COMPARISON OF CATEGORICAL PROPERTIES OFFERED BY MULTIPLE MOOC PLATFORMS

Using automated Web Crawler in Python with Scrapy

Bachelor Thesis - Introduction Presentation

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

25.01.18

AGENDA

1. Motivation
2. Research topic
3. Project Plan
4. Technical Details
5. Challenges
6. Demo

1. Motivation

Motivation

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

The goal of Irom

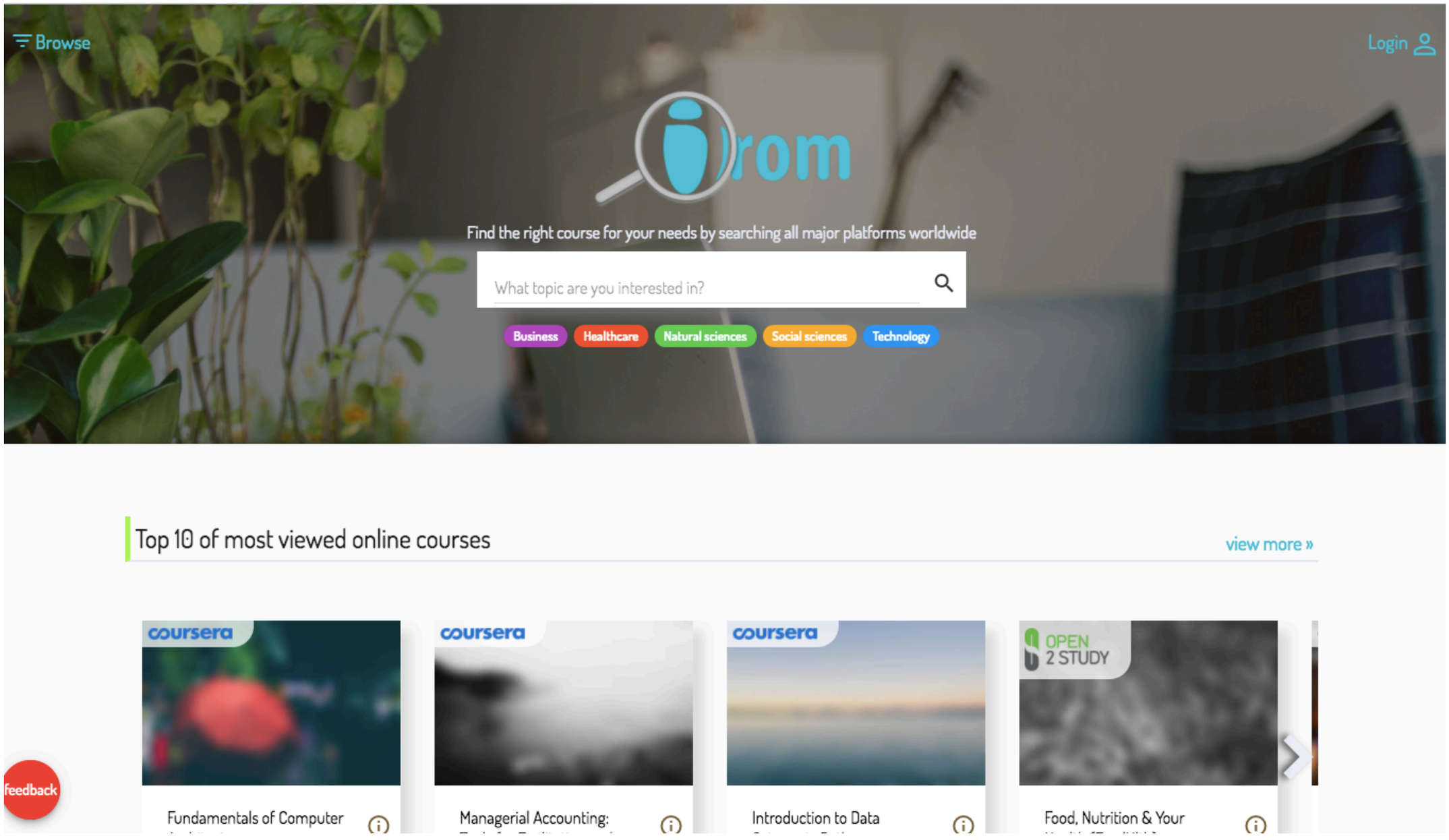
- To improve the learning and studying at the university.
- To develop an intelligent MOOCs search engine

My Motivation

- Define unified categorical set across all MOOC platforms.

Motivation

Link: https://irom.pms.ifi.lmu.de/#/home



Motivation - MOOC

MOOC (Massive Open Online Course)

Massive - Unlimited learners

Open - No requirements

Online - Open access via the web

Course - Filmed lectures/Videos, Readings

Motivation - Popular MOOC platforms

coursera

Future Learn

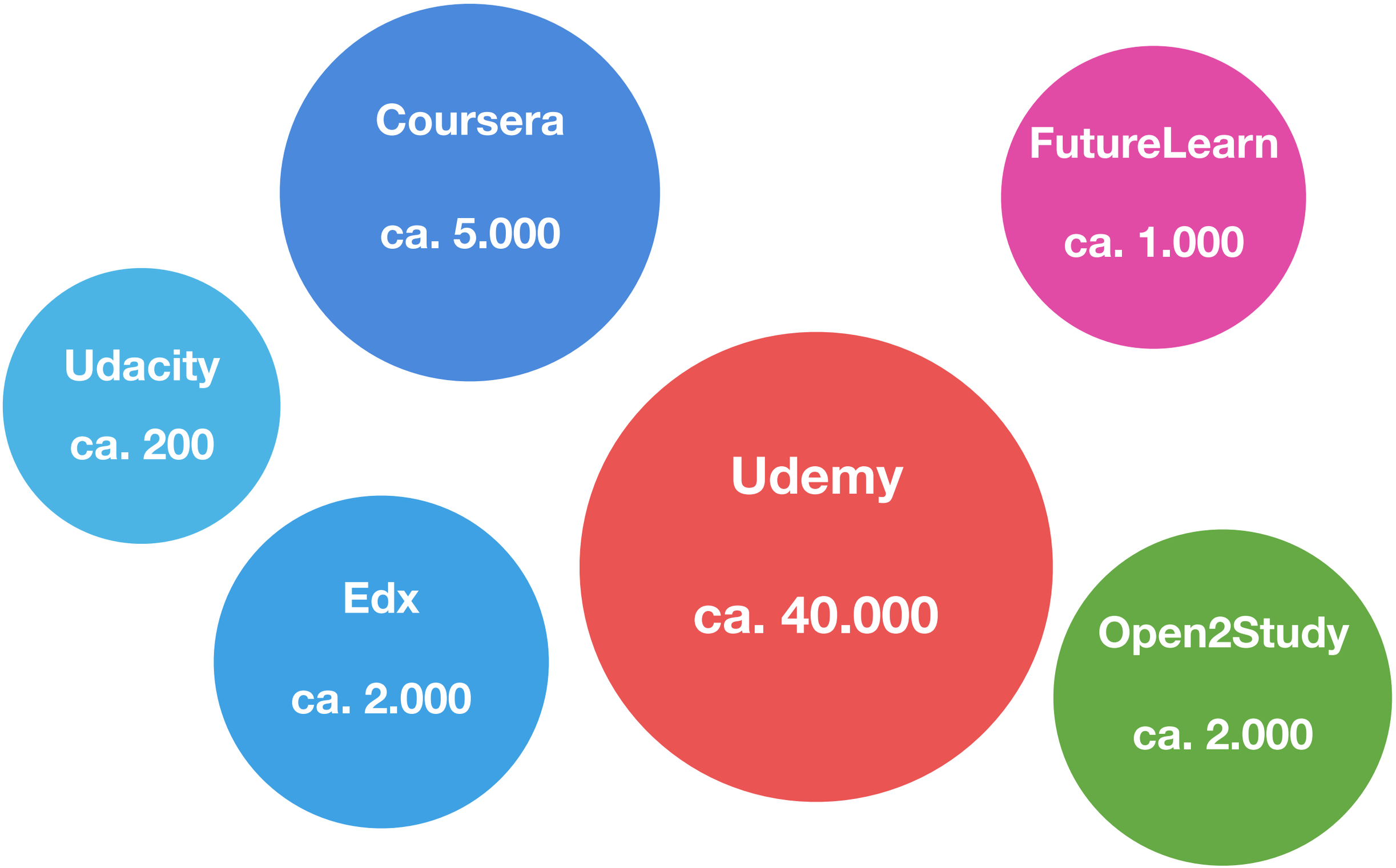
OPEN 2 STUDY

UDACITY

Udemy

edX

Motivation - MOOC platforms by size



Motivation - Diverse categories

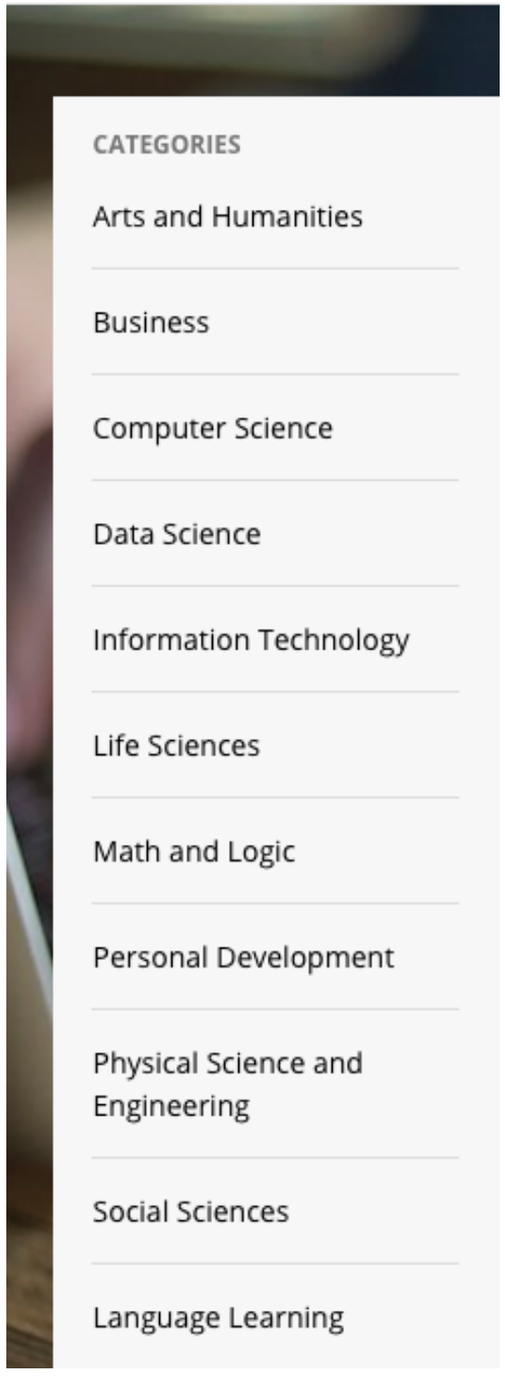
Udemy

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

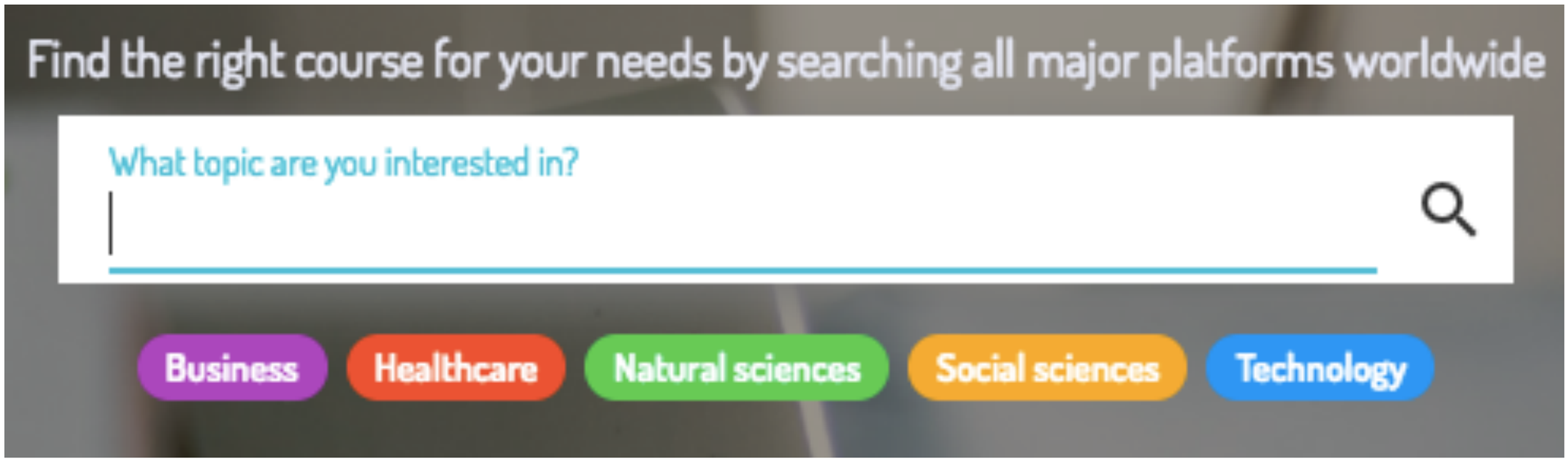


edX

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

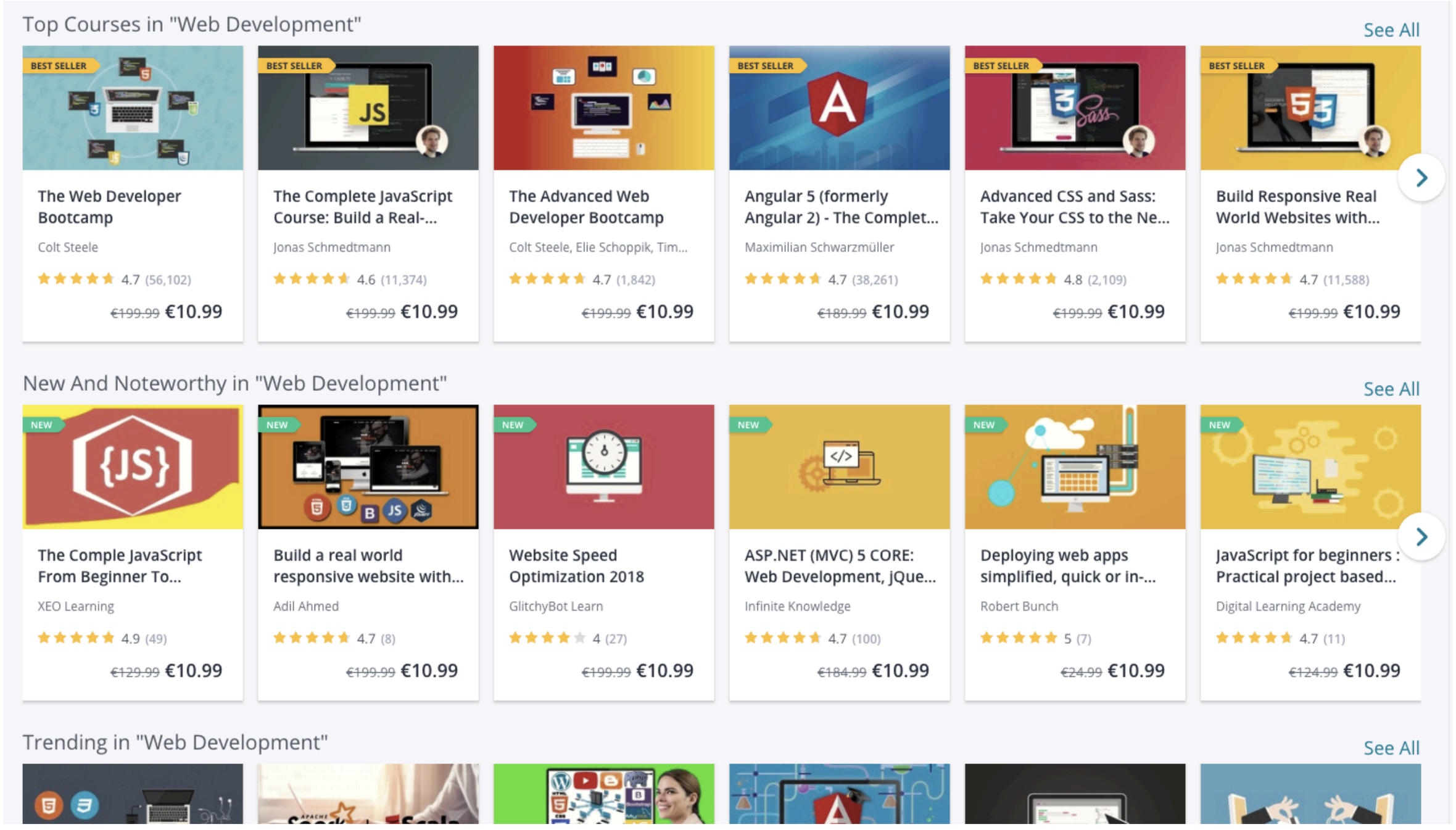
Motivation - Behind my research question

Unified categorical set across all the platforms to allow users to browse through the categories on Irom



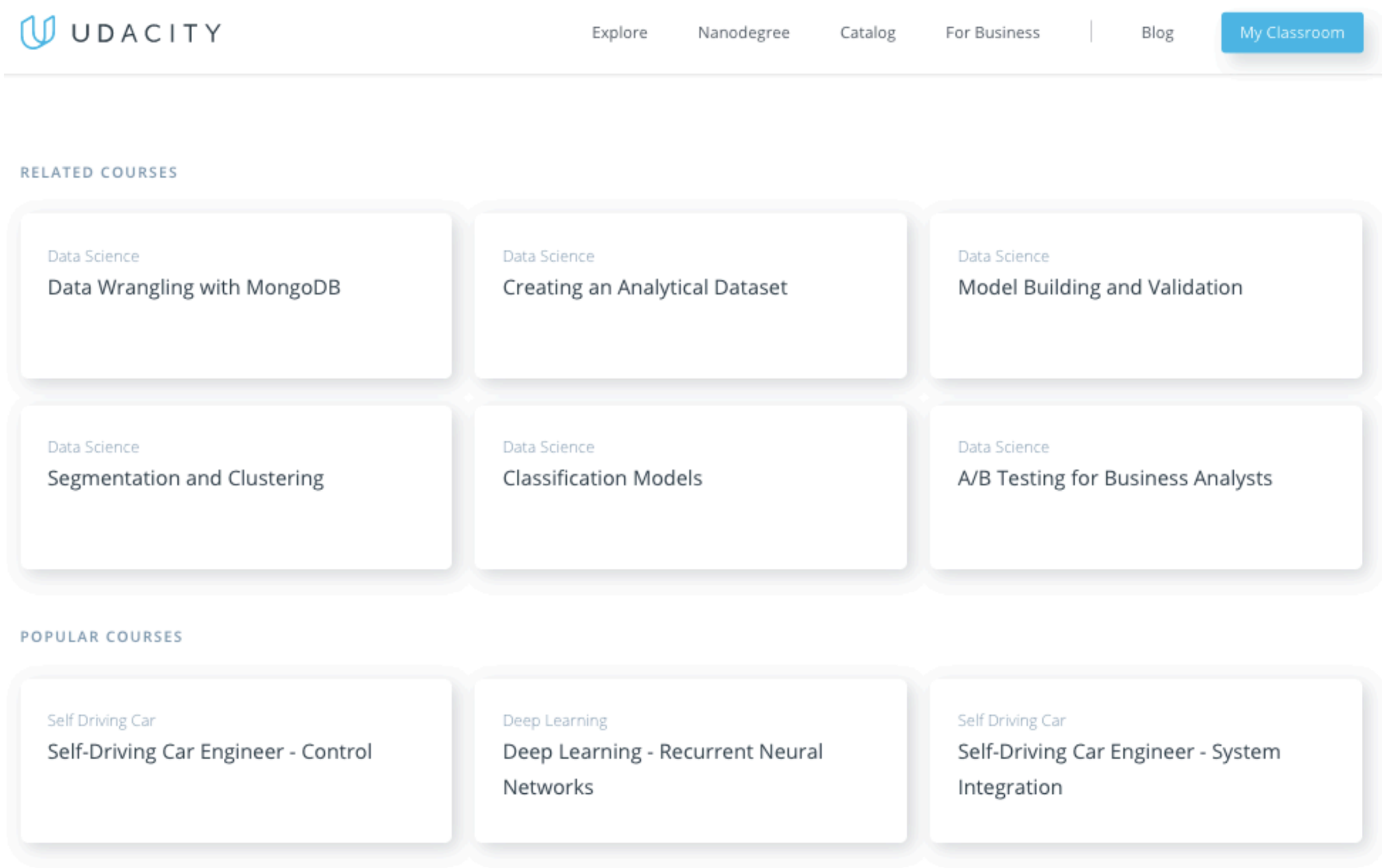
Motivation - Advantages

- Browse & create new subcategories e.g.: “Top Courses”



Motivation - Advantages

- Easier to recommend similar courses



2. Research Topic

Research Topic

COMPARISON OF CATEGORICAL PROPERTIES OFFERED BY MULTIPLE MOOC PLATFORMS

- **Tasks:**
  - Define unified MOOC Model.
  - Web crawl 6 platforms and extract ca. 40.000 courses.
  - Unified Categorical set across all platforms

Research Topic - Unified MOOC Model

```
{
  “title”: String,
  “courseUrl”: String,
  “imageUrl”: String,
  “description”: String,
  “duration”: Int,
  “category”: String,
  ...
}
```

Research Topic

Data Science

Courses

Most n occurring words

[Data,

Science,

Machine,

Learning,

Python,

…,

n]

course 1

```
{
    “title”: String,
    “courseUrl”: String,
    “imageUrl”: String,
    “description”: String,
    “duration”: Int,
    “category”: String
}
```

course 2

```
{
    “title”: String,
    “courseUrl”: String,
    “imageUrl”: String,
    “description”: String,
    “duration”: Int,
    “category”: String
}
```

course 3

```
{
    “title”: String,
    “courseUrl”: String,
    “imageUrl”: String,
    “description”: String,
    “duration”: Int,
    “category”: String
}
```

course m

```
{
    “title”: String,
    “courseUrl”: String,
    “imageUrl”: String,
    “description”: String,
    “duration”: Int,
    “category”: String
}
```

Research Topic

Data Science

↓

Most n occurring words

[**Data**,

Science,

Machine,

Learning,

Python,

…,

n]

Compare ⟷

New course from Udacity

↓

Most n occurring in description words

[**Data**,

Science,

Machine,

Driving,

Car,

…,

n]

3. Project Plan

Project Plan - Timeline

Build web crawlers

Crawler for 6 platforms.

Define unified MOOC model.

Dec 17

Analyse of categories

Remove stop words.

most occurring words.

Jan 18

Compare & Evaluate

Categorical properties.

Define unified categorical set.

Feb 18

Deadline

12 Mar 18

4. Technical details

Technical Details

5. Challenges

Challenges

- Web crawling websites with Javascript
- Defining unified MOOC model
- Websites changing their layout

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

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
- Copy a repository from GitHub onto a local computer for using with a project
- Create a new repository on GitHub
- 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

6. Quick Demo of web crawling

Catalog | Search catalog | For Enterprise | Log In

Home > Data Science > Machine Learning

Practical Machine Learning

- Overview
- Syllabus
- FAQs
- Creators
- Pricing
- Ratings and Reviews

About this course: One of the most common tasks performed by data scientists and data analysts are prediction and machine learning. This course will cover the basic components of building and applying prediction functions with an emphasis on practical applications. The course will provide basic grounding in concepts such as training and tests sets, overfitting, and error rates. The course will also introduce a range of model based and algorithmic machine learning methods including regression, classification trees, Naive Bayes, and random forests. The course will cover the complete process of building prediction functions including data collection, feature creation, algorithms, and evaluation.

Show less

Enroll
Started Jan 22

Apply for Financial Aid

Created by: Johns Hopkins University

JOHNS HOPKINS UNIVERSITY

Training 5 or more people?

Get your team access to Coursera's Catalog.

Try Coursera for Business

Taught by: Jeff Leek, PhD, Associate Professor, Biostatistics
Bloomberg School of Public Health