PRÀCTICA 1

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

El lloc web escollit per a la recol·lecció d'informació és TrustPilot, en la seva versió espanyola (https://es.trustpilot.com/). Aquesta plataforma digital, com ells mateixos expliquen, va néixer per posar en contacte a empreses i consumidors per a fomentar la confiança mútua i estimular la col·laboració. Consta de multitud de negocis dels quals els usuaris poden donar la seva opinió i una puntuació basada en la seva experiència amb l'empresa o negoci en qüestió.

Les empreses o negocis estan dividits en diferents categories per les quals es pot navegar. A nivell d'usuari, això permet trobar aquelles empreses més ben valorades dins d'un àmbit concret i llegir què n'opinen els altres consumidors. També es poden aplicar diferents filtres per trobar més específicament aquelles empreses d'interès per a l'usuari com el nombre d'opinions mínim, el període de temps de les dades o l'estat de l'empresa (si està invitant activament a opinar i si té el perfil reclamat o no). Per a les empreses, suposa un *feedback* de primera mà dels seus serveis, cosa que pot ajudar a millorar-ne la qualitat. També poden optar per oferir informació útil com l'adreça del negoci i formes de posar-se en contacte amb ells.

TrustPilot és una plataforma transparent, oberta a tot el món i que es pot utilitzar de manera gratuïta. Consta de més de 120 milions d'opinions, més de 520.000 empreses ressenyades, aproximadament 6,9 mil milions d'impressions mensuals de les seves TrustBoxes i forma part del top 1% d'empreses més populars del món segons el rànquing Alexa (dades del 31 de desembre del 2020). Per tant, la informació extreta d'aquest lloc web és pot considerar molt rellevant, gràcies a la seva àmplia base d'usuaris i l'impacte que pot causar en els consumidors i les empreses.

Així doncs, en aquest lloc web proporciona informació sobre diferents negocis de tot tipus (vint categories diferents i moltes subcategories) i la popularitat d'aquests entre els consumidors.

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

Puntuació TrustPilot de diferents negocis amb activitat a Espanya per categories.

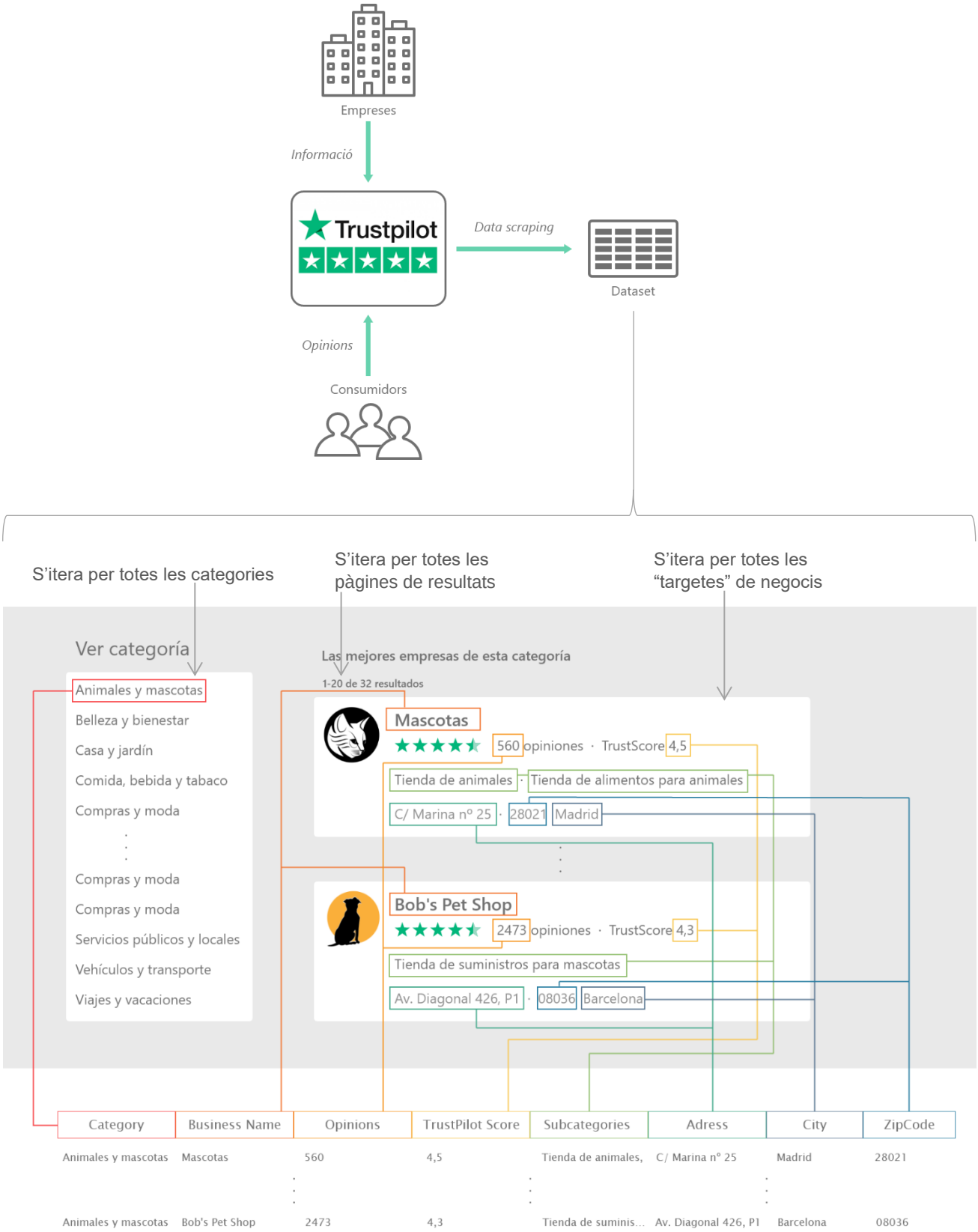
3. **Descripció del dataset.** Desenvolupar una descripció breu del conjunt de dades que s'ha extret. És necessari que aquesta descripció tingui sentit amb el títol escollit.

El conjunt de dades consta de diferents negocis o empreses amb activitat a Espanya les quals estan organitzades en un total de vint categories en la plataforma TrustPilot. Per a cada negoci o empresa es proporcionen el nombre d'opinions rebudes i la seva puntuació mitjana. També proporciona informació sobre les subcategories a les que pertany cada negoci i la seva adreça, ciutat i codi postal.

Aquest *dataset* té un total de 8 camps o atributs i 1686 files a dia 5 de novembre del 2021.

4. **Representació gràfica.** Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.

El següent diagrama mostra de forma visual el projecte, indicant d'on s'ha extret la informació per a cada camp del *dataset* i com s'ha generat aquest:



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

El conjunt de dades compta amb els següents 8 camps:

- Category: És la categoria principal o més general a la qual pertany el negoci o l'empresa. Hi ha un total de vint valors diferents.

- Business Name: És el nom del negoci o empresa. També poden ser pàgines web.

- Opinions: És el nombre total d'opinions que ha rebut el negoci o empresa. En aquest *dataset* només s'han inclòs aquells casos amb un mínim de 25 opinions.

- TrustPilot Score: És el valor mitjà de la puntuació que han posat els diferents usuaris al negoci o empresa en concret. Està en una escala de l'1 al 5.

- Subcategories: Són les diferents subcategories a les que pertany el negoci o empresa. Són categories que defineixen més concretament el negoci. N'hi pot haver una o més i en el cas que n'hi hagi més d'una estan separades per comes.

- Adress: És l'adreça de l'empresa o el negoci, si n'hi ha. No segueix un format unificat o estàndard, ja que cada empresa ha introduït les seves dades pròpies. No totes les files inclouen aquesta informació, per indicar aquesta absència de dades s'ha usat el valor buit o None.

- City: És la ciutat de l'empresa o negoci, si n'hi ha. No totes les files inclouen aquesta informació, per indicar aquesta absència de dades s'ha usat el valor buit o None.

- ZipCode: És el codi postal de l'empresa o negoci, si n'hi ha. No totes les files inclouen aquesta informació, per indicar aquesta absència de dades s'ha usat el valor buit o None.

Totes les dades incloses en el conjunt de dades corresponen als últims 12 mesos, ja que és el filtre que s'ha aplicat en el moment de fer-ne la recol·lecció. En aquest cas anirien concretament des del 5/11/2020 fins al 5/11/2021.

El conjunt de dades s'ha generat per mitjà d'un script amb Python utilitzant principalment les llibreries requests i BeautifulSoup.

El primer pas ha sigut el d'investigar l'estructura del lloc web per veure on estaven les parts rellevants de la informació que es volia extreure. Un cop fet això, s'ha anat al punt inicial per on començar el web scraping, que ha sigut el lloc web https://es.trustpilot.com/categories. Mitjançant requests s'ha obtingut la resposta del lloc web, i amb BeautifulSoup s'ha convertit en un objecte pel qual s'ha iterat per totes les categories principals per obtenir-ne el nom i l'enllaç.

Posteriorment, s'ha fet el mateix per obtenir l'objecte *soup* de cada un dels links de les categories. En cada categoria s'ha comprovat el nombre d'empreses que conté i s'ha extret la informació relativa a cada una d'aquestes, que està inclosa en un tipus de "targetes" on hi ha el nom de l'empresa, el nombre d'opinions, la puntuació TrustPilot, les subcategories i opcionalment l'adreça, la ciutat i el codi postal. En cada pàgina,

només es mostren un total de 20 empreses, per tant, per a les categories que en tenen més de 20 s'ha hagut d'anar iterant per les següents pàgines fins a obtenir la informació de totes les "targetes".

Per a cada empresa s'ha creat una fila, amb les dades ja mencionades, i s'ha afegit en un DataFrame de la llibreria Pandas.

Finalment, un cop s'ha iterat per totes les categories i les seves múltiples pàgines, s'ha convertit el DataFrame en un document CSV.

El temps d'execució aproximat de tots aquests passos ha sigut d'uns 6 minuts i 30 segons.

6. **Agraïments.** Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-n'hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte.

Agrair a TrustPilot la creació d'una plataforma oberta, transparent i accessible per a tot el món de forma gratuïta de la qual es poden obtenir dades útils per a diferents anàlisis.

Els principals anàlisis que es solen fer en aquest tipus de pàgines web són els de les opinions dels usuaris per a diferents empreses o productes. És a dir, seleccionen una plataforma on hi hagi opinions d'usuaris com podria ser TrustPilot, SiteJabber o Amazon, per exemple, i posteriorment seleccionen una empresa o producte en concret del qual recol·lecten totes les opinions que ha rebut, per a posteriorment fer un anàlisi de sentiments o de polaritat.

Dos exemples d'aquestes eines són els que es poden trobar en els següents repositoris:

https://github.com/toxtli/company-reputation-reviews-ratings-extractor-scraper

https://github.com/hakimkhalafi/trustpilot-scraper

En el primer cas l'autor Carlos Toxtli ha creat un script amb Python, mitjançant les llibreries requests i BeautifulSoup per a l'extracció de les opinions publicades en els lloc web SiteJabber i ConsumerAffairs amb l'opció d'afegir arguments per a reduir la cerca a les puntuacions (entre 1 i 5) i les companyies desitjades.

En el segon cas l'autor Hakim Khalafi ha fet servir l'entorn de Jupyter Notebooks i també ha fet servir Python i la llibreria requests per a recollir la informació de la plataforma TrustPilot i generar un arxiu CSV amb l'objectiu de fer tasques de classificació de textos posteriorment.

Els passos que s'han seguit per actuar d'acord amb els principis ètics i legals han estat en primer lloc revisar les condicions d'us i les polítiques de privacitat de la plataforma TrustPilot. Pel que fa al web scraping, s'han afegit 5 segons de pausa entre requests de les diferents categories, i entre 1 i 3 segons de pausa entre les diferents pàgines dins de cada categoria per a evitar spammejar la plataforma amb requests. També s'ha escollit una llicència respectuosa amb els drets d'autor.

7. **Inspiració.** Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l'apartat 6.

Tot i que s'han trobat molts anàlisis anteriors sobre web scraping de les opinions de plataformes com TrustPilot, no n'hi ha tants que es centrin en fer web scraping entre totes les empreses o negocis i les seves puntuacions, enlloc del text de les opinions d'una empresa en concret.

Tot i això, crec que la informació que pot aportar aquest conjunt de dades és molt interessant i pot permetre fer anàlisis de la competència, així com descobrir noves oportunitats de negoci.

El conjunt de dades pot respondre a un seguit de preguntes com les següents:

- Quin tipus de negocis reben més opinions.
- Quines són les categories que reben més opinions.
- El tipus de negocis o empreses més ben valorats.
- Dins d'una categoria en concret quines són les subcategories en que hi ha més opinions, i en quines hi ha més bones (o dolentes) puntuacions.
- Hi ha més opinions en negocis físics o en negocis online (llocs web), i quins són valorats més positivament.
- El nombre d'empreses d'una categoria en una ciutat en concret.

També es podrien crear mapes amb els negocis que inclouen la informació de l'adreça per veure si hi ha alguns patrons entre la ubicació dels negocis i la seva puntuació, per exemple.

Com es pot observar, el *dataset* permet respondre moltes preguntes diverses depenent dels objectius que tingui l'usuari que el vulgui fer servir.

Aquesta informació obtinguda amb el conjunt de dades difereix una mica de les anàlisis anteriors, en que les preguntes que pretenien respondre es centraven més en un negoci en concret, com saber els aspectes que més agraden d'un negoci i els que menys, per exemple. Si en els anàlisis anteriors es fes la comparació de diferents negocis, també es podria arribar a resultats similars, que segurament serien més acurats gràcies a la inclusió dels textos de les opinions, però també seria un procés molt més costós.

8. **Llicència.** Seleccionar una d'aquestes llicències pel dataset resultant i justificar 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

La llicència escollida ha estat Reconeixement-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0) ja que tot i que algunes parts del conjunt de dades són de domini públic com els noms de les empreses i les seves adreces, hi ha altres parts que són propietat de la plataforma TrustPilot, com la puntuació mitjana i el nombre

d’opinions rebudes i en les condicions d’ús de la plataforma s’indica que no es pot fer servir el contingut per a fins comercials tret que s'arribi a un acord per escrit amb ells, o que ho permeti la legislació obligatòria aplicable.

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

El *dataset* s’ha generat mitjançant Python, es pot trobar al repositori GitHub de l’enllaç següent:

https://github.com/aidabg/prac1

10. **Dataset.** Publicar el dataset obtingut en format CSV a Zenodo amb una breu descripció. Obtenir i adjuntar l’enllaç del DOI.

Enllaç del DOI:

https://doi.org/10.5281/zenodo.5647693

Contribucions	***Signatura***
Investigació prèvia	A.B.G.
Redacció de les respostes	A.B.G.
Desenvolupament del codi	A.B.G.