فاز صفر – جمع آوری داده ها

هدف از این فاز آشنایی با روش های جمع آوری داده از صفحات وب و ساخت یک دیتاست متناسب برای انجام عملیات داده کاوی بر روی آن است.

انتخاب دانشگاه هدف

شما ابتدا میبایست از میان دانشگاه هایی که استخراج اطلاعات دروس از صفحات course catalog آنها امکان پذیر بوده یک دانشگاه را انتخاب نموده و سپس طبق توضیحات اطلاعات دروس را جمع آوری نمایید. دانشگاه مورد نظر خود را میبایست از طریق این لیست انتخاب و نام خود را ثبت نمایید.

Crawl صفحات

ابزارهای مورد استفاده ما در این پروژه Selenium و Beautiful Soup هستند که نحوه کار با هریک در کلاس آموزش داده شده است. به این منظور شما می بایست کدهای خود را با ساخت یک کلاس که از کلاس BaseCrawler ارث بری میکند که یک interface را برای کدهایی که قرار است بنویسید مشخص میکند، شروع نمایید. از آنجایی که اکثر دانشگاه ها دروس را به تفکیک دپارتمان ها در صفحات مختلف قرار میدهند شما می بایست ابتدا این لیست را استخراج و سپس برای هر یک از آنها تابع get_courses_of_department(self, department) را فراخوانی کنید. سپس از طریق یک حلقه برای هریک از دروس موجود در صفحه هر دپارتمان تابع get_course_data(self, course) را فراخوانی نمایید. ویژگی های مهم در این پروژه Prerequisite, Objective, Outcome, Required Skill هستند.
کدهای کلاس BaseCrawler و یک نمونه کد صحیح از اینجا قابل مشاهده است.

فایل خروجی

فایل خروجی میبایست با فرمت **csv** و با نام مختصر دانشگاهی که اطلاعات آنرا استخراج مینمایید باشد. به این منظور مثلا اگر شما در حال crawl اطلاعات دانشگاه Berkeley هستید کافیست تا یک فایل به نام UBC ایجاد نموده و کلاس خود را به همین نام در داخل همین فایل و با ارث بری از کلاس پایه به صورت زیر ایجاد نمایید. در اینصورت هنگامی که شما تابع save_course_data را فراخوانی نمایید اطلاعات به صورتی که مطلوب است در یک فایل در کنار کدهای شما ذخیره میگردد.

```
class UCB(BaseCrawler):
```

نکات تحویل

- این فاز در قالب گروه های ۲ نفره و یا به صورت انفرادی قابل انجام میباشد.
- استفاده از دیگر ابزارهای خزش همانند scrapy و ... غیرمجاز می باشد.
- در حین نوشتن برنامه شما مجاز به تغییر کلاس BaseCrawler و همچنین تابع save_course_data نبوده و میبایست کدهای خود را در دیگر توابع پیاده سازی نمایید.
- پیاده سازی توابع اضافه تر بلامانع میباشد اما باید به نحوی باشد که منجر به کاهش خوانایی کد نشود.
- انتظار میرود تمرکز بیشتری برروی استخراج داده های مربوط به ویژگی ها مهمتر نمایید.
- فایل ارسالی شما میبایست تنها شامل فایل های py. و csv. به نام اختصاری دانشگاه هدف شما باشد.
- فایل ها باید در قالب Student۱Name-Student۲Name-CourseCode-phase۰.zip ارسال شود.
  - CourseCode برای گروه ساعت ۴ عدد ۱ و برای گروه ساعت ۶ عدد ۲ می باشد.
- ارسال فایل تنها از طریق سامانه VU مورد قبول بوده و فایل های ارسال شده در تلگرام و ... تصحیح نخواهند شد.