**Pràctica 1 (35% nota final)**

**Presentació**

En aquesta practica 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 practica haureu de treballar en grups de 2 persones. 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 del vostre lliurament. Cada membre de l'equip haurà de contribuir amb el seu usuari Github. Podeu revisar aquests exemples com a 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 resoldre-ho.
* Capacitat per aplicar les tècniques especifiques de web scraping.

**Objectius**

Els objectius concrets d'aquesta practica son:

* Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contextos mes 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) 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 us 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 en una web. Per a la seva realització, s'han de complir els següents punts:

1. Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per que el lloc web triat proporciona aquesta informació.
2. Definir un títol pel dataset. Triar un títol que sigui descriptiu.
3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (es necessari que aquesta descripció tingui sentit amb el títol triat).
4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment
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. Es necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).
7. Inspiració. Explicar per que es interessant aquest conjunt de dades i quines preguntes es pretenen respondre.
8. Llicencia. Seleccionar una d'aquestes llicencies pel dataset resultant i explicar 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

1. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.
2. Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.

**Recursos**

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

* 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 Meisner, 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 son obligatoris. La ponderació dels exercicis es la següent:

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

Altres criteris que es prendran en compte per a l’avaluació son:

* Idoneïtat de les respostes (hauran de ser clares i completes).
* Complexitat del lloc web triat per a l'extracció.
* Síntesis i claredat, a traves 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.

**Format i data de lliurament**

Durant la setmana del 26 al 30 d’octubre, el grup podrà lliurar al professor un lliurament parcial opcional. Aquest lliurament parcial es molt recomanable per rebre assessorament sobre la practica i verificar que la direcció presa es la correcta. Es lliuraran comentaris als estudiants que hagin efectuat el lliurament parcial però no comptaran per a la nota de la practica. En el lliurament parcial els estudiants hauran de lliurar per correu electrònic, al professor encarregat de l’aula, l'enllaç al repositori Github amb el que hagin avançat.

En referent al lliurament final, cal lliurar un únic fitxer que contingui l'enllaç a Github on hi hagi:

1. Una Wiki on estiguin els noms dels components del grup i una descripció dels fitxers.
2. Un document PDF amb les respostes a les preguntes i els noms dels components del grup. A mes, 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 a cada apartat, per la qual cosa, idealment, els apartats haurien d'estar signats per tots els integrants.

|  |  |
| --- | --- |
| **Contribucions** | **Signa** |
| Recerca prèvia | Integrant 1, Integrant 2, ... |
| Redacció de les respostes | Integrant 1, Integrant 2, ... |
| Desevolupament codi | Integrant 1, Integrant 2, ... |

1. Una carpeta amb el codi Python o R generat per obtenir les dades.
2. El DOI a les dades.

Aquest document del lliurament final s'ha de lliurar a l'espai de Lliurament i Registre d'AC de l'aula abans de les **23:59 del dia 9 de novembre**. No s'acceptaran lliuraments fora de termini.

**Respostes**

1. **Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per que el lloc web triat proporciona aquesta informació.**

Catalonia rural housing.

1. **Definir un títol pel dataset. Triar un títol que sigui descriptiu.**

Catalonia rural housing.

1. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (es necessari que aquesta descripció tingui sentit amb el títol triat).
2. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment
3. Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.
4. Agraïments. Presentar el propietari del conjunt de dades. Es necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).
5. Inspiració. Explicar per que es interessant aquest conjunt de dades i quines preguntes es pretenen respondre.
6. Llicencia. Seleccionar una d'aquestes llicencies pel dataset resultant i explicar 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

1. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.
2. Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.