Estudiants: Roger Peris Serrano i Albert Cámara Viñals

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.

Estudiants: Roger Peris Serrano i Albert Cámara Viñals

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

9) Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

10) 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.

Estudiants: Roger Peris Serrano i Albert Cámara Viñals

- 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, ...

3. Una carpeta amb el codi Python o R generat per obtenir les dades.
4. 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.

Estudiants: Roger Peris Serrano i Albert Cámara Viñals

Respostes

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

La informació s'ha recol·lectat en el context de la pandèmia originada pel Covid-19, on el sector del turisme ha patit una forta davallada del volum del negoci i donades les restriccions de mobilitat s'ha vist incrementat el turisme nacional o local, jugant un paper important el turisme rural donades les seves característiques. En aquest sentit, s'ha volgut generar un joc de dades o dataset que permeti d'una banda estudiar i/o conèixer la situació actual del turisme rural així com també per a permetre la realització d'estudis futurs per avaluar la introducció de nous agents en aquest sector o d'una nova oferta turística. Per això s'ha cercat un lloc web consolidat en aquest mercat *EscapaRural*, amb més de 13 anys d'existència i sent un dels principals cercadors de cases rurals a Espanya i Portugal. En aquest portal podem trobar gairebé 2420 referències de cases rurals a Catalunya, amb forces detalls sobre les cases, incloent el tipus de lloguer, les valoracions, preus, ubicació i altres característiques pròpies de la casa com el nombre de dormitoris o el nombre de llits. Per la qual cosa sembla un portal web adient per l'obtenció de dades referides al turisme rural a Catalunya.

2) Definir un títol pel dataset. Triar un títol que sigui descriptiu.

Fent al·lusió a un dels joc de dades, segurament més coneguts a l'àmbit de la iniciació de la ciència de dades, el *Boston Housing Dataset*, em considerat oportú anomenar el nostre dataset com a: *Catalonia Rural Housing Dataset*.

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).

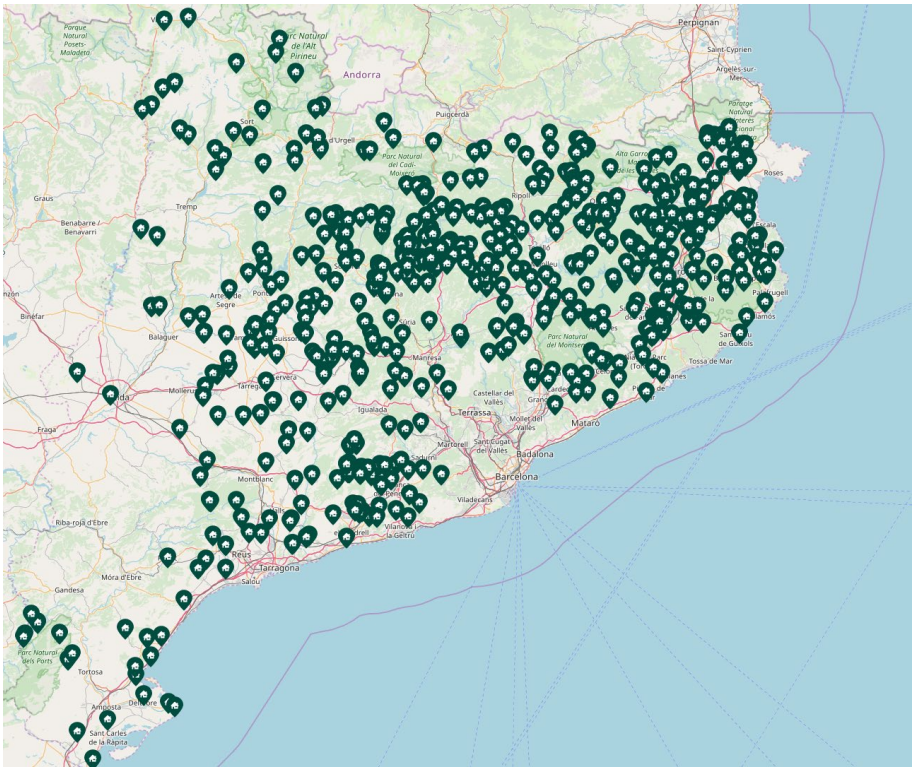
El dataset *Catalonia Rural Housing Dataset* està format pel conjunt de característiques extretes de les cases rurals de Catalunya disponibles al portal *EscapaRural*. Per a cada casa rural es pot trobar un registre amb 17 atributs, que descriuen les principals característiques de la casa, com ara el tipus de lloguer, les valoracions, el preu, la seva ubicació i altres característiques pròpies de la casa com el nombre de dormitoris o el nombre de llits.

4) Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment

Font: EscapadaRural (https://www.escapadarural.com/casa-rural/tarragona/mas-llagostera, consultat el 17-10-2020)

Estudiants: Roger Peris Serrano i Albert Cámara Viñals

I a continuació un altre imatge, que pensem que pot ajudar encara més que l'anterior a ubicar les dades presents al joc de dades.



Font: EscapadaRural (https://www.escapadarural.com/casas-rurales?l=cataluna&viz=map, consultat el 17-10-2020)

5) Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

El dataset *Catalonia Rural Housing Dataset* està format per 2415 registres, referits a cases rurals de la regió de Catalunya disponibles al portal *EscapaRural*. Per a cada casa rural es pot trobar un registre amb 17 atributs, que són:

- *url*: Direcció web original d'on s'han extret les dades, (Variable de tipus text).
- *name*: Nom de la casa rural, (Variable de tipus text).
- *town*: Nom de la població on es localitza la casa rural, (Variable de tipus text).
- *stars*: Qualificació generada a partir de les valoracions realitzades pels usuaris, (Variable de tipus text).
- *score*: Conversió de l'atribut anterior, stars, a un valor numèric, (Variable de tipus numèric).
- *reviews*: Nombre de valoracions que ha rebut la casa rural per part dels usuaris. (Variable de tipus numèric).
- *rent_type*: Tipus de lloguer disponible, per habitacions, complert o ambdues opcions, (Variable de tipus text).
- *capacity*: Nombre de persones que pot allotjar la casa rural, (Variable de tipus text).

Estudiants: Roger Peris Serrano i Albert Cámara Viñals

- *bedrooms*: Nombre d'habitacions de la casa rural, (Variable de tipus text).
- *beds*: Nombre de llits de la casa rural, (Variable de tipus text).
- *price*: Preu de lloguer aproximat per persona i nit, (Variable de tipus text).
- *longitude*: Coordenada geogràfica, longitud, de la ubicació de la casa, (Variable de tipus text).
- *latitude*: Coordenada geogràfica, latitud, de la ubicació de la casa, (Variable de tipus text).
- *street*: Nom del carrer on es localitza la casa rural, (Variable de tipus text).
- *municipality*: Nom del municipi on es localitza la casa rural, (Variable de tipus text)
- *province*: Nom de la província on es localitza la casa rural, (Variable de tipus text).
- *url_image*: Direcció web amb una imatge representativa de la casa rural, (Variable de tipus text).

Pel que fa al període de temps de les dades, podem dir que disposem de dades de cases rurals des del any 2007, que és quan va començar a operar el portal *EscapaRural*, fins al dia al qual s'ha realitzat l'extracció, 25 d'octubre del 2020. Només com apunt, cal tenir present que es tracta d'un portal dinàmic i per tant el nombre de cases rurals pot variar al llarg dels temps. De fet en el transcurs de la pràctica em pogut veure com apareixien noves cases.

Per acabar, cal dir que les dades s'han recollit en un sol procés d'extracció, realitzat el 25 d'octubre del 2020. i que ha requerit uns 2 minuts.

6) Agraïments. Presentar el propietari del conjunt de dades. Es necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).

El propietari del conjunt de dades és *ESCAPADA RURAL SERVICIOS PARA PROPIETARIOS SL,* per a més informació es pot visitar el seu lloc web *EscapadaRural*. És tracta d'un dels principals cercadors especialitzat al sector de les cases rurals que cobreix l'estat Espanyol i Portugal, amb més de 1.400.000 usuaris registrats.

Respecte a la recerca en l'àmbit del turisme rural les referències podem dir que no són gaire extenses, encara que si que és cert que és més fàcil trobar estudis relacionats amb el desenvolupament en l'entorn rural, on segurament el turisme i juga un paper important, per exemple podem destacar les següents referències:

- The Housing Assistance Council (HAC), The Rural Data Portal (http://www.ruraldataportal.org/, consultat el 23-10-2020)
- The world Bank, Agriculture & Rural Development (https://data.worldbank.org/topic/agriculture-and-rural-development, consultat el 23-10-2020)

Tot i així cal destacar la iniciativa desenvolupada pels propis propietaris del portal web *EscapadaRural* conjuntament amb la *EUHT CETT-UB* i *Netquest* anomenada ***'El Observatorio del Turismo* Rural'**. La finalitat d'aquesta iniciativa és generar coneixement i aportar informació de valor sobre el sector del turisme rural a Espanya tant des de l'àmbit de la oferta ("propietaris") como des de la demanda ("viatgers"). Per a més informació es pot consultar la següent referència:

- Observatorio del Turismo Rural (http://www.escapadarural.com/observatorio/, consultat el 23-10-2020)

Estudiants: Roger Peris Serrano i Albert Cámara Viñals

7) Inspiració. Explicar per que es interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

El dataset *Catalonia Rural Housing Dataset* és interessant ja que aporta una visió aproximada de l'oferta del sector del turisme rural a Catalunya. Gràcies a aquestes dades es podrien tractar preguntes com:

- Quin patrons de ubicació segueixen les cases de turisme rural? Les puntuacions i preus depenen de la ubicació?
- Es pot crear un model que estimi el preu de l'allotjament en base a les característiques de la casa, com el nombre de llits, la ubicació, etc?
- Hi ha alguna similitud entre les cases que es troben en una mateixa zona geogràfica?
- Quines són les característiques de les cases que fan que aquestes siguin més ben valorades o rebin més recomanacions.
- Donada una casa particular, amb les seves característiques i ubicació, seria una bona opció de negoci convertir-la en un allotjament turístic.
- Seria possible estendre el conjunt de dades a un conjunt de dades visuals (imatges), que permetés predir si donada una imatge d'una casa qualsevol representa un allotjament rural o no.

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

Per al joc de dades s'ha escollit una llicència Creative Commons, CC BY-NC-ND 4.0. Donat que el propietari del portal i de les dades presents és *EscapadaRural Servicios para Propietarios S.L., amb CIF B-64823032 i amb domicili fiscal al C/Bailén 50, local 7, 08009 - Barcelona* i que a les condicions d'ús del seu portal expressà l'ús que es poden fer del seu contingut, s'ha escollit una llicència que faciliti el seu compliment. La llicència Creative Commons, CC BY-NC-ND 4.0 permet la distribució del joc de dades però alhora assegura l'atribució dels autors, que no es faci un us comercial d'aquest o és limita la distribució d'obres derivades d'aquest. I d'aquesta manera ens permet limitar l'ús del dataset, preservant al propietari original de les dades.

9) Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

El codi font, desenvolupat en Python, es pot consular al següent repositori de gitHub:

https://github.com/rugeps/CataloniaRuralHousingScraper/

10) Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.

El dataset es troba publicat a Zenodo (**DOI:** 10.5281/zenodo.4164777), i es pot consultar a:

https://zenodo.org/record/4164777#.X50LQq7OHJU

Estudiants: Roger Peris Serrano i Albert Cámara Viñals

Contribucions al treball

Contribucions	Signa
Recerca prèvia	RPS, ACV
Redacció de les respostes	RPS, ACV
Desenvolupament codi	RPS, ACV