Master en Ciència de Dades
M2.951 - Tipologia i cicle de vida de les dades

Pràctica 1: Web Scraping.

Dataset:

"Eleccions locals Barcelona. Distribució i evolució del vot per Seccions Censals 2011-2019"

Alumne: Josep Maria Sabaté Ibarz

Barcelona, Novembre del 2019

Índex

1. Context.

(Explicar en quin context s'ha recol·lectat la informació. Explicar perquè el lloc web triat proporciona aquesta informació)

Es pretén realitzar un estudi del resultat electoral de les últimes eleccions locals a Barcelona (25 de maig del 2019) així com l'evolució del sentit del vot comparant aquests resultats amb els d'anteriors eleccions (22 de maig del 2011 i 24 de maig del 2015).

La web de l'Ajuntament de Barcelona, departament d'Estadística i Difusió de Dades, (https://www.bcn.cat/estadistica/catala/index.htm), proporciona aquestes dades.

Es realitza aquest treball de web scraping sobre la web citada per extreure la informació objecte d'estudi: vots obtinguts per cada candidatura a cada Districte, Barri, Àrea Estadística Bàsica i Secció Censal de la ciutat, així com el fitxer descriptiu de la llista de candidatures. Es recuperen, així mateix, gràfics al·lusius als resultats de l'any 2019.

Com a avaluació preliminar es realitza:

1.1. Fitxer robots.txt

D'una primera anàlisi manual es planifica realizar web scraping sobre pàgines del tipus https://www.bcn.cat/estadistica/catala/dades/telec/loc/.../… com ara:

Pàgines de resultats de les eleccions dels anys 2011, 2015, 2019:
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc19/t310.htm
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc15/cloc1599.htm
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc11/cloc1199.htm

Pàgines amb la relació de candidatures:
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc19/t24.htm
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc15/cloc1517.htm
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc11/cloc1117.htm

Pàgines amb mapes amb els resultats electorals per barris i seccions censals:
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc19/t311.htm
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc19/t312.htm
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc19/t313.htm
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc19/t314.htm

El fitxer *robots.txt* està publicat a la mateixa web (http://ajuntament.barcelona.cat/robots.txt) i no assenyala restriccions per a l'accés a través de tècniques de scraping a les pàgines citades.

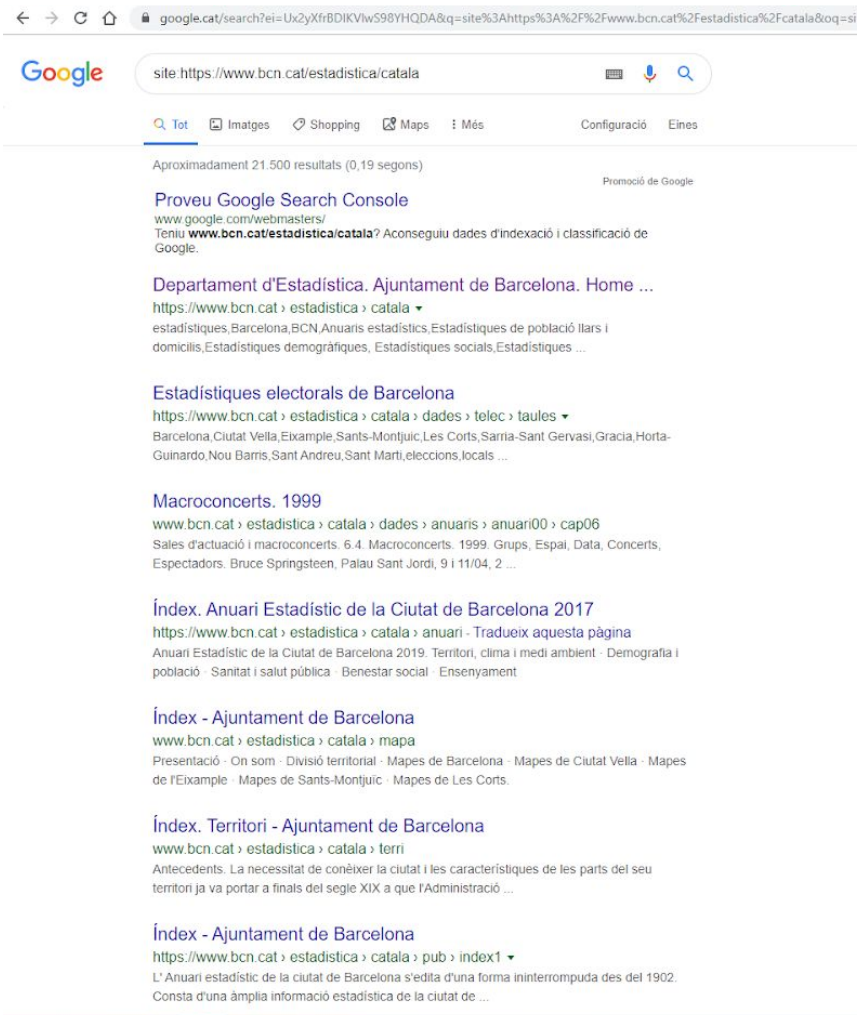
barcelona.cat/robots.txt

```
# FILE: http://www.barcelona.cat/robots.txt (& lameva)  v15-5-19

User-agent: *
Disallow: */resources/
Disallow: */brand/
Allow: /resources/hu/
Disallow: */download/
Disallow: */thumbnails/
Disallow: */cgi-bin/
Disallow: */includes/
Disallow: */misc/
Disallow: */modules/
Disallow: */profiles/
Disallow: */scripts/
Disallow: */themes/
Disallow: */CHANGELOG.txt
Disallow: */cron.php
Disallow: */INSTALL.mysql.txt
Disallow: */INSTALL.pgsql.txt
Disallow: */INSTALL.sqlite.txt
Disallow: */install.php
Disallow: */INSTALL.txt
Disallow: */LICENSE.txt
Disallow: */MAINTAINERS.txt
Disallow: */update.php
Disallow: */UPGRADE.txt
Disallow: */xmlrpc.php
Disallow: */admin/
Disallow: */comment/reply/
Disallow: */filter/tips/
Disallow: */search/
Disallow: */user/register/
Disallow: */user/password/
Disallow: */user/login/
Disallow: */user/logout/
# Paths (no clean URLs)
Disallow: */?q=admin/
Disallow: */?q=comment/reply/
Disallow: */?q=filter/tips/
Disallow: */?q=node/add/
Disallow: */?q=search/
Disallow: */?q=user/password/
Disallow: */?q=user/register/
Disallow: */?q=user/login/
Disallow: */?q=user/logout/
# wp
Disallow: */wp-admin/
Disallow: */wp-includes/
# dr
Disallow: */node/
Disallow: */taxonomy/
Disallow: */comment/
Disallow: */image_captcha/
Disallow: */tags/
Disallow: */tags-carab/
Disallow: */cataleg-col%C2%B7lectiu-de-les-biblioteques-dels-museus
Disallow: /barcelonacultura/en/collective-catalogue-museum-libraries
Disallow: /barcelonacultura/es/catalogo-colectivo-de-las-bibliotecas-de-los-museos
Disallow: /barcelonacultura/sites/default/files/
Disallow: /grec/api/
Disallow: */event-created/
Disallow: */llistat/
Disallow: */llistat?
Disallow: */listado?
Disallow: */listado/
Disallow: */list?
Disallow: */list/
Disallow: *?pg=search&
Disallow: */participantsorteig/
Disallow: */estat-participant/
Disallow: /barcelonacultura/*?page=
Disallow: /grec/*?page=
```

1.2 Grandària de la web

Per a l'anàlisi a realitzar, el nombre de pàgines web a les que s'accedrà serà limitat. No obstant això i en previsió d'algun error del propi analitzador o existència d'altres imprevistos, a través de Google s'observa que la mida d'aquesta web és relativament reduïda (21.500 pàgines):



1.3 Tecnologia.

La tecnologia utilitzada en el disseny del lloc web és:

```
In [10]: builtwith.builtwith("https://www.bcn.cat/estadistica/catala/index.htm")
Out[10]: {u'editors': [u'DreamWeaver'], u'web-servers': [u'Nginx']}
```

1.4 Propietari.

El propietari del lloc web és el propi Ajuntament de la ciutat de Barcelona com es mostra amb la següent funció Python:

```
In [11]: whois.whois ("https://www.bcn.cat/estadistica/catala/index.htm")
Out[11]:
{u'address': u'PL.SANT JAUME S/N',
 u'city': u'Barcelona',
 u'country': u'ES',
 u'creation_date': datetime.datetime(2006, 2, 16, 11, 54, 21, 401000),
 u'dnssec': u'unsigned',
 u'domain_name': u'bcn.cat',
 u'emails': [u'abuse@entorno.es', u'dominis@bcn.cat'],
 u'expiration_date': datetime.datetime(2020, 2, 16, 11, 54, 21, 401000),
 u'name': u"Departament d'Internet - Direccio de Comunicacio",
 u'name_servers': [u'ns2.bcn.cat', u'ns1.bcn.cat', u'ns3bcn.entorno.es'],
 u'org': u'Ajuntament de Barcelona',
 u'referral_url': None,
 u'registrar': u'Entorno Digital',
 u'state': u'Barcelona',
 u'status': u'ok https://icann.org/epp#ok',
 u'updated_date': datetime.datetime(2019, 1, 20, 11, 47, 10, 140000),
 u'whois_server': None,
 u'zipcode': u'08002'}
```

2. Definir un títol pel dataset.

Triar un títol que sigui descriptiu.

Per al conjunt de Datasets obtinguts per a aquest exercici:

“Eleccions locals Barcelona. Distribució i evolució del vot per Seccions Censals 2011-2019”

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).

Per a cada votació corresponent a les 3 últimes eleccions locals al municipi de Barcelona (22 de maig del 2011, 24 de maig del 2015 i 26 de maig del 2019) s’han obtingut dues taules:

- Taula amb la relació de candidatures presents a la votació (Nota 1).
- Taula amb els resultats electorals de cada candidatura segons la divisió territorial (Veure Nota 2) del municipi de Barcelona:
    - Districte
    - Barri
    - Área Estadística Bàsica
    - Secció Censal.
- En quant als resultats de l’any 2019, s’ha obtingut, a més, diferent informació gràfica per barris i seccions censals.

Nota 1: Al realitzar aquest exercici de Web Scraping s’ha observat que hi ha alguna inconsistència entre la pàgina web de relació de candidatures i la pàgina de resultats per candidatura, especialment en els anys 2011 i 2015. En qualsevol cas, aquesta és la

informació subministrada per l'Ajuntament. Així mateix, cal tenir en compte que els noms de les candidatures entre elecció i elecció, a nivell local, tenen una gran variació. Si cal fer un estudi a partir d'aquests fitxers, caldria establir alguna mena d'equivalència.

Nota 2: La descripció de la divisió territorial del municipi de Barcelona es pot trobar a https://www.bcn.cat/estadistica/catala/terri/index.htm, en resum la ciutat està dividida en:

- 10 Districtes Municipals (Denominació oficial numèrica i també nominal). En el fitxer obtingut en aquest estudi, el valor és el numèric.
- 73 Barris (Denominació oficial numèrica i nominal). En el fitxer obtingut en aquest estudi, el valor és el numèric. Van començar a haver dades estadístiques a nivell de Barris a partir de l'any 2007.
- 233 Àrees Estadístiques Bàsiques (AEB). Nivell territorial intermig entre Barris i les seccions censals. La denominació és numèrica i seguida, de l'1 al 233.
- 1068 Seccions Censals (SC). L'Ajuntament ha definit seccions el més homogènies possible, de formes regulars i amb un nombre d'electors aproximat de 1.000. Han variat al llarg del temps. La denominació de les seccions censals és numèrica: núm. Districte i núm. Secció censal; començant en cada districte la numeració de les seccions des del número 1. El seccionat està vigent des del gener de 2009 i tenia 1.061 seccions censals. L'1 de gener de 2014 el seccionat ha passat a tenir 1.068 seccions censals.

Cada nivell territorial està contingut en el nivell immediatament superior.

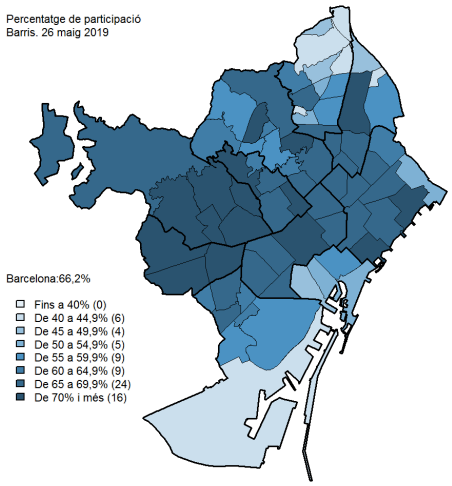
4. Representació gràfica.

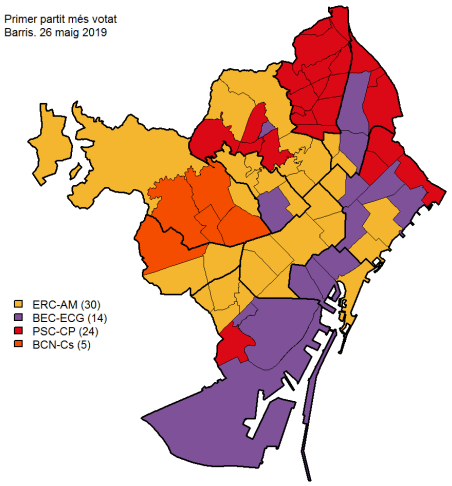
Presentar una imatge o esquema que identifiqui el dataset visualment.

La representació gràfica que es proposa és una composició de mapes extrets de la web de l'Ajuntament de Barcelona:
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc19/t312.htm i
https://www.bcn.cat/estadistica/catala/dades/telec/loc/loc19/t311.htm

sobre la participació per barris i resultats del partir més votat a cada barri l'any 2019.

La imatge justifica el treball que es podria realitzar com a proposta d'aquest estudi: un estudi de l'evolució del vot a les eleccions locals del municipi, el primer pas d'aquest estudi seria l'extracció de dades proporcionades pel mateix Ajuntament de Barcelona a la web:
https://www.bcn.cat/estadistica/catala/dades/telec/loc/index.htm





5. Contingut.

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

Dues taules per cada votació (2011, 2015, 2019) tenen el següent contingut:

5.1. Taula de relacions de candidatures.

Conté:

- Sigles de les Candidatures. Alfanumèric
- Noms de les Candidatures. Alfanumèric.
- Addicionalment, la relació de candidatures corresponent a l'any 2019 conté una tercera columna: Nom del candidat a Alcalde.

Cal tenir en compte la manca d'homogeneïtat de sigles de candidatures d'una elecció a una altra. En cas de posteriors estudis, caldria establir algun tipus d'equivalència.

La taula conté un primer registre amb el nom de les columnes

El nom dels fitxers és:
"Eleccions_locals_Barcelona_*aaaammdd*_Candidatures.csv"

on *aaaammdd* correspon a la data de les 4 eleccions: 20190526, 20150524 i 20110522.

5.2. Taula de resultats de les eleccions locals

Conté (veure nota):

- Els 4 camps de la Divisió territorial: Districte - Barri - AEB i SC.
- Nombre d'Electors.
- Nombre de Votants.

- Abstenció (taules dels anys 2011 i 2019)
- Vots nuls.
- Vots vàlids (només taula de l'any 2019).
- Vots en Blanc.
- Un camp per a cada candidatura amb el nombre de vots obtinguts en cada SC.
- Una última columna amb la sigla del partir més votat (només taula de l'any 2019)

Nota: A part de les diferències mencionades al descriure els camps, s'observa que entre els diferents anys, l'ordre de les columnes varia. Cal comprovar el contingut de la columna segons el primer registre.

La taula conté un primer registre amb el nom de les columnes i un segon registre on Districte és "Barcelona", els restants 3 elements de la Divisió Territorial a blancs i per a cada candidatura el nombre de vots obtinguts en tot el municipi. Aquesta fila es redundant, es podria obtenir com la suma de la resta de files de detall de cada element de la Divisió Territorial

El nom dels fitxers és:
"Eleccions_locals_Barcelona_*aaaammdd*_Resultats.csv"

on *aaaammdd* correspon a la data de les 4 eleccions: 20190526, 20150524 i 20110522.

6. Imatges

La Web de l'Ajuntament de Barcelona proporciona, també, informació gràfica dels resultats de les eleccions local per a l'any 2019 (no així per als anys anteriors).

L'exercici de Web Scraping ha finalitzat recuperant la informació de 3 pàgines que contenen aquestes imatges:

- 2 Mapes de participació per barris i seccions censals
  Fitxers: "part_barris.png" i "part_sc.png"
- 2 Mapes del primer partit més votat per barris i seccions censals.
  Fitxers: "pguany_barris.png" i "pguany_sc.png
- 2 Mapes del segon partit més votat per barris i seccions censals.
  Fitxers: "pguany_barris2.png" i "pguany_sc2.png"
- 12 Mapes d'implantació de les principals candidatures per barris i seccions censals.
  Fitxers: "*xxx*_barris.png" i "*xxx*_sc.png" on *"xxx"* correspon la candidatura. com ara erc (Esquerra Republicana de Catalunya), bc (Barcelona en Comú) , psc, ...

7. Agraïments.

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

A l'apartat 1.4 ja s'ha exposat que les dades s'han obtingut de la web de l'Ajuntament de Barcelona, a través de tècniques de Web Scraping amb Python.

Cal posar en valor i agrair a l'Ajuntament de Barcelona per la creació i manteniment d'aquest servei públic que proporciona informació sobre diferents aspectes del municipi de Barcelona. En aquest sentit, la pàgina https://www.bcn.cat/estadistica/catala/index.htm contè l'índex de tota la informació subministrada per l'Ajuntament.

8. Inspiració.

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

Aquest conjunt de dades pot ser un punt de partida per a realitzar diferents estudis:

- Sobre les tendències electorals en els diferents àmbits territorials de la ciutat de Barcelona.
- L'evolució en el temps d'aquestes tendències electorals.
- Creuar aquesta informació amb altra informació de la ciutat com ara estadístiques demogràfiques o econòmiques (també presents a la web de l'Ajuntament) que permetin una millor comprensió i explicació de les tendències electorals per a cada àmbit territorial de la ciutat.

Es podria, així, plantejar-se qüestions com ara:

- Implantació territorial de cada partit.
- Influència de dades demogràfiques o de població (sexe, edat, …) sobre el sentit del vot.
- Influència de dades econòmiques sobre el sentit del vot.
- Predicció de l'evolució del vot en funció d'alguns dels paràmetres anteriors que es demostri que influeixen en el sentit del vot.

9. 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)
- Unknown License

Per al data set resultant escolliria la *"Released Under CC BY-NC-SA 4.0 License"*, donat que:

- la informació inicial que dóna lloc al dataset correspon a dades totalment públiques
- el futur usuari, llicenciatari, ha de referir-se al llicenciador quan faci ús del fitxer, respectant així el treball realitzat pel llicenciador.

- el treball realitzat de Web Scraping és de recollida de la informació per motius estrictament acadèmics i es desitja poder compartir (share) el fitxer així com no posar restriccions a que tercers adaptin les dades sense cap mena de restricció.
- El llicenciador, per les raons anteriors, no desitja permetre que es faci un ús comercial del dataset.

10. Codi.

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

S'adjunta el codi en Python: "Pràctica_1_jsabatei_Web_Scraping.py.

11. Dataset.

Presentar el dataset en format CSV

Els dataset obtinguts en aquest treball i descrits anteriorment són:

Resultats de les Eleccions Locals al municipi de Barcelona:

- Any 2019: Eleccions_locals_Barcelona_20190526_Resultats.csv
- Any 2015: Eleccions_locals_Barcelona_20150524_Resultats.csv
- Any 2011: Eleccions_locals_Barcelona_20110522_Resultats.csv

Candidatures de les Eleccions Locals al municipi de Barcelona:

- Any 2019: Eleccions_locals_Barcelona_20190526_Candidatures.csv
- Any 2015: Eleccions_locals_Barcelona_20150524_Candidatures.csv
- Any 2011: Eleccions_locals_Barcelona_20110522_Candidatures.csv

Imatges: Veure apartat 6. Imatges

12. Autors Dataset

Contribucions	Signa
Treball en la seva totalitat (prenent com a referència els documents mencionats a la Bibliografia)	Josep Maria Sabaté

13. Bibliografia

Aquest exercici es basa majoritàriament en els documents suggerits a l'assignatura de "Tipologia i cicle de vida de les dades", Bloc 2:

[1] Subirats, L., Pérez, D., Calvo, M. (2019). Introducció al cicle de vida de les dades. Editorial UOC
[2] Subirats, L., Calvo, M. (2019). Web scraping. Editorial UOC.

Addicionalment, s'ha consultat:

[3] Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 1. Introduction to Web Scraping

Barcelona, 9 de novembre del 2019