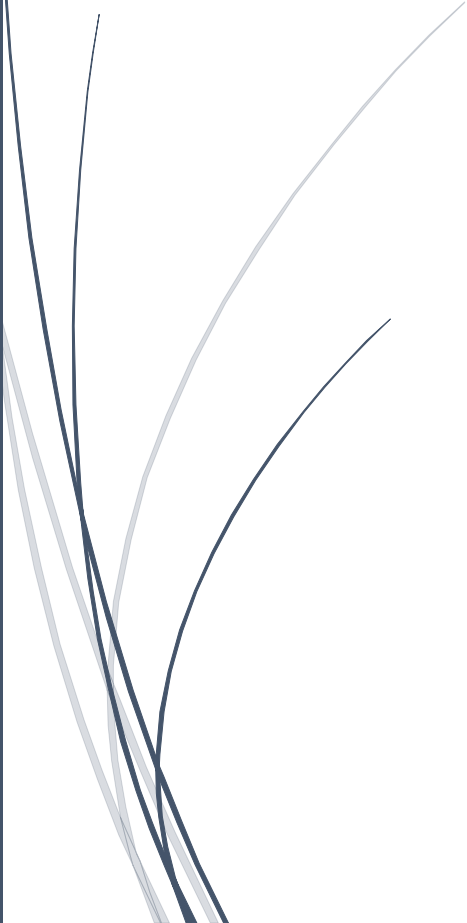


A dark blue vertical bar runs down the left side of the page. A blue arrow points to the right from this bar, containing the date.

14-4-2019

PRAC 1 -Tipologia i Cicle de Dades

WEB SCRAPING

Several thin, curved lines in dark blue and light grey originate from the bottom left and sweep upwards and to the right.

Joan Maggi Gómez, Carles Maggi Gómez
UOC – TICD -PAC1

Pràctica de Tipologia i Cicle de Vida de les Dades

Descripció de la Pràctica a realitzar	2
1. Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació	2
Consideracions.txt	2
urls.txt	3
2. Definir un títol pel dataset. Triar un títol que sigui descriptiu.	3
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).	3
4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment	3
5. Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.	4
HotelsBarcelonaBooking	4
ComentarisXHotelsBarcelonaBooking	4
CategoriesXComentariBooking	5
6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).	5
7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.	5
8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:	5
9. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.	5
10. Dataset. Presentar el dataset en format CSV	6
CONTRIBUCIONS:	6

Descripció de la Pràctica a realitzar

L'objectiu d'aquesta practica és la creació d'un dataset a partir de les dades contingudes en un enllaç de la web **www.booking.com** , concretament del hotels de Barcelona, per tal de limitar la mida del resultats.

A partir del número d'hotels demanats es creen aquest fitxers :

HotelsBarcelonaBooking.csv	Dataset que inclou la llista d'hotels trobats.
ComentariosXHotelsBarcelonaBooking.csv	Dataset de comentaris per hotel de hotels.csv
CategoriesXComentariBooking.csv	Categories associades a un comentari d'un hotel, una fila de comentarios.csv.

També es creen aquest dos fitxers .txt amb la informació sobre el context.

Connsideracions.txt: Contingut de totes les dades relacionades amb el context.

urls.txt: Urls dels sitemap, per si cal tractar-les més endavant.

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

Primerament, en el propi codi de l'aplicació, hem realitzat extracció de la informació de context que hem consultat per determinar la viabilitat o no de fer l'scrapping. Primer ens vam plantejar la idea d'extreure informació de comentaris d'hotels de Barcelon del site de booking. Vam fer una ullada a la informació de robots.txt i vam veure que el link que vam trobar per obtenir la informació d'hotels de Barcelona no era territori prohibit. En un principi no trobàvem la manera de poder consultar els comentaris (s'havia d'emular una navegació i no fer un scarping). Consultant el sitemap ens vam adonar que hi havia una forma lògica d'enumerar la url per accedir a la llista de comentaris per hotel i per tant eren accessibles a nivell d'scrapeig. Finalment, veient la informació que teniem vam considerar oportú estructurar-la en tres csv.

Al Github, a la carpeta de FitxersContext trobarem els següents:

Consideracions.txt

En el fitxer de text **consideracions.txt** estan totes les dades relatives el context que envolta aquest url.

Com a dades més significatives tenin :

Propietari :

```
{ "domain_name": [ "BOOKING.COM", "booking.com" ], "registrar":  
"MarkMonitor, Inc.", ...}
```

Grandaria :

Aproximadamente 53.400.000 resultados

Tecnologia :

```
{'web-servers': ['Nginx'], 'javascript-frameworks': ['Prototype', 'RequireJS', 'jQuery']}
```

Robots.txt :

El contingut del fitxer estar dins del fitxer *****consideracions.txt*****

urls.txt

El contingut del sitemap està dins del fitxer *****robots.txt*****, però hem fet el fitxer *****urls.txt***** a part on només hi ha les urls del fitxers '.xml' per si cal tractar-les després.

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

HotelsBarcelonaBooking

ComentarisXHotelsBarcelonaBooking

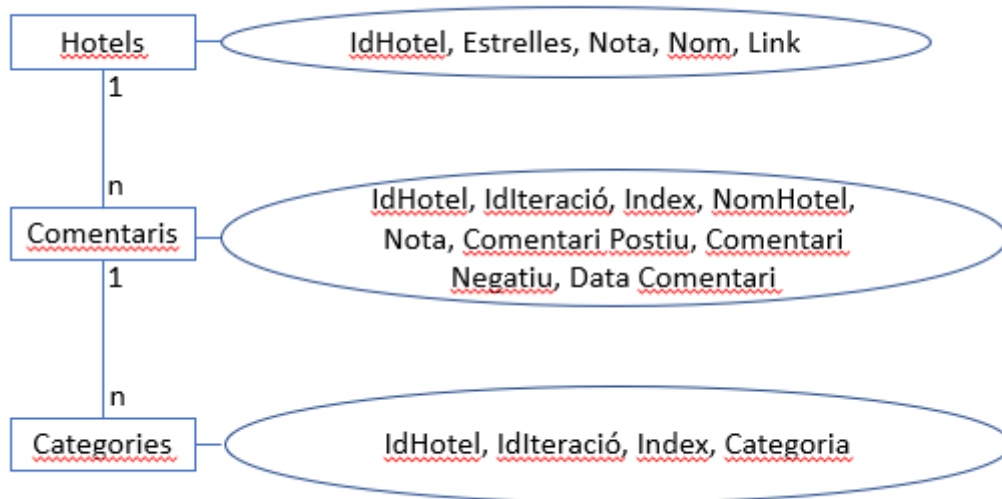
CategoriesXComentariBooking

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).

Hem creat 3 datasets donada la informació que consideràvem interessant descarregar i la naturalesa de la mateixa

Conceptualment hem creat el dataset de *HotelsBarcelonaBooking* que estreu informació sobre els Hotels de Barcelona, el dataset *ComentarisXHotelsBarcelonaBooking* que extreu tots els comentaris que hem trobat per cada hotel i finalment el Dataset *CategoriesXComentariBooking* que obté la categorització per cadascun dels comentaris.

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.

Agafem a partir d'un enllaç de cerca genèric (que no té una relació temporal) el conjunt de hotels de Barcelona, i d'aquí recollim el conjunt de comentaris que estan vinculats a aquest hotels accessibles (creiem que no tots són accessibles) i per cada comentari agafem les categories que el categoritzem i creem un fitxer per tal d'establir un "model relacional" entre datasets.

HotelsBarcelonaBooking

IdHotel	: BigInt	que identifica de manera unívoca el hotel
Estrelles	: Int	Número d'estrelles, en cas que n'hi hagi
Nota	: Float	mitja de l'hotel
Nom	: String	Nom de l'hotel
Link	: String	Url de la pàgina de l'hotel

ComentariosXHotelsBarcelonaBooking

IdHotel	: BigInt	Identifica de manera unívoca el hotel
IdIteració	: Int	Primera part que identifica un comentari (iteració llista comentaris)
Index	: Int*	Segon part d'identificació de comentari (número de comentari dins la iteració)
NomHotel	: String	Nom de l'hotel
Nota	: Float	Nota que en qualifica el comentari de l'hotel
Comentari Positiu	: Text	Comentari positiu si n'hi ha
Comentari Negatiu	: Text	Comentari negatiu si n'hi ha
Data Comentari	: Date	Data enregistrada del comentari

CategoriesXComentariBooking

Un comentari pot estar categoritzat per una o més categories.

IdHotel	: BigInt	Identifica de manera unívoca el hotel
IdIteració	: Int	Primera part que identifica un comentari (iteració llista comentaris)
Index	: Int	Segon part d'identificació de comentari (número de comentari dins la iteració)
Categoria	: Text	Categoria amb la que s'ha categoritzat el comentari

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

Agraïm a Booking.com poder scrapejar aquests datasets.

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

Farem una explicació en funció de les persones que poden tenir un interès respecte l'activitat comercial en qüestió :

Propietari hotel: Poder fer un seguiment, planificant un scraping diari, de com evoluciona la nota del seu hotel i la de les seues competidors.

Client Hotel: Comparar en funció de la categorització dels comentaris aquells hotels que tinguin una millor nota.

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

Triem la llicència **Released Under CC0: Public Domain License** perquè de la mateixa manera que nosaltres hem obtingut les dades en obert, nosaltres pensem que hem de seguir la mateixa filosofia, deixar-ho en obert, i que pugui ser emprat per tecers

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

En el directori src podem trobar el codi generat :

Main.py: Programa principal amb un paràmetre. Cal executar-lo : `main.py --nhotels 30`

Comment.py: Classe que recull i guarda els comentaris al data set
ComentarisXHotelsBarcelonaBookin

Hotels.py: Classe que recull i guarda les dades genèriques de l'Hotel

10. Dataset. Presentar el dataset en format CSV

En el directori CSV podrem trobar els tres datasets :

HotelsBarcelonaBooking .CSV

CategoriesXComentariBooking.csv

ComentarisXHotelsBarcelonaBooking.csv

CONTRIBUCIONS:

Recerca Prèvia	Carles Maggi, Joan Maggi
Redacció de les Respostes	Carles Maggi, Joan Maggi
Desenvolupament del Codi	Carles Maggi, Joan Maggi