Pràctica 1 (35% nota final)

Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'extracció de dades. Per fer aquesta pràctica haureu de treballar en grups de 3 o 2 persones, o si preferiu, també podeu fer-ho de manera individual. Haureu de lliurar un sol fitxer amb l'enllaç Github (https://github.com) on hi hagi les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius que corresponen la vostra entrega. Cada membre de l'equip haurà de contribuir amb el seu usuari Github. Podeu mirar aquests exemples com guia:

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

Competències

En aquesta PAC es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- 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 dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants que el seu tractament aporten 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) i mitjançant diferents mecanismes (tals com queries, API i scraping).
- Actuar amb 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 al web. Heu d'indicar les següents característiques del dataset general:

1. Títol del dataset. Cal que poseu un títol que sigui descriptiu.
2. Subtítol del dataset. Agregueu una descripció àgil del vostre conjunt de dades pel vostre subtítol.
3. Imatge. Agregueu una imatge que identifiqui el vostre dataset visualment
4. Context. Quina és la matèria del conjunt de dades?
5. Contingut. Quins camps inclou? Quin és el període de temps de les dades i com s'ha recollit?
6. Agraïments. Qui és propietari del conjunt de dades? Inclou cites de recerca o anàlisi anteriors.
7. Inspiració. Per què és interessant aquest conjunt de dades? Quines preguntes li agradaria respondre la comunitat?
8. Llicència. Cal que seleccioneu una d'aquestes llicències i cal dir perquè l'heu seleccionada:
    - 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: Cal adjuntar el codi amb el que heu generat el dataset, preferiblement amb R o Python, que us ha ajudat a generar el dataset
10. Dataset: Dataset en format CSV

Recursos

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

- El llenguatge Python
- 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 apartat són obligatoris. La ponderació dels exercicis és la següent:

- Els apartats 1, 2, 3 i 4 valen 0,25 punts cadascun.
- Els apartats 5, 6, 7, 8 valen 1 punt cadascun.
- Els apartats 9 i 10 valen 2,5 punts cadascun.

Format i data de lliurament

Durant la setmana del 19 de març el grup podrà lliurar al professor una entrega parcial opcional. Aquesta entrega parcial és molt recomanable per tal de rebre assessorament sobre la pràctica i verificar que la direcció presa és la correcta. Es lliuraran comentaris als estudiants que hagin efectuat l'entrega parcial però no comptarà per la nota de la pràctica. En l'entrega parcial els estudiants hauran de lliurar per correu electrònic (mcalvogonza@uoc.edu) l'enllaç al repositori Github amb el que hagin avançat.

Pel que fa a l'entrega final, cal lliurar un únic fitxer que contingui l'enllaç a Github on hi hagi:

1. Una Wiki on hi hagi els noms dels components del grup i una descripció dels fitxers.
2. Un document Word, Open Office o PDF amb les respostes a les preguntes i els noms dels components del grup.
3. Una carpeta amb el codi Python o R generat per obtenir les dades.
4. El fitxer CSV amb les dades.

Aquest document de l'entrega final de la Pràctica 1 s'ha de lliurar a l'espai de Lliurament i Registre d'AC de l'aula abans de les **23:59 del dia 16 d'abril**. No s'acceptaran lliuraments fora de termini.