Pràctica 1

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

   Per realitzar la pràctica s'ha escollit el lloguer d'immobles de Barcelona ciutat. A Internet hi han diversos portals de referència de publicació d'immobles i en el moment de cercar la pàgina web, el criteri de selecció va ser:

   "Quin dels portals immobiliaris a Internet tenen major número de publicacions d'habitatges en lloguer a la capital de Barcelona?"

   El resultat va ser que habitaclia.com contenia el major número de publicacions d'habitatges en lloguer de Barcelona, sent aproximadament un 6900 habitatges en lloguer.

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

   Habitatges en lloguer a Barcelona Abril 2019

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 triat).**

   El dataset és un recull dels últims habitatges en lloguer a Barcelona que s'han publicat a través del portal web habitaclia.com durant el mes d'abril del 2019.
   Les dades que s'han extret poden conduir a la realització d'un anàlisi sobre la tendència de la oferta d'habitatges de lloguer a Barcelona i de les seves característiques en relació al preu.

   Les dades extretes corresponen al tipus d'habitatge (per exemple si és un pis, apartament o estudi), metres quadrats, número d'habitacions i lavabos, preu de lloguer i una breu descripció personalitzada de l'anunciant.

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.**

Les dades del dataset generat pertanyen als últims habitatges de lloguer publicats al portal de habitaclia.com en el mes d'abril de l'any 2019.

Les dades s'han recollit a través d'un script generat amb *Python* i juntament amb les llibreries de *BeautifulSoup* i *Panda*. En el moment de l'extracció, Habitaclia tenia aproximadament 6900 habitatges publicats, desgranats en 450 pàgines, és a dir 15 publicacions per pàgina.

Per això, ha estat necessari un bucle que recorri les 450 pàgines. Dintre de cada pàgina l'script captura les etiquetes *html* que contenen la informació rellevant i les emmagatzema en variables que seran emmagatzemades en una llista. A través de *Panda*, es crea un *DataFrame* per poder extreure les dades en format CSV. A més, mitjançant expressions regulars s'ha pogut extreure amb més precisió les dades necessàries.

Val a dir, que abans de realitzar *l'sraping*, s'ha comprovat que el fitxer de robots.txt no contenia cap restricció que afectés a l'extracció de dades. No obstant això, per evitar *banejos* s'ha camuflat el *User Agent* de l'script simulant ser un navegador de Mozilla.

El dataset inclou els següents camps:

Camp	Descripció
Barri	Barri on pertany l'habitatge de lloguer publicat
Habitacions	Número d'habitacions que té l'habitatge
Lavabo	Número de lavabos que té l'habitatge
Metres	Número de metres quadrats que té l'habitatge
Preu	Preu que és demanar per llogar l'habitatge
TipusImmoble	Tipus d'immoble que s'està oferint, que pot ser tipus pis, apartament, dúplex, àtic, casa, loft, estudi, xalet, entre d'altres. És una variable categòrica
Descripció	Breu resum de l'habitatge publicat

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

   Habitaclia és un portal immobiliari digital amb 14 anys d'experiència, el qual s'encarrega de posar en contacte a compradors i venedors, arrendataris i arrendadors, tant a nivell particular com professional. Aquest servei l'ofereix sense cobrar cap tipus de comissió d'intermediació.

   Agraïm el seu treball per estructura i recollir les dades atès que gràcies al seu portal s'ha pogut dur a terme l'objectiu de la pràctica.

7. **Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.**

   El conjunt de dades recollit permet analitzar amb més detall quina és la tendència dels habitatges en lloguer i quina variació ha patit en el mes d'abril de l'any 2019. Concretament, les dades del dataset permetran respondre les següents preguntes:

   - Quin tipus d'immoble és el més publicat?
   - Quina relació de preus per metre quadrat hi ha a cada barri?
   - Donat una franja econòmica i tipus d'immoble, a quin barri aconseguiríem major número de prestacions? És a dir, major nombre de metres quadrats, habitacions i lavabos.
   - Quin barri cotitzen els lloguers més cars / barats?

Més enllà de les preguntes proposades, si el periode temporal del data set fos més llarg també es podria respondre a tipus de preguntes predictives, descriptives, exploratòries o inferencials.

8. **Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:**

L'objectiu pel qual s'ha creat el dataset és per realitzar estudis sobre l'evolució del mercat immobiliari de lloguer a Barcelona i així poder detectar tendències, per aquest motiu, s'ha escollit la llicència Released Under CC BY-NC-SA 4.0 License[1], atès que estem d'acord en què altres persones utilitzin el nostre dataset per realitzar les seves investigacions però sense motivacions comercials.

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

   Codi adjunt al fitxer **habitaclia.py**
   https://github.com/jlchanjlchan/practica1/blob/master/habitaclia.py

10. **Dataset. Presentar el dataset en format CSV**

    habitaclia_pisos_barcelona_abril_2019.csv

Taula de contribucions del treball.

Contribucions	Signa
Recerca prèvia	earinos, jlchan
Redacció de les respostes	earinos, jlchan
Desenvolupament codi	earinos, jlchan

[1] https://creativecommons.org/licenses/by-nc-sa/4.0/