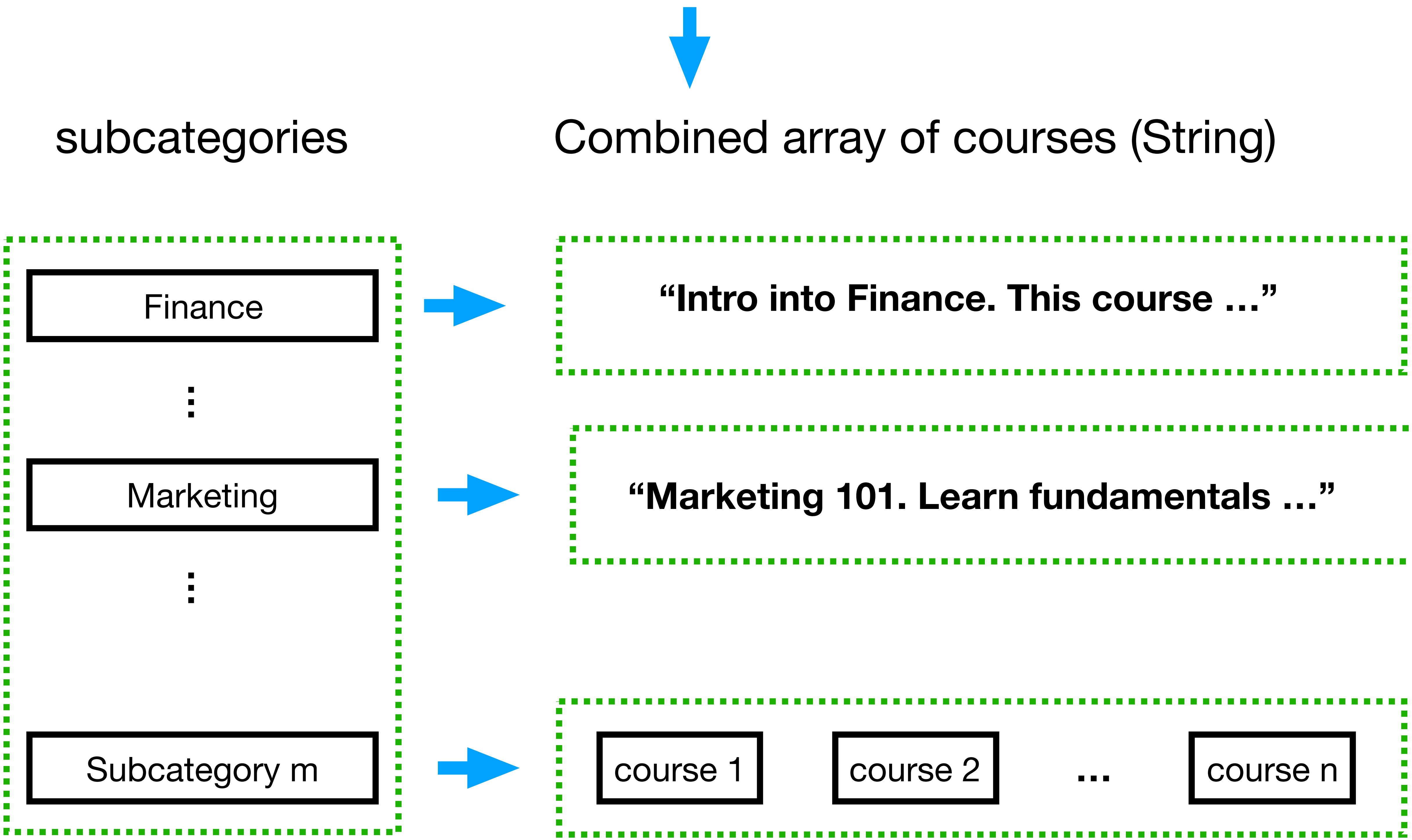
Text categorisation approach (Step 3)

Iterate through courses and extract and combine the ‘title’, ‘Summary’, ‘Description’



Text categorisation approach (Step 4)

Join all arrays/list of strings into one string

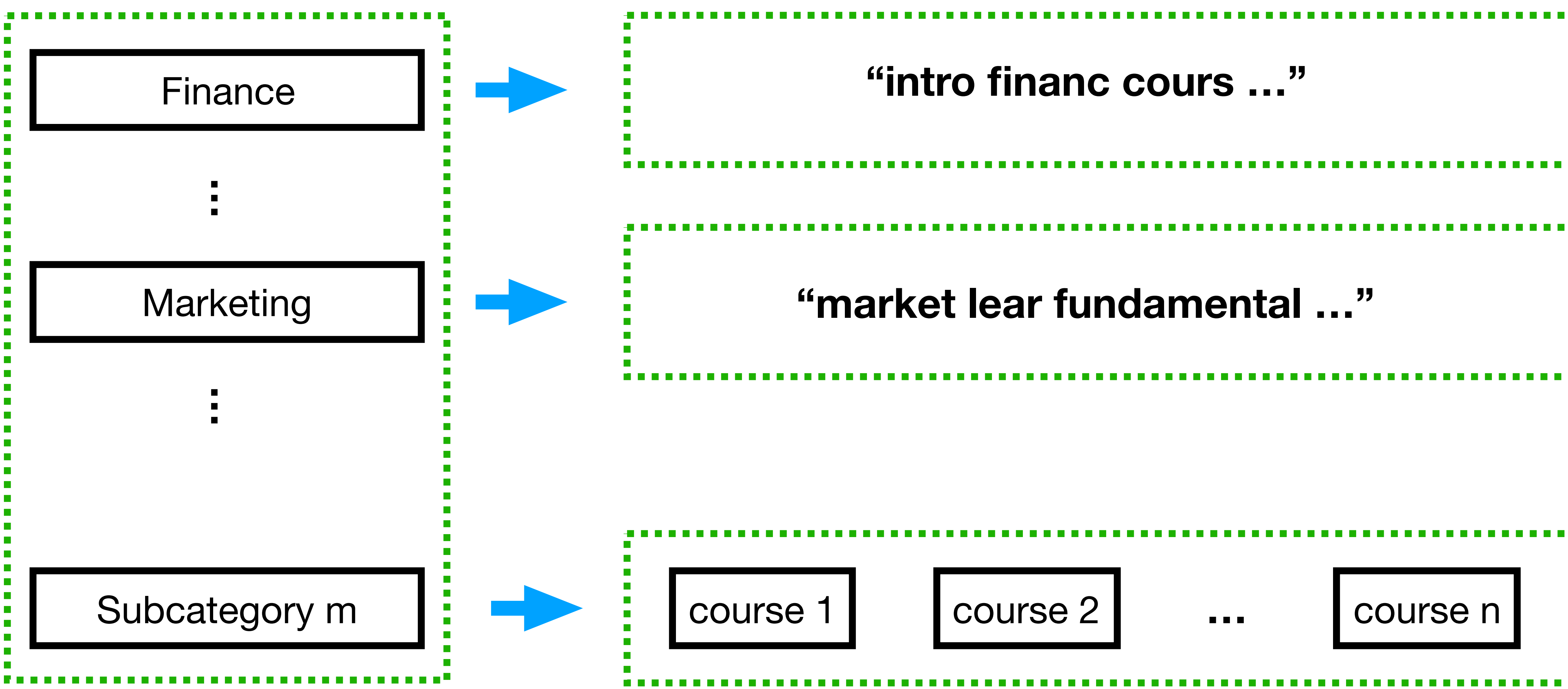


Text categorisation approach (Step 5)

Preprocess Data: Remove all stop words and punctuations, All words to lowercase, All words are stemmed

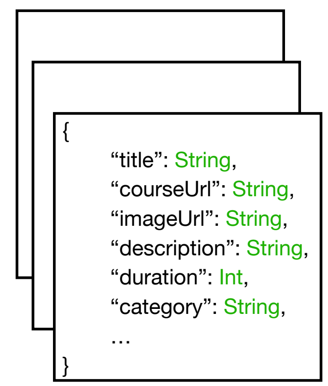
subcategories

Preprocessed combined array of courses



Text categorisation approach (Step 6)

Course (Query) from another platform, that needs to be categorised

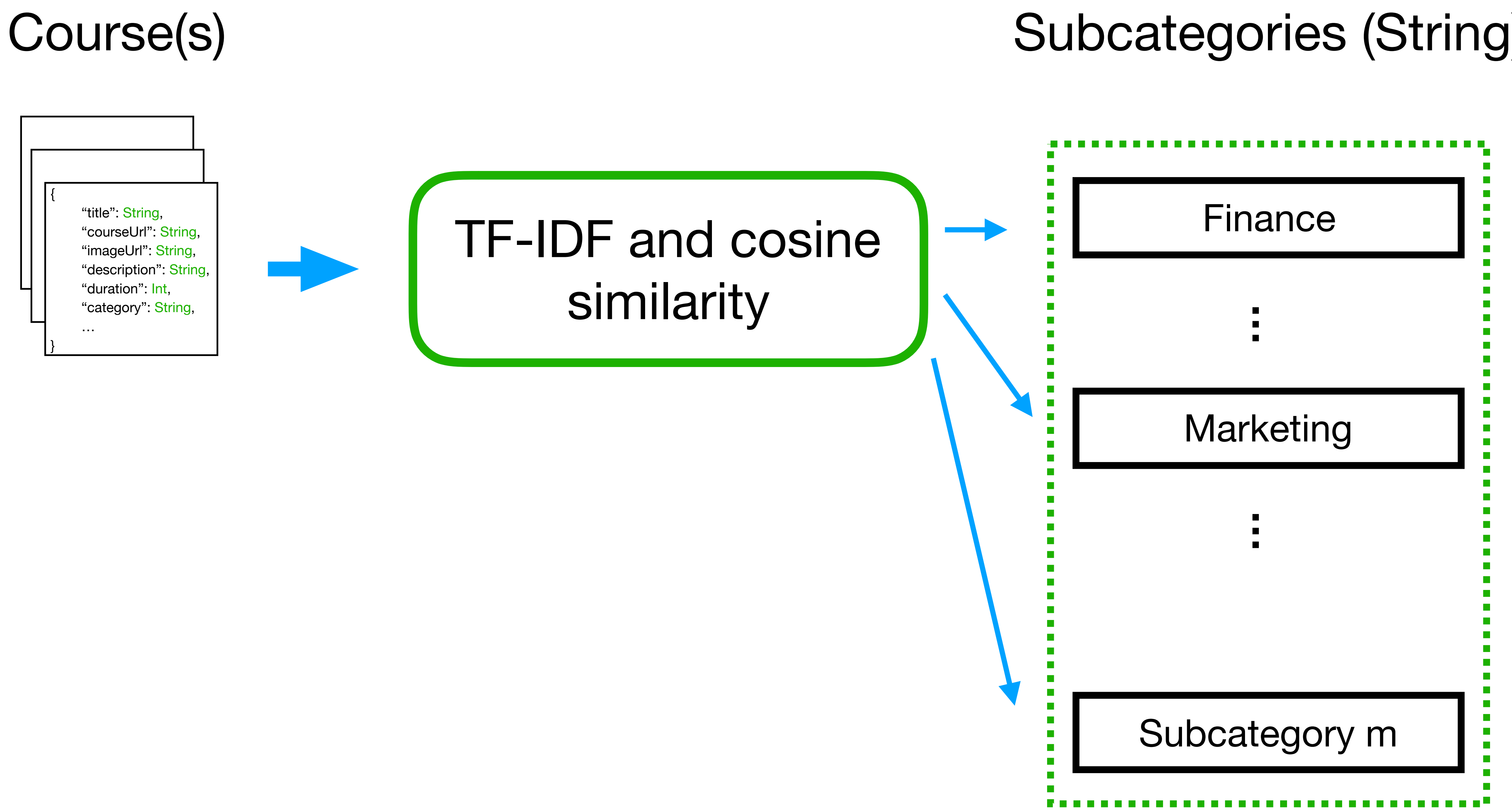


Extract and combine the 'title', 'Summary', 'Description'

Preprocess Data: Remove all stop words and punctuations, All words to lowercase, All words are stemmed

Text categorisation approach (Step 6)

Calculate **TF-IDF** and **Cosine similarity** for all subcategories. Course is categorised to the subcategory with the highest value.



TF-IDF & Cosine Similarity

TF-IDF - Term frequency inverse document frequency

Term Frequency - How frequent a term appears in a given document

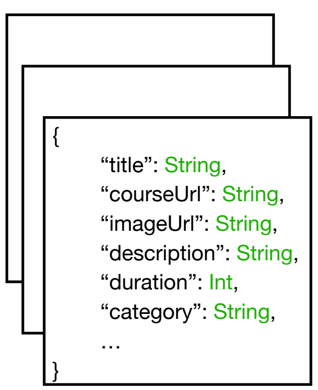
Inverse document frequency - diminishes the weight of terms that appear very frequently in the corpus and increases the weight of terms that appear rarely.

Cosine similarity - a measure of similarity between two vectors, that measures the cosine of the angle between them.

6. Approach evaluation

Approach evaluation

Coursera courses



TF-IDF and cosine similarity approach

New Category

Approach evaluation

Accuracy - Accuracy is a ratio of total correctly categorised courses to the total number of courses

$$\text{Gold standard accuracy} = \frac{2625}{3032} \approx \mathbf{0.87}$$

7. Evaluation of all platforms

Evaluation of all platforms (Udacity)

Intro. to Android

Category: Android

TF-IDF and cosine similarity approach

Computer Science Category

Good or bad outcome?

Evaluation of all platforms (Udacity)

Gold standard (Coursera) categories

Udacity categories

	B	CS	DS	SS	PSE	LS	AH	LL	IT	ML	PD	DPC
Android	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
iOS	0.0%	95.45%	4.55%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Digital Marketing	90.0%	0.0%	10.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Web Development	0.0%	85.0%	5.0%	0.0%	0.0%	0.0%	0.0%	0.0%	10.0%	0.0%	0.0%	0.0%
Artificial Intelligence	0.0%	71.43%	28.57%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Data Science	3.85%	15.38%	80.77%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Data Analytics	0.0%	6.67%	93.33%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Software Engineering	0.0%	77.27%	13.64%	0.0%	0.0%	0.0%	0.0%	0.0%	9.09%	0.0%	0.0%	0.0%
Deep Learning	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Non-Tech	25.0%	31.25%	12.5%	6.25%	6.25%	6.25%	0.0%	0.0%	0.0%	12.5%	0.0%	0.0%
Virtual Reality	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Machine Learning	0.0%	10.0%	90.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Mobile App Development	0.0%	92.86%	7.14%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Self Driving Car	0.0%	77.78%	0.0%	0.0%	22.22%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Developer Essentials	25.0%	62.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	12.5%	0.0%	0.0%	0.0%

Evaluation of all platforms (Udacity)

	B	CS	DS	SS	PSE	LS	AH	LL	IT	ML	PD	DPC
Android	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
iOS	0.0%	95.45%	4.55%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Digital Marketing	90.0%	0.0%	10.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Web Development	0.0%	85.0%	5.0%	0.0%	0.0%	0.0%	0.0%	0.0%	10.0%	0.0%	0.0%	0.0%
Artificial Intelligence	0.0%	71.43%	28.57%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Data Science	3.85%	15.38%	80.77%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Data Analytics	0.0%	6.67%	93.33%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Software Engineering	0.0%	77.27%	13.64%	0.0%	0.0%	0.0%	0.0%	0.0%	9.09%	0.0%	0.0%	0.0%
Deep Learning	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Non-Tech	25.0%	31.25%	12.5%	6.25%	6.25%	6.25%	0.0%	0.0%	0.0%	12.5%	0.0%	0.0%
Virtual Reality	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Machine Learning	0.0%	10.0%	90.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Mobile App Development	0.0%	92.86%	7.14%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Self Driving Car	0.0%	77.78%	0.0%	0.0%	22.22%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Developer Essentials	25.0%	62.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	12.5%	0.0%	0.0%	0.0%

*The heat-map shows the percentages of courses categorised to that particular category, with darker colours indicating greater percentage.

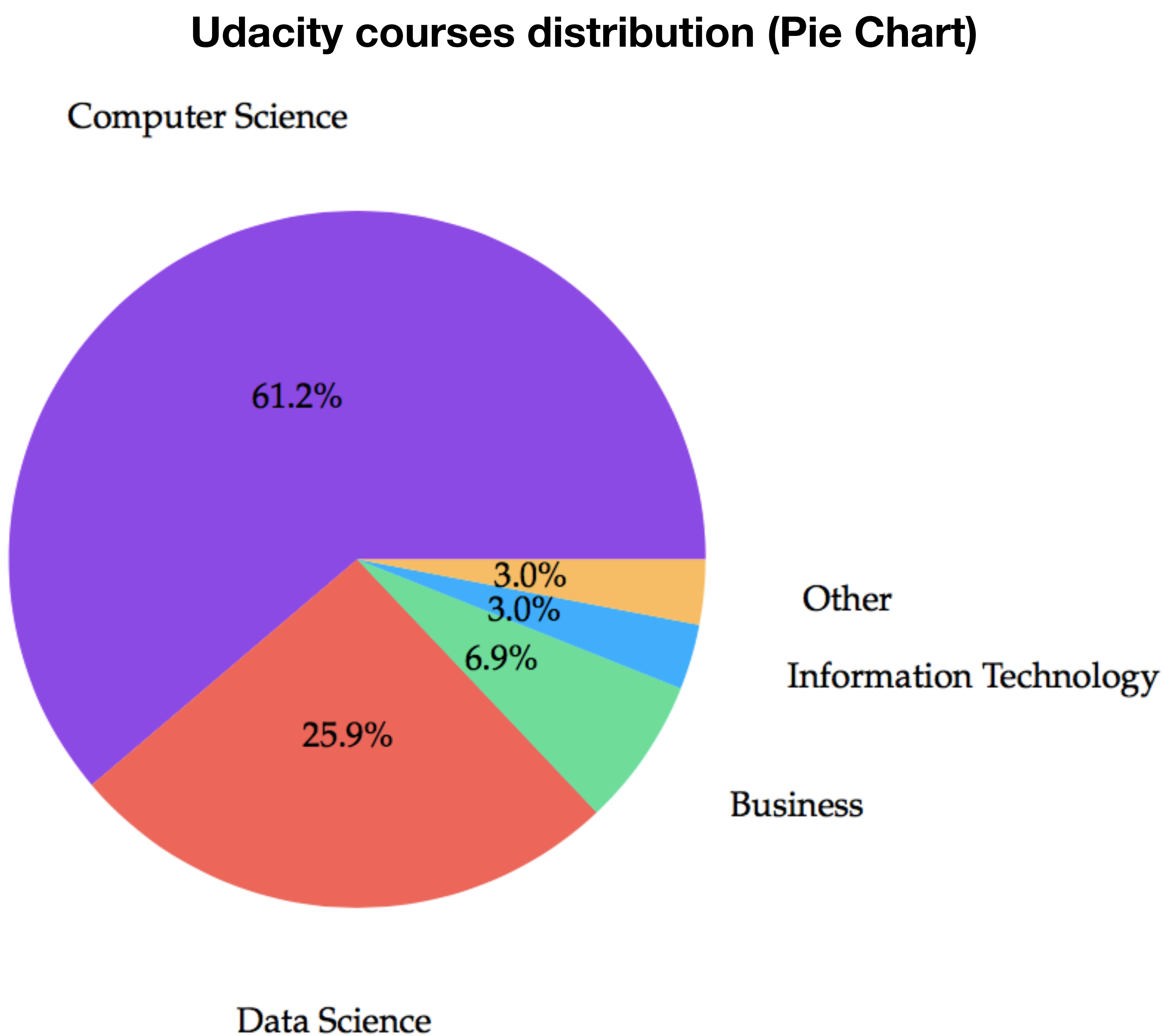
Grading schema

Evaluation of all platforms (Udacity)

Udacity evaluation table

	B	CS	DS	SS	PSE	LS	AH	LL	IT	ML	PD	DPC	Rating
Android	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
iOS	0.0%	95.45%	4.55%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Digital Marketing	90.0%	0.0%	10.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Web Development	0.0%	85.0%	5.0%	0.0%	0.0%	0.0%	0.0%	0.0%	10.0%	0.0%	0.0%	0.0%	1
Artificial Intelligence	0.0%	71.43%	28.57%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3
Data Science	3.85%	15.38%	80.77%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Data Analytics	0.0%	6.67%	93.33%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Software Engineering	0.0%	77.27%	13.64%	0.0%	0.0%	0.0%	0.0%	0.0%	9.09%	0.0%	0.0%	0.0%	1
Deep Learning	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Non-Tech	25.0%	31.25%	12.5%	6.25%	6.25%	6.25%	0.0%	0.0%	0.0%	12.5%	0.0%	0.0%	5
Virtual Reality	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2
Machine Learning	0.0%	10.0%	90.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Mobile App Development	0.0%	92.86%	7.14%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Self Driving Car	0.0%	77.78%	0.0%	0.0%	22.22%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2
Developer Essentials	25.0%	62.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	12.5%	0.0%	0.0%	0.0%	1
Overall rating	-	-	-	-	-	-	-	-	-	-	-	-	1.53

Evaluation of all platforms (Udacity)



Evaluation of all platforms (Udacity)

Udacity courses distribution (Table)

	Total no. of courses per category	Percentage of Total no. of all courses
Degrees & Professional Certificates	0	0.0%
Language Learning	0	0.0%
Arts & Humanities	0	0.0%
Personal Development	0	0.0%
Social Science	1	0.43%
Life Sciences	1	0.43%
Math & Logic	2	0.86%
Physical Science and Engineering	3	1.29%
Information Technology	7	3.02%
Business	16	6.9%
Data Science	60	25.86%
Computer Science	142	61.21%
Total	232	100%

Evaluation of all platforms (Edx)

	B	CS	DS	SS	PSE	LS	AH	LL	IT	ML	PD	DPC	Rating
Computer Science	6.88%	49.57%	18.05%	0.29%	4.58%	0.86%	0.0%	0.0%	19.48%	0.29%	0.0%	0.0%	2
Business & Management	77.87%	0.4%	9.49%	1.98%	2.37%	0.0%	0.4%	5.93%	1.58%	0.0%	0.0%	0.0%	1
Biology & Life Sciences	1.82%	1.82%	12.73%	0.0%	7.27%	76.36%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Humanities	13.04%	0.0%	0.0%	21.74%	0.0%	4.35%	39.13%	17.39%	0.0%	0.0%	4.35%	0.0%	1
Economics & Finance	88.46%	0.0%	1.92%	5.77%	0.0%	0.0%	0.0%	0.0%	3.85%	0.0%	0.0%	0.0%	1
Engineering	4.76%	4.76%	9.52%	4.76%	66.67%	2.38%	4.76%	0.0%	0.0%	2.38%	0.0%	0.0%	1
Physics	0.0%	0.0%	7.41%	0.0%	85.19%	7.41%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Social Sciences	21.05%	0.0%	0.0%	65.79%	2.63%	2.63%	7.89%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Education & Teacher Training	5.88%	8.82%	0.0%	76.47%	0.0%	0.0%	0.0%	2.94%	0.0%	0.0%	5.88%	0.0%	1
Math	0.0%	9.09%	9.09%	9.09%	0.0%	0.0%	0.0%	0.0%	0.0%	72.73%	0.0%	0.0%	1
Science	0.0%	0.0%	7.14%	50.0%	35.71%	7.14%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Language	0.0%	0.0%	0.0%	25.0%	0.0%	0.0%	0.0%	75.0%	0.0%	0.0%	0.0%	0.0%	1
Data Analysis & Statistics	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Architecture	0.0%	11.11%	0.0%	77.78%	0.0%	0.0%	11.11%	0.0%	0.0%	0.0%	0.0%	0.0%	1
History	0.0%	0.0%	0.0%	13.33%	0.0%	0.0%	80.0%	6.67%	0.0%	0.0%	0.0%	0.0%	1
Energy & Earth Sciences	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Literature	0.0%	0.0%	0.0%	50.0%	0.0%	0.0%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Medicine	0.0%	0.0%	0.0%	22.22%	11.11%	66.67%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Health & Safety	0.0%	0.0%	0.0%	50.0%	0.0%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Art & Culture	10.0%	3.33%	0.0%	3.33%	0.0%	0.0%	66.67%	16.67%	0.0%	0.0%	0.0%	0.0%	1
Music	0.0%	0.0%	0.0%	14.29%	0.0%	0.0%	85.71%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Chemistry	12.5%	25.0%	0.0%	12.5%	25.0%	25.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5
Philosophy & Ethics	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Communication	6.67%	0.0%	0.0%	20.0%	6.67%	0.0%	0.0%	60.0%	0.0%	0.0%	6.67%	0.0%	1
Environmental Studies	9.09%	0.0%	0.0%	0.0%	90.91%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Ethics	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2
Electronics	0.0%	7.14%	0.0%	0.0%	92.86%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Design	9.09%	72.73%	0.0%	9.09%	0.0%	9.09%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2
Law	20.0%	0.0%	0.0%	80.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2
Food & Nutrition	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1
Overall rating	-	-	-	-	-	-	-	-	-	-	-	-	1.26

8. Conclusion & Future work

Conclusion

1. ca 45.000 courses scraped and indexed for IROM.
2. Coursera's categories as the gold standard was a great outcome.
3. Tf-idf and cosine similarity measure was also a positive outcome.

Future work

1. Measure the quality of data scraped

2. Better approach - machine learning (neural networks, etc)

3. Evaluating text categorisation

Thank you.