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
AGENDA

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

• Irom - Intelligent Recommender Of MOOCs
• MOOC - Massive Open Online Course

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
Motivation - MOOC platforms by size

- Udemy: ca. 40.000
- Coursera: ca. 5.000
- Edx: ca. 2.000
- Udacity: ca. 200
- FutureLearn: ca. 1.000
- Open2Study: ca. 2.000
Motivation - Diverse categories
**Motivation** - Behind my research question

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

![Search interface for courses](image)
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

Most n occurring words

[ Data, Science, Machine, Learning, Python, …, n ]

Courses

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

course 1

course 2

course 3

course m
Research Topic

Data Science

Most \( n \) occurring words

\[
[ \text{Data, Science, Machine, Learning, Python, ..., n} ]
\]

Compare

New course from Udacity

Most \( n \) occurring in description words

\[
[ \text{Data, Science, Machine, Driving, Car, ..., n} ]
\]
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

mongoDB

matplotlib

python

Selenium

Scrapy

Python Library
5. Challenges
Challenges

• Web crawling websites with Javascript

• Defining unified MOOC model

• Websites changing their layout
Intro to Machine Learning
Pattern Recognition for Fun and Profit

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

NANODEGREE PROGRAM

Machine Learning Engineer by kaggle
Make Predictive Models

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

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

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
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.

Created by: Johns Hopkins University

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