Pràctica 1 (25% nota final)

Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per a un projecte analític i utilitzar eines d'extracció de dades. Per fer aquesta pràctica haureu de treballar en grups de 2 persones. Haureu de lliurar un sol fitxer amb l'enllaç al repositori Git on hi hagi les solucions, incloent-hi els noms dels components del grup. Podeu utilitzar la Wiki o README.md del repositori per descriure el vostre grup i els diferents arxius del vostre lliurament. Cada membre del grup haurà de contribuir amb el seu usuari del repositori. Podeu revisar aquests exemples com a guia (recordeu que es tracta d'exemples i no de respostes perfectes per a la práctica):

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

A més, heu de lliurar un vídeo explicatiu de la pràctica on cadascun dels integrants del grup expliqui amb les seves pròpies paraules tant les respostes del projecte com el codi utilitzat per a dur a terme l'extracció. El vídeo ha de ser enviat a través d'un enllaç a **Google Drive** que heu de proporcionar, juntament amb l'enllaç al repositori Git, al moment de lliurar la práctica.

Recordeu que és obligatori i queda com a responsabilitat de l'estudiant revisar que el fitxer lliurat és el correcte. Un fitxer buit o no pertinent es considerarà com no lliurat. Així mateix, perquè un lliurament es consideri com realitzat, s'ha de completar almenys un 25% de tota l'activitat.

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. L'idioma del lloc web escollit haurà de ser castellà, anglès o català. S'hauran de resoldre els següents apartats:

1. **Context.** Explicar en quin context s'ha recollit la informació. Explicar per què el lloc web triat proporciona aquesta informació.
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, el període de temps de les dades i com s'ha recollit.
6. **Agraïments.** 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:
    - 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
    - Other (specified above)
    - Unknown License
9. **Codi.** Adjuntar al repositori Git el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

10. **Dataset.** Publicar el dataset obtingut*(*)* en format CSV a Zenodo amb una breu descripció. Obtenir i adjuntar l'enllaç del DOI.
11. **Vídeo**. S'ha de lliurar un vídeo explicatiu de la pràctica on cadascun dels integrants del grup expliqui amb les seves pròpies paraules tant les respostes del projecte com el codi utilitzat per a dur a terme l'extracció. El vídeo ha de ser enviat a través d'un enllaç a Google Drive que heu de proporcionar, juntament amb l'enllaç al repositori Git, al moment de lliurar la pràctica.

(*) Si existeix qualsevol impediment per publicar el dataset real, s'haurà de justificar aquesta situació i publicar a Zenodo un dataset simulat. En aquest cas, el dataset real es comunicarà al professor de forma privada (p.ex., enllaç de Google Drive).

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 exercicis 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

Durant la setmana **del 28 de març al 01 d'abril**, el grup podrà fer un lliurament parcial opcional. Aquest lliurament parcial és molt recomanable per rebre assessorament sobre la pràctica i verificar que la direcció presa és la correcta. Es lliuraran comentaris als estudiants que hagin efectuat el lliurament parcial, però no comptaran per a la nota de la pràctica. En el lliurament parcial els estudiants hauran de lliurar per correu electrònic, al professor encarregat de l'aula, l'enllaç al repositori Git amb allò que hagin avançat. És important destacar que aquest lliurament és una guia per a verificar que la pràctica s'està realitzant en la direcció correcta i no ha de prendre's com una primera correcció de les preguntes. El professor verificarà principalment que el lloc web triat i la temàtica és adequada però no es valorarà si les preguntes s'estan responent correctament.

En referència al lliurament final, es demana:

a. **Un únic document** (.txt, .pdf, .docx) que contingui **l'enllaç al repositori Git** del projecte (apartat b) i **l'enllaç al vídeo del projecte** (apartat c). Aquest document es lliurarà a l'espai de Lliurament i Registre d'AC de l'aula.
b. Un **repositori Git** amb les solucions de la pràctica. El repositori Git es crearà a Github (https://github.com/), i podrà ser un repositori públic o privat, a elecció del grup. Si s'utilitza un repositori privat, s'haurà de facilitar accés al professor, mitjançant el nom d'usuari que s'indicarà al Tauler de l'aula o per email. **El repositori no es podrà modificar passada la data de lliurament**, i haurà de contenir:
    b.1. Una **Wiki** o **README.md** on estiguin els noms dels components del grup, una descripció dels fitxers i el DOI de Zenodo del dataset generat.
    b.2. Un **document PDF** (no s'acceptaran altres formats) amb les respostes als apartats 1-10 i els noms dels components del grup. **L'extensió d'aquest document 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 en 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

b.3. Una carpeta amb el **codi Python** o **R** generat per obtenir les dades.

c. Un **breu vídeo** amb la participació dels dos components del grup, on es realitzarà una presentació del projecte, destacant els punts més rellevants. El vídeo s'haurà de compartir mitjançant un enllaç del Google Drive de la UOC o incloure-ho al repositori Git. **La durada d'aquest vídeo no ha de superar els 10 minuts.**

El document del lliurament final s'ha de pujar a l'espai de Lliurament i Registre d'AC de l'aula abans de les **23:59 CET del dia 11 d'abril**. No s'acceptaran lliuraments fora de termini.

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.