Pràctica 1 (25% nota final)

Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar dades rellevants per a un projecte analític i utilitzar eines d'extracció de dades. Per fer aquesta pràctica haureu de treballar en grups de dues persones.

De manera orientativa, es poden consultar els següents exemples, tenint en compte que poden no ser respostes perfectes per a la pràctica que es planteja aquest semestre:

- Exemple: https://github.com/rafoelhonrado/foodPriceScraper
- Exemple complex: https://github.com/tteguayco/Web-scraping

El lliurament d'aquesta pràctica s'ha de realitzar tal com s'especifica a l'apartat Format i data de lliurament. Haureu de lliurar al RAC un únic arxiu amb l'enllaç al repositori on hi hagi el codi font utilitzat per a dur a terme l'extracció, el dataset, i un document PDF amb les respostes als apartats. A més, haureu de realitzar un vídeo explicatiu de la pràctica, en el qual tots dos integrants del grup han de comentar els aspectes més rellevants del projecte, tant de les respostes als apartats com del codi utilitzat per a extreure les dades.

És important tenir en compte les següents consideracions a l'hora de lliurar la pràctica:

- És obligatori i **queda com a responsabilitat de l'estudiant revisar que l'arxiu lliurat al RAC és el correcte.** Un arxiu buit o no pertinent es considerarà com no lliurat.
- Perquè el lliurament es consideri com realitzat, s'ha de completar almenys el 25% de tota l'activitat.
- No podrà modificar-se cap element de la pràctica passada la data de lliurament (repositori, arxius de Google Drive, etc.).
- Així mateix, també és responsabilitat de l'estudiant assegurar-se que, en el moment del lliurament de la pràctica, **s'ha donat accés al professor als diferents elements privats que es lliurin** (p. ex., repositori GitHub privat o arxius restringits de Google Drive). El professor indicarà al Tauler de l'aula el seu nom d'usuari en aquestes plataformes.

Competències

En aquesta pràctica es desenvolupen les següents competències del Màster universitari en Ciència de Dades:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per resoldre-ho.
- Capacitat per aplicar les tècniques específiques de web scraping.

Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants que el seu tractament aporta valor a una empresa i la identificació de nous projectes analítics.
- Saber identificar les dades rellevants per dur a terme un projecte analític.
- Capturar dades de diferents fonts de dades (tals com a xarxes socials, web de dades o repositoris).
- Actuar segons els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Descripció de la pràctica a realitzar

L'objectiu d'aquesta activitat serà la creació d'un dataset a partir de les dades contingudes a un lloc web. Han de tenir-se en compte les consideracions sobre el lloc web triat, el codi i el dataset que s'indiquen més endavant. S'haurà de presentar un document PDF **(màxim 20 pàgines)** en el qual es resolguin els següents apartats:

1. **Context.** Explicar en quin context s'ha recollit la informació. Explicar per què el lloc web triat proporciona aquesta informació. Indicar l'adreça del lloc web.
2. **Títol.** Definir un títol que sigui descriptiu pel dataset.
3. **Descripció del dataset.** Desenvolupar una descripció breu del conjunt de dades que s'ha extret. És necessari que aquesta descripció tingui sentit amb el títol escollit.
4. **Representació gràfica.** Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.
5. **Contingut.** Explicar els camps que inclou el dataset i el període de temps de les dades.
6. **Propietari.** Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-n'hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte.
7. **Inspiració.** Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l'apartat 6.
8. **Llicència.** Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció. Exemples de llicències que poden considerar-se:
    - Released Under CC0: Public Domain License.
    - Released Under CC BY-NC-SA 4.0 License.
    - Released Under CC BY-SA 4.0 License.

- Database released under Open Database License, individual contents under Database Contents License.
- Altres (especificar quina).

9. **Codi.** Codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.
   - El codi haurà de situar-se a la carpeta **/source** del repositori.
   - S'han d'indicar les llibreries i versions utilitzades. P. ex., en Python poden obtenir-se mitjançant la comanda
     `pip3 freeze > requirements.txt`
   - Al document PDF s'han de comentar els aspectes més rellevants sobre com el codi realitza el procés de recol·lecció de dades, quines dificultats presenta el lloc web triat, i com les heu resolt.
10. **Dataset.** Publicar el dataset obtingut en format CSV a Zenodo, incloent-hi una breu descripció. Obtenir i adjuntar l'enllaç del DOI del dataset (https://doi.org/…). El dataset també haurà d'incloure's a la carpeta **/dataset** del repositori.

    Si existeix qualsevol circumstància que impedeixi publicar obertament el dataset real a Zenodo, s'haurà de: (1) comentar aquesta circumstància i justificar el motiu en aquest apartat; (2) generar un dataset simulat i publicar-lo a Zenodo, obtenint l'enllaç del DOI; i (3) comunicar al professor el dataset real de manera privada (p. ex., utilitzant un repositori privat).

11. **Vídeo**. Realitzar un breu vídeo explicatiu de la pràctica **(màxim 10 minuts)**, que haurà de comptar amb la participació dels dos integrants del grup. Al vídeo s'haurà de realitzar una presentació del projecte, destacant els punts més rellevants, tant de les respostes als apartats com del codi utilitzat per a extreure les dades. Indicar l'enllaç del vídeo (https://drive.google.com/…), que haurà d'estar al Google Drive de la UOC.

Consideracions sobre el lloc web triat, el codi i el dataset

A l'hora de triar un lloc web per a realitzar aquesta pràctica, és important tenir en compte que l'objectiu és que el lloc web permeti extreure un dataset "interessant" i que el procés d'extracció no sigui completament "trivial":

- L'idioma del lloc web triat ha de ser **castellà**, **anglès** o **català**.
- El codi generat per a obtenir el dataset ha de realitzar **descobriment d'enllaços i navegació autònoma**. P. ex., no és suficient amb processar el contingut d'una única pàgina web on aparegui tot el dataset a una taula.
- El codi ha d'implementar mecanismes per a fer un **bon ús del web scraping** (p. ex., evitar saturar al servidor).
- Ha de comprovar-se **quin User-Agent està utilitzant el codi**, encara que s'utilitzi un WebDriver.

- **No es permet l'ús d'APIs com a part principal de la pràctica.** En cas que el lloc web triat ofereixi alguna API per a accedir a les dades, s'haurà de prescindir d'aquesta. Es permet l'ús d'APIs com a part complementària a la pràctica, p. ex., per a realitzar consultes a algun servei addicional per a tractar/completar les dades recollides.
- El codi ha de tenir un **nivell adequat de modularitat i estar degudament comentat**. No es tracta d'introduir un comentari per cada línia de codi, sinó de comentar punts clau que ajudin a seguir i entendre què està realitzant.
- **No és necessari realitzar neteja del dataset resultant**, ja que això serà objecte de la Pràctica 2. És interessant que el dataset resultant contingui varietat de dades numèriques i categòriques si es desitja utilitzar per a la Pràctica 2 (no és obligatori utilitzar aquest dataset).

La nota final tindrà en compte les dificultats abordades en la recol·lecció del dataset. Alguns aspectes que incrementen la dificultat són:

- Ús de tecnologies avançades com Selenium o Scrapy.
- Recol·lecció de dades de llocs web amb contingut dinàmic (p. ex., *infinite scroll*, *mouseover*).
- Ús de mètodes avançats per a saltar-se la prevenció de web scraping.
- Gestió de contingut audiovisual.
- Gestió d'usuaris i contrasenyes.
- Gestió de codi JavaScript.

Recursos

Els següents recursos són d'utilitat per la realització de la pràctica:

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. El lenguaje Python. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
- Tutorial de Github https://guides.github.com/activities/hello-world.

Criteris de valoració

Tots els apartats són obligatoris. La ponderació dels apartats és la següent:

Apartat	1	2	3	4	5	6	7	8	9	10	11
Punts	0,25	0,25	0,25	0,5	1	1,5	1,25	0,5	2	2	0,5

Criteris que es tindran en compte per a l'avaluació de la pràctica són:

- Idoneïtat de les respostes (hauran de ser clares i completes).
- **Complexitat** del lloc web triat per a l'extracció. És important tenir en compte que la complexitat serà un factor que s'avaluarà i dependrà tant del lloc triat com de l'anàlisi realitzat a la pràctica.
- Síntesi i claredat, a través de l'ús de comentaris, del codi resultant.
- Presentació adequada de les dades.
- Organització i claredat dels documents de lliurament final.
- Completitut dels documents requerits per al lliurament final.
- Seguiment de recomanacions per al bon ús del web scraping.

Format i data de lliurament

En referència al lliurament de la pràctica, es demana:

a. **Un únic document (.txt o .pdf)** que contingui l'enllaç al repositori Git del projecte (https://github.com/…). Aquest document es lliurarà, per cadascun dels integrants del grup a l'espai de Lliurament i Registre d'AC de l'aula.

b. **Un repositori Git** amb la resolució de la pràctica a la branca "**main**". El repositori es crearà a GitHub (https://github.com/), i podrà ser un repositori públic o privat, a elecció del grup. Si es fa servir un repositori privat, s'haurà de donar accés al professor. El repositori haurà de contenir:

   b.1. **Un document README.md**: Estarà situat a la carpeta arrel i haurà de contenir els noms dels integrants del grup, una descripció dels arxius que componen el repositori i el DOI de Zenodo del dataset generat.

   b.2. **Un document memòria.pdf**: Estarà situat a la carpeta arrel i haurà de contenir els noms dels integrants del grup i les respostes als apartats 1 a 11. Ha de ser un **document PDF** (no s'accepten altres formats), l'extensió del qual **no ha de superar les 20 pàgines**. A més, al final del document, ha d'aparèixer la següent taula de contribucions al treball, la qual ha de signar cada integrant del grup amb les seves inicials. Les inicials representen la confirmació per part del grup que l'integrant ha participat a aquest apartat. Tots els integrants han de participar en cada apartat, per la qual cosa, idealment, els apartats haurien d'estar signats per tots els integrants.

Contribucions	Signatura
Investigació prèvia	Integrant 1, Integrant 2
Redacció de les respostes	Integrant 1, Integrant 2
Desenvolupament del codi	Integrant 1, Integrant 2
Participació al vídeo	Integrant 1, Integrant 2

b.3. **Carpeta /source**: Haurà de contenir el codi Python o R generat per a obtenir les dades.

b.4. **Carpeta /dataset**: Haurà de contenir el dataset resultant en format CSV.

c. **Un vídeo explicatiu**, la durada del qual **no ha de superar els 10 minuts**. L'enllaç del mateix s'ha d'indicar a l'apartat 11 del document PDF.

El document del lliurament s'ha de pujar a l'espai de Lliurament i Registre d'AC de l'aula abans de les **23:59 CET del dia 22 de novembre**. No s'acceptaran lliuraments fora de termini. **No podrà modificar-se cap element de la pràctica passada la data de lliurament** (repositori, arxius de Google Drive, etc.).

Si s'estima oportú, el professor sol·licitarà als integrants del grup una entrevista remota (de manera conjunta o individual) mitjançant Google Meet, en referència a la pràctica realitzada, en un dia i hora acordats.