درس مبانی بازیابی اطلاعات و جستجوی وب

استاد: دکتر محمود نشاطی

ترم اول سال تحصیلی 1400/1399

دانشکده مهندسی و علوم کامپیوتر

تمرین اول: آشنایی با خزش روی صفحات وب

آخرین تاریخ تحویل: 2 آبان ماه

توجه: هر روز تاخیر درتحویل 5 درصد جریمه دارد.

## مقدمه:

در اولین تمرین درس شما با خزش روی صفحات وب و جمع آوری اطلاعات مورد نیاز آشنا خواهید شد. موفقیت در تمرین های بعدی وابسته به انجام صحیح این تمرین است. در پروژه های آینده شما یک موتور جستجوی ساده را راه اندازی خواهید کرد و اطلاعات مورد نیاز برای ساخت و تست موتور جستجو، از تمرین اول شما می‌آید.

## خواسته ها:

در این تمرین از شما انتظار می‌رود یک کرالر، ترجیحاً با استفاده از زبان پایتون و پکیج scrapy پیاده سازی کنید. برای آشنایی با کرالر و scrapy می‌توانید از منابعی که در اطلاعیه اول Courseware مطرح شد استفاده کنید. برای هر گروه یک سایت درنظر گرفته شده است و اهداف و اطلاعاتی که باید از آن سایت استخراج کنید به صورت متفاوتی طرح شده است. در آدرس زیر اطلاعات مربوط لیست پروژه ها آورده شده است. در این صفحه لینک وبسایتی که قرار است کرال شود، اطلاعات خواسته شده، راهنمایی ها و دوره آپدیت آورده شده است. برای ذخیره سازی اطلاعات در پایگاه داده از هر پایگاه داده ای که خواستید می‌توانید استفاده کنید. همچنین برنامه شما باید در دوره های زمانی خواسته شده اطلاعات جدید را به صورت خودکار دریافت کند. برنامه شما باید خروجی .json داشته باشد؛ و اطلاعات کرال شده را در لحظه ی خواسته شده درون یک فایل json بریزد. این فایل را به عنوان یکی از خروجی های برنامه در کنار بقیه فایل ها به صورت زیپ بفرستید. درصورتی که حجم آن بالا بود میتوانید آن را در یکی از سایت های آپلود مانند گوگل درایو ذخیره کرده و لینک آنرا در گزارش خود بیاورید.

آدرس صورت مسائل: لینک

## لیست تمرین گروه ها:

از آنجا که تمرین ها متفاوت هستد، آن ها را به صورت کاملا رندوم به گروه ها اختصاص دادیم.

assignment.py > ...

```python
import random

group_numbers = list(range(1, 24))
assignment_numbers = list(range(1, 24))

gc = len(group_numbers)
ac = len(assignment_numbers)

group_sampled_list = random.sample(group_numbers, gc)
assignment_sampled_list = random.sample(assignment_numbers, ac)

print(group_sampled_list)
print(assignment_sampled_list)
```

PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\Hamed\Documents\Code> python .\assignment.py
[16, 21, 13, 2, 7, 14, 9, 6, 23, 1, 22, 12, 11, 20, 3, 10, 5, 18, 19, 8, 15, 4, 17]
[22, 17, 1, 3, 18, 7, 14, 12, 15, 23, 4, 11, 20, 16, 19, 8, 13, 10, 6, 21, 9, 2, 5]
PS C:\Users\Hamed\Documents\Code>
```

## نتایج تمرین گروه‌ها

نتایج به شرح زیر است:

| شماره گروه | شماره تمرین |
|---|---|
| **16** | 22 |
| **21** | 17 |
| **13** | 1 |
| **2** | 3 |
| **7** | 18 |
| **14** | 7 |
| **9** | 14 |
| **6** | 12 |
| **23** | 15 |
| **1** | 23 |
| **22** | 4 |
| **12** | 11 |
| **11** | 20 |
| **20** | 16 |
| **3** | 19 |
| **10** | 8 |
| **5** | 13 |
| **18** | 10 |
| **19** | 6 |
| **8** | 21 |
| **15** | 9 |
| **4** | 2 |
| **17** | 5 |

جهت یادآوری شماره گروه شما از طریق صفحه دوم این لینک قابل دسترسی است.

## نکات پیاده سازی:

1. توجه کنید که خزش روی صفحات وب توسط کرالر با سرعت بالایی انجام می‌شود. بنابرین کرالر شما ممکن است در مدت زمان کوتاهی تعدادی زیادی درخواست http را به سمت سایت مربوطه ارسال کند. این حجم زیاد و فرکانس بالا از درخواست ها می‌تواند باعث شود که سایت مربوطه (و یا فایروال سرور)، ارسال کننده (کامپیوتر شما با آیپی مشخص) را یک مهاجم تشخیص دهد و درخواست های شما را در صورت گذشتن از حد مجاز رد کند. در این صورت درخواست شما با یک پاسخ 403 (Forbidden) رد می‌شود و اطلاعات وبسایت برای شما ارسال نخواهد شد. لذا باید

توجه کنید اگر که با چنین مشـکلی روبرو شـدید، به برنامه تان یک خط کد خواب(sleep) اضـافه کنید و ترد اصلی برنامه را پس از بازیابی صفحه مربوطه برای چند ثانیه( معمولا 3 ثانیه کافی است اما درصورت مواجه مجدد با این مشکل میتوانید آنرا افزایش دهید). راه هوشمندانه تر این است پس از مواجهه با این پاسخ ترد مربوطه را به خواب ببرید و سپس درخواست رد شده را تکرار کنید.

2. در خزش روی وبسایت شما عملاً با دو نوع وبسایت روبرو هستید: وبسایت های دارای درخواست ajax و وبسایت های فاقد آن. به طور خلاصه، صفحات دارای Ajax صفحاتی هستند که برای به روز رسانی بخشی از صفحه، بدون رفرش آن، یک درخواست http به سرور مربوطه ارسال و صرفا اطلاعات همان بخش را دریافت می‌کند و در صـفحه قرار می‌دهد. این کار با اسـتفاده از زبان جاوا اسکریپت و api سایت مربوطه انجام می‌شود. در تمرین بعضی از گروه ها مانند تمرین شماره 23 که در آن هدف جمع آوری اطلاعات قیمت دلار و ... است، شما عملا با یک جدول(یا هر آیتم html دیگر) روبرو هستید که خود را با استفاده از درخواست ajax به روز رسانی می‌کند. در این نوع از تمرین ها شما باید به ارتباطات برقرار شده بین مرورگر و سایت مربوطه دقت کنید تا api مربوطه را کشـف کنید و اطلاعات را از آن api به طور مسـتقیم دریافت کنید. برای اینکار ابتدا صـفحه inspect element را فعال کنید. سپس کاری که باعث فعال سازی و شلیک درخواست ajax می‌شود را شناسایی کنید و آنرا انجام دهید. این کار می‌تواند شامل اسکرول کردن، کلیک روی یک دکمه و ... باشد. سپس در تب Network گزینه‌ی XHR را فعال کنید تا این ارتباطات را کشف کنید. از طریق این درخواست های ارسال شده میتوانید api را شناسایی کنید. برای شناسایی api ( یا همان لینک آن) اطلاعات مربوط به هدر این درخواست http و پاسخ http و حتی بررسی بدنه این پیام، اطلاعات مربوطه را به شما می‌دهد. پس از شناسایی این api ها، به راحتی میتوانید آن ها را صـدا زده و فایل json را تحویل بگیرید و پس از پارس کردن آن، آنرا در دیتابیس ذخیره کنید. در چند تمرین این کشف api توسط ما برای راهنمایی شما قرار داده شده است.
3. در بعضی از مواقع که شما api را صدا می‌زنید، سرور یک پاسخ مبتنی بر کاراکتر های یونیکد میفرستد که برای ما خوانایی ندارد. از آنجا که تمرین شما از سایت های فارسی طرح شده است، کاراکتر های فارسی به احتمال بالا توسط api به صورت ناخوانا و در قالب کاراکترهای یونیکد ارسال می‌شوند. برای خوانایی حتما آنها را به utf-8 تبدیل کنید و سپس در پایگاه داده ذخیره کنید. توجه داشته باشید که احتمالاً اندکی از گروه ها با این مشکل روبرو خواهند شد و این بند شامل همه گروه ها نمی‌شود. برای مثال دو متن زیر یکسان هستند و متن انگلیسی نشان دهنده ی متن فارسی، با کاراکتر های یونیکد متناظر است:

قیمت کل

\u0642\u06cc\u0645\u062a \u06a9\u0644

## نکات مهم:

- پروژه به صورت تیم های حداکثر دونفری انجام خواهد گرفت.
- گرفتن نمره ی کامل منوط به **حضور** در جلسه‌ی احتمالی تحویل اسکایپی و تسلط بر روی کد می باشد که زمان آن متعاقبا اعلام خواهد شد. در این جلسه شما باید کدی که در کورس ویر آپلود کردید را بر روی رایانه ی خود، مجددا کامپایل کرده و از برنامه خروجی بگیرید.
- همچنین هر دو عضو گروه باید بر روی کد کاملا مسلط باشند. **دقت داشته باشید که عدم شرکت در جلسه اسکایپی در صورت تشکیل، موجب از دست دادن درصد زیادی از نمره خواهد شد.**
- برای برنامه نویسی از زبان های پایتون، جاوا و C# میتوانید استفاده کنید. درصورت استفاده از زبان پایتون و پکیج Scrapy، ده درصد نمره‌ی پایه، به شما نمره ی اضافی تعلق می‌گیرد.
- تمیزی و خوانایی کد، نمره اضافی خواهد داشت.
- سورس کد برنامه ، خروجی برنامه( فایل اجرایی) و گزارش را به صورت ir_proj1_{groupNO}.zip بر روی کورس ویر آپلود کنید.
- دقت کنید که در صورت استفاده از زبان جاوا ، فایل های *.java و *.class یک فایل خروجی محسوب نمی شود و باید یک فایل *.jar به عنوان فایل اجرایی بفرستید.
- در صوورتی که از زبان پایتون استفاده می کنید، حتماً باید پکیج هایی که استفاده کردید را در فایل requirements.txt ذخیره کرده باشید. همچنین در مستندات خوود به ورژن پایتونی که استفاده کرده‌اید، اشاره کنید.
- ترجیحا ازآخرین ورژن پکیج های کرالر و زبان های برنامه نویسی استفاده نمایید.
- از خروجی خود یک گزارش با فرمت pdf. تهیه کنید. تعداد صفحات و فرمت گزارش مهم نیست صرفا توضیحاتی کوتاه به همراه چند اسکرین شات از اجرای برنامه و خروجی آن در گزارش گنجانده شود.
- در صورت داشتن هرگونه سوال مرتبط با این پروژه، می توانید با ایمیل h.sanaei@mail.sbu.ac.ir در ارتباط باشید. لطفا در نظر داشته باشید که سوال هایتان را نهایتاً تا سه روز مانده به ددلاین مطرح کنید، چرا که پس از آن احتمال پاسخ به شما بسیار کم می باشد.

با آرزوی موفقیت و سربلندی

حامد ثنائی