Tipologia i cicle de vida de les dades
PRAC1

Josep Alòs Pascual, Daniel Galan Vilella

November 11, 2019

1 Descripció del projecte

1.1 Context i inspiració

Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

Ens trobem en els inicis de la nostra vida laboral, on el nostre poder adquisitiu és superior que en la nostra època d'estudiants de grau. Cansats de freqüentar els mateixos bars i restaurants autoanomenats 'low cost', ens hem començat a interessar en restaurants d'un altre tipus.

Utilitzant la pàgina de TripAdvisor hem anat explorant nous restaurants. Tot i això, com a bons informàtics, no som amics de les tasques manuals repetitives. Aprofitant aquesta pràctica, hem decidit que podem solucionar aquest problema mitjançant les tècniques d'analisi de dades, que estem aprenent en aquest master, de forma que ens pugui recomanar restaurants nous sense haver de pensar nosaltres on anar.

TripAdvisor és l'opció ideal per obtenir dades sobre restaurants ja que té informació sobre (1) la ubicació, (2) el preu aproximat, (3) comentaris i puntuacions, (4) i detalls de la seva cuina.

1.2 Descripció del dataset

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

Restaurants Lleida

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).

Aquest dataset conté dades relatives als restaurants de Lleida. Aquesta informació fa referència a dos blocs: (1) dades informatives del restaurant (horaris, ubicació, telèfon...), i (2) dades de qualitat (puntuacions i comentaris). En l'apartat 1.4 es mostra en detall quines dades es guarden.

Aquestes dades s'han extret de TripAdvisor, i per tant no són dades proporcionades pel restaurant directament sinó que estan basades en les aportacions dels clients, fet que pot aportar soroll però a la vegada informació més imparcial.

1.3 Representació gràfica

Presentar una imatge o esquema que identifiqui el dataset visualment En la figura 1 podem veure la distribució de les puntuacions dels diferents restaurants del dataset.

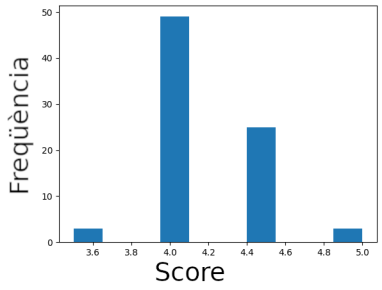

Figure 1: Distribució de les puntuacions dels restaurants

1.4 Contingut

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

El dataset estarà separat en dos fitxers. En el primer s'inclourà la informació referent als restaurants, i el segon sobre els comentaris dels usuaris. S'ha decidit separar els dos conjunts per tal de mantenir la consistència dins d'un mateix fitxer i evitar tenir dades duplicades.

Dataset 1: restaurants Aquest dataset conté informació sobre els restaurants. Les dades i el seu tipus es mostren en la taula 1.

Dataset 2: comentaris Aquest dataset contindrà la informació sobre els comentaris d'un restaurant. Les dades guardades i el seu tipus es mostren en la taula 1.

Aquestes dades s'han col·leccionat al novembre de 2019. S'han agafat dades de comentaris de l'últim any per reduïr la quantitat d'informació amb la que es treballava, però el codi proporciona un paràmetre per generar datasets més amb altres dates.

Referent als restaurants, TripAdvisor proporciona dades en una llista ordenada per rellevància. Per tant, s'ha limitat als 80 primers restaurants d'aquesta llista, pel mateix motiu esmentat anteriorment. Aquest valor també es parametritzable.

Per tal de recollir aquestes dades, s'ha fet un *web scrapping* de la pàgina de la llista de restaurants de Lleida, i per cada restaurant s'ha agafat l'enllaç de

Nom	Tipus	Descripció
Nom	string	Nom del restaurant.
Adreça	string	Direcció del restaurant.
Telèfon	string	Telèfon del restaurant.
Puntuació	float	La puntuació que té el restaurant. Aquesta puntuació és la mitjana de les puntuacions de menjar, servei, preu, i atmosfera. El seu valor va d'1 a 5, i pot ser un nombre decimal.
Puntuació menjar	float	
Puntuació servei	float	
Puntuació preu	float	
Puntuació atmòsfera	float	
Rang de preu	string	El rang de preus que es troben els plats.
Detalls cuina	string	Detalls de la cuina d'aquest restaurant (estil, plats...).
Certificat excelència	boolean	Si el restaurant té un certificat d'excelència.

Table 1: Dades de restaurants

la seva pàgina de detalls i se n'ha analitzat el contingut. Per últim, aquestes dades s'han guardat en dos fitxers *csv*.

1.5 Agraïments

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

El propietari d'aquestes dades és TripAdvisor, Inc.

1.6 Llicència

Seleccionar una d'aquestes llicències 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)*

Nom	Tipus	Descripció
Restaurant	string	Identificador del restaurant
Usuari	string	L'usuari que ha escrit el comentari
Titol	string	El titol del comentari
Text	string	El contingut del comentari
Data de visita	date	La data de la visita (mes i any)
Puntuació	int	La puntuació que ha donat aquest usuari
Resposta	string	Resposta del restaurant (pot existir o no)

Table 2: Dades dels comentaris

- *Unknown License*

Aquestes dades són propietat de TripAdvisor, Inc. Segons la seva pàgina web, qualsevol ús d'aquestes dades s'ha de consultar amb els seus creadors:

> Please note that any use of TripAdvisor content, with the exception of the self-service widgets, must be approved by TripAdvisor. For more information on the use of our self-service widgets, please see our Widget Terms of Use for specific requirements. For complete instructions about how to find and use self-service widgets, please refer to our Insights guide.

Tot i això, sota la clàusula de *Fair Use* del Copyright, entenem que aquestes dades es poden fer servir de forma educativa durant un procès d'aprenentatge. Per qualsevol altre ús de les dades s'ha de contactar amb els seus propietaris. Per tant, aquest dataset està protegit sota una llicència de Copyright, propietat de TripAdvisor, Inc.

2 Taula de contribucions

Les contribucions a aquest projecte es poden veure a la taula 3.

Contribucions	Signa
Recerca prèvia	DGV, JAP
Redacció de les respostes	DGV, JAP
Desenvolupament del codi	DGV, JAP

Table 3: Taula de contribucions