Pràctica 1: Web scraping

Assignatura: Tipologia i cicle de vida de les dades

Alumnes: Jesús Marí i Víctor Boix

Data d'entrega: 15 d'abril de 2019

Taula de continguts

1. Context

El Departament d'Ensenyament publica regularment els nomenaments de tot el professorat interí i substitut que forma part de la borsa de treball docent i que s'incorpora a treballar a centres educatius públics. Aquesta informació, juntament amb les dades públiques de la borsa de treball, és recopilada per diversos sindicats per tal de generar un llistat ordenat de tots els professors que indica quins d'ells estan treballant i les característiques de cada adjudicació. Aquest llistat es publica a la web dels sindicats i resulta molt important per als professionals que volen començar treballar o que volen canviar de centre, perquè els permet conèixer les possibilitats d'incorporar-se aviat o saber quines zones i especialitats ofereixen més oportunitats. Malgrat tot, tant la informació que publica el Departament com la publicada pels sindicats es troba segmentada per zones i especialitats, no ofereix una visió global de la borsa de treball.

Aquest conjunt de dades pretén recopilar tota aquesta informació en únic fitxer per tal de poder realitzar anàlisis més generals i que puguin resultar útils per a tot el professorat. A més, també s'han recuperat dades de cursos anteriors per poder estudiar l'evolució temporal de les dades.

2. Títol

Dades anuals dels nomenaments del professorat interí i substitut a Catalunya

3. Descripció del dataset

El conjunt de dades està format per 6 *datasets* corresponents a les dades de cada curs escolar entre 2013 i 2019. Aquests *datasets* es guarden en format CSV i s'anomenen *dadesyyYY.csv*, on *yy* és l'any d'inici de curs i *YY* l'any de finalització (per exemple *dades1819.csv* conté les dades del curs 2018-2019). A més, el codi permet capturar totes les dades en un únic *dataset* anomenat *dadesALL.csv* que engloba la informació de tots els cursos.

Cada *dataset* conté un registre per a cada professor que opta a un lloc de treball per a una especialitat concreta a un servei territorial. Cada registre conté informació sobre el curs, el número de la borsa que identifica al treballador, l'especialitat demanada i el servei territorial preferent. A més, per als professors que ja han sigut nomenats, es

disposa d'informació sobre l'adjudicació: especialitat, jornada, centre, dates d'inici i finalització.

Podem esquematitzar l'estructura del *dataset* amb la següent taula.

Informació	Nom	Tipus de dada	Descripció
Borsa	**curs**	*integer*	Curs escolar de les dades
	sstt	*integer*	Servei territorial (zona)
	especialitat	*string*	Codi de l'especialitat demanada
	inicials	*string*	Inicials del docent
	bloc	*integer*	Bloc de la borsa
	n_interi	*integer*	Número de la borsa
Nomenament	**data_ini**	*date*	Data d'inici del nomenament
	especialitat_dest	*string*	Codi de l'especialitat adjudicada
	codi_centre	*integer*	Codi del centre adjudicat
	centre	*string*	Nom del centre adjudicat
	tipus_jornada	*float*	Tipus de jornada laboral
	data_fi	*date*	Data de finalització del nom.

Descripció dels camps i tipus de dades del dataset

La web que s'ha utilitzat per extreure la informació té publicades dades des del curs 2013-2014 fins a l'actual. Les dades de cursos finalitzats ja no es modifiquen, per tant serà suficient amb capturar-les una única vegada; en canvi, les dades del curs actual (2018-2019) s'actualitzen setmanalment incorporant els nomenaments que s'han produït durant la darrera setmana. Això vol dir que serà necessari rastrejar aquest curs de manera periòdica, a més, aquesta captura haurà de sobreescriure necessàriament totes les dades del curs perquè resulta inviable obtenir només les dades noves i pot ser necessari incorporar alguna correcció o actualització.

Com es veurà més endavant, les dades recopilades són filtrades a partir d'un procés de neteja. Aquest es realitza per uniformitzar l'estructura dels fitxers, ja que el format de les dades publicades no és sempre el mateix a tots els cursos. A més, el conjunt de dades requerirà una neteja posterior que realitzi tasques com:

- Uniformitzar el nom dels centres
- Corregir el valor del tipus de jornada

- Completar el codi de centre
- Completar el valor del bloc.

Les dades poden contenir inconsistències, per exemple poden donar-se solapaments entre dates, és a dir, que un mateix docent estigui en dos centres als mateix temps, degut a errors en les dates d'adjudicació o degut a nomenaments que han acabat abans del previst i no consten a les dades. També poden donar-se problemes de valors nuls. Quan un professor no ha començat a treballar tots els seus atributs corresponents als nomenaments es troben en blanc, però pot donar-se la situació que un docent en actiu tingui valors nuls en algun dels camps. Per exemple, hem detectat valors incomplets o incorrectes en camps com el tipus de jornada o la data. Aquest tipus de problemàtiques també caldrà resoldre-les en una fase posterior de neteja de dades.

4. Representació gràfica

Imatge que pretén il·lustrar la llista d'espera del professorat a Catalunya

5. Contingut

5.1. Avaluació inicial

Les dades a recopilar es troben totes allotjades a la pàgina web del sindicat USTEC-STEs, al domini http://sindicat.net. El fitxer *robots.txt* es troba a la URL http://sindicat.net/robots.txt i permet un accés complet a tots els visitants. De totes

maneres, per precaució, hem utilitzat el mòdul *robotparser* per comprovar l'accés a cada enllaç abans de descarregar-lo. D'altra banda, hem modificat el *User-agent* per evitar bloquejos per defecte.

La web no disposa d'un mapa del lloc (*sitemap*), però els enllaços a rastrejar es troben tots organitzats seguint una mateixa estructura. La informació es troba classificada per cursos en diferents pàgines identificades per l'any, seguint el format http://sindicat.net/borsa/yyYY on yy és l'any d'inici del curs i YY l'any de finalització. Les dades disponibles van des del curs 2013-2014 (https://sindicat.net/borsa/1314/) fins al 2018-2019 (https://sindicat.net/borsa/1819/).

Cadascuna d'aquestes pàgines conté un llistat d'enllaços agrupats per especialitat i servei territorial, com es veu a la taula següent.

Altres codis

Els números en negreta són el codi del SSTT i els números en blau la quantitat d'interins. Al clicar surt la llista íntegra i la plaça adjudicada.

133 --cat-> 1=22 - 2=9 - 3=16 - 4=9 - 5=5 - 6=7 - 17=8 - 25=2 - 43=8 - 44=2 -
134 --cat-> 1=8 - 2=1 - 17=2 - 25=2 - 43=1 -
135 --cat-> 1=7 - 2=1 - 3=2 - 4=1 - 17=3 - 25=1 -
136 --cat-> 1=1 -
137 --cat-> 1=3 -
138 --cat-> 1=5 - 5=1 - 17=1 - 25=1 - 43=3 -
190 --cat-> 1=30 - 2=6 - 4=3 - 17=3 - 25=2 - 43=6 -
192 --cat-> 1=30 - 2=8 - 3=10 - 4=18 - 5=13 - 6=6 - 17=15 - 25=5 - 43=9 - 44=6 -
193 --cat-> 1=78 - 2=26 - 3=45 - 4=48 - 5=27 - 6=24 - 17=45 - 25=20 - 43=49 - 44=16 -
195 --cat-> 1=19 - 2=4 - 3=2 - 4=13 - 6=3 - 17=8 - 25=6 - 43=6 -
198 --cat-> 1=1 -
501 --cat-> 1=189 - 2=137 - 3=181 - 4=265 - 5=281 - 6=168 - 17=313 - 25=135 - 43=201 - 44=112 -
502 --cat-> 1=173 - 2=94 - 3=125 - 4=142 - 5=116 - 6=105 - 17=172 - 25=72 - 43=116 - 44=56 -

Llistat d'enllaços corresponent al curs 2018-19

Cada enllaç permet accedir a una pàgina diferent que conté totes les dades. L'especialitat i el servei territorial poden capturar-se del títol, mentre que la resta d'informació es troba agrupada en una taula.

SSTT: 3 - ESPECIALITAT: MA

1	**TG,J 1 143**				
2	**MM,S 1 219**	17/07 1	**MA**	Institut Joan Oró - FITXA	31/08
3	**CT,MC 1 224**	17/07 1	**517**	Institut Miquel Martí i Pol - FITXA	31/08
4	**SG,J 1 266**	17/07 1	**TEC**	Institut Josep Lluís Sert - FITXA	31/08
5	**AF,MP 1 275**	17/07 1	**MA**	Institut Francesc Macià - FITXA	31/08
6	**CS,J 1 627**	17/07 1	**MA**	Institut Estany de la Ricarda - FITXA	31/08
7	**FL,J 1 915**	17/07 1	**TEC**	Institut de Sales - FITXA	31/08
8	**TR,M 1 1.102**	17/07 1	**MA**	Institut Voltrera - FITXA	31/08
9	**RL,J 1 1.472**	17/07 1	**507**	Institut Eugeni d'Ors - FITXA	31/08
10	**VE,MC 1 1.563**	17/07 1	**MA**	Institut Esteve Terradas i Illa - FITXA	31/08
11	**MM,AJ 1 1.773**	17/07 1	**MA**	Institut Estany de la Ricarda - FITXA	31/08
12	**AM,AM 1 2.082**	17/07 1	**ECO**	Institut Rafael Casanova - FITXA	31/08
13	**DLRM,S 1 3.001**	17/07 1	**MA**	Institut Josep Mestres i Busquets - FITXA	31/08
14	**JC,V 1 3.054**				

Dades de la borsa per a l'especialitat de matemàtiques (MA)

al Baix Llobregat (SSTT:3)

5.2. Extracció de les dades

El programa, a través d'un paràmetre, permet recollir les dades corresponents a un únic curs o bé a tots. A partir d'aquest paràmetre, s'accedeix a la pàgina corresponent i es guarden els enllaços a totes les taules de dades (identificats per contenir *'ctot.php'* a la URL). A continuació, es descarrega el codi HTML utilitzant el mòdul *requests* i es genera una estructura en forma d'arbre amb la llibreria *BeautifulSoup* per tal de seleccionar el contingut necessari. La descàrrega de cada pàgina es troba separada per un temps d'espera proporcional a la resposta del servidor, d'aquesta manera les peticions s'adapten a la resposta del servidor i s'evita sobrecarregar-lo amb moltes descàrregues consecutives.

La informació de l'especialitat i el servei territorial l'obtenim del títol (*h1*), mentre que la resta de dades es troben dins de l'etiqueta '*table*'. Alguns atributs s'han obtingut directament de la cel·la corresponent a cada fila de la taula (*td*), mentre que en altres atributs ha sigut necessari separar la informació continguda en una mateixa cel·la.

Una de les principals dificultats que s'ha trobat a l'hora de capturar les dades és la diferència en el format i ordre dels camps per als diferents cursos. Això s'ha resolt modificant, en funció del curs que es vol rastrejar, el paràmetre *columns* que indica el nom de les columnes utilitzat per afegir les dades al *DataFrame*.

5.3. Neteja de dades

Una vegada carregades les dades sense processar, utilitzant un *DataFrame* de *Pandas,* s'ha realitzat una primera neteja de dades per uniformitzar el format i tipus de dada de les columnes i corregir alguns valors erronis. Per a cada columna s'ha definit un filtre, implementat a les diferents classes del mòdul *columnFilter*, que s'encarrega de netejar i transformar les dades de la columna. El filtres aplicats a cada columna s'assignen amb el diccionari *COL_TRANSFORMER_MAP,* definit com a constant de la classe *WebScraper.* Per fer-ho, es crea una instància de la classe *RowTransformer* amb les dades i el diccionari, i s'executa el mètode *transform()*, que aplica els filtres de manera successiva a les diferents columnes i retorna un *DataFrame* amb les dades processades. Aquest procés de neteja inclou transformacions com corregir el tipus de dada, eliminar caràcters no vàlids, igualar el format de les dates o uniformitzar la nomenclatura per al tipus de jornada. També ha resultat necessari determinar l'any per als valors de les columnes *data_ini* i *data_fi*, ja que només es publica el dia i el mes. Aquest s'ha calculat a partir del mes i el curs que s'està rastrejant tenint en compte el calendari escolar.

5.4. Descripció del contingut

Una vegada extret i netejat, cada *dataset* està format per 12 camps que inclouen totes les dades de les taules inspeccionades:

- **curs (integer).** Curs escolar al què correspon el registre seguint el format yyYY (per exemple: 1819).
- **sstt (integer).** Codi numèric del servei territorial preferent (zona) per al qual s'opta a treballar (veure Taula 1, Annex 1).
- **especialitat (string).** Codi de l'especialitat docent sol·licitada pel professor, format per 2 o 3 dígits.
- **inicials (string).** Inicials del professor. Tres lletres, per exemple CC, N.
- **bloc (integer).** Indica el bloc de la borsa de què forma part el professor (1 si ha treballat anteriorment al Departament i 2 si no ha treballat mai).
- **n_interi (integer).** Número d'ordre de tot el professorat a la borsa de treball en funció del temps treballat.
- **data_ini (date).** Data d'incorporació al centre en format *yyy-mm-dd*.

- **especialitat_dest (string).** Codi de l'especialitat docent per a la qual ha sigut nomenat un professor (2 o 3 dígits).
- **codi_centre (integer).** Codi de 8 xifres que identifica a cada centre de manera única.
- **centre (string).** Nom del centre de treball de l'adjudicació.
- **tipus_jornada (float)**. Durada de la jornada de treball adjudicada (1 sencera, 0'5 mitja, 0'33 terç...).
- **data_fi (date).** Data de finalització del nomenament en format dd/mm/yyyy.

6. Agraïments

Les dades utilitzades per generar el *dataset* són publicades regularment pel Departament d'Ensenyament a través d'un aplicació de la seva pàgina web. Aquestes dades són recollides, processades i publicades de manera oberta per diferents sindicats per tal d'oferir una informació més completa al conjunt del professorat.

Aquest conjunt de dades ha sigut generat a partir de la informació publicada pel sindicat USTEC-STEs perquè disposa d'una informació molt completa: les dades s'han creuat amb el llistat de tot el professorat de la borsa, el període de temps és força extens (disposa de dades des del 2013), la informació es troba ben estructurada, resulta fàcil de capturar i és completament accessible a tots els rastrejadors.

Hi ha altres sindicats especialitzats en el sector de l'ensenyament que també ofereixen informació sobre la borsa de treball a la seva pàgina web; per exemple, el Sindicat Professors de Secundària (*aspepc-sps*) disposa d'una aplicació web que permet consultar i filtrar dades sobre tots els nomenaments a l'enllaç https://secundaria.info/nomenaments.php. Malgrat tot, aquestes dades sempre es troben segmentades per especialitats i serveis territorials i no permeten realitzar anàlisis com els que proposem.

Durant la recerca anterior a la realització d'aquest treball no hem trobat cap pàgina web que ofereixi aquesta informació de manera global com proposem nosaltres, ni cap estudi que abordi preguntes com les plantejades en aquest apartat.

7. Inspiració

Aquest conjunt de dades vol oferir al professorat de Catalunya una informació integrada de la borsa de treball que l'ajudi a prendre decisions sobre el seu futur professional. Aquestes dades han de permetre realitzar anàlisis més generals, com per exemple comparar la oferta i la demanda de professorat per zones i especialitats, estudiar la freqüència, distribució i característiques de les vacants adjudicades, avaluar la mobilitat de professors entre serveis territorials o analitzar l'evolució temporal del conjunt de la borsa. Algunes de les preguntes a les quals es vol donar resposta poden ser:

- Especialitats i zones amb més possibilitats de ser nomenat.
- Temps mitjà d'espera en funció de la posició a les llistes.
- Percentatges d'ocupació per especialitats i per zones.
- Freqüència dels nomenaments, tant a nivell global com per especialitat i localització.
- Mobilitat del professorat: interins que són nomenats per a un servei territorial que no es correspon amb el sol·licitat preferentment.
- Evolució anual del nombre de nomenaments i del percentatge d'ocupació de les llistes.

Per tal de fer accessible aquesta informació a tota la gent interessada, sense necessitat de disposar de coneixements d'anàlisi de dades, pot resultar interessant publicar-la a través d'una API que permeti realitzar-hi consultes específiques. Això vol dir que es podria publicar a través d'una aplicació web on els usuaris puguin consultar i filtrar el contingut de manera dinàmica.

8. Llicència

Per al conjunt de dades hem triat la llicència *"Creative Commons Attribution Share Alike 4.0 International" CC-BY-SA-4.0*, que pertany al tipus *Open Source* aplicades a materials que no són programari, com conjunts de dades i material multimèdia.

Aquesta llicència permet l'ús comercial, distribució, modificació i ús privat amb certes condicions: cal adjuntar una còpia de la llicència, les modificacions sobre el conjunt de dades s'han de publicar amb una llicència del mateix tipus i documentant

tots els canvis realitzats, i els contribuïdors del projecte no adquireixen cap dret sobre les dades. D'altra banda, també especifica que no s'ofereix cap garantia o responsabilitat pels danys causat pel seu ús.

Hem triat aquesta llicència perquè ens interessa que s'hi puguin realitzar canvis i aportacions, però conservant un control dels canvis i mantenint una llicència oberta. Ens interessa que tot el col·lectiu pugui fer ús de la informació continguda i aportada per tercers en el futur. Per aquest motiu, hem descartat llicències més permissives, com CC0-1.0, on les modificacions poden ser alliberades amb llicències comercials tancades.

9. Codi

Per a l'extracció de les dades s'ha utilitzat *Python3* juntament amb les llibreries *requests*, *re*, *time*, *os*, *pandas, robotparser* i *BeautifulSoup*. Tot el codi font i el conjunt de dades generat es troba disponible a l'enllaç següent:

https://github.com/jmari/BorsaDensenyamentCat

El codi s'ha estructurat seguint el paradigma de la programació orientada a objectes. Per extreure les dades corresponents a un curs concret cal instanciar un objecte de la classe *WebScraper* tot especificant el curs com a paràmetre d'entrada. A continuació pot utilitzar-se el mètode *scrape()* per a obtenir les dades del curs corresponent, el mètode *write_csv()* per guardar-les en un fitxer CSV i el mètode *get_data()* per a obtenir un *dataframe* amb tota la informació. Per exemple:

```
ws = WebScraper("1819")  # Creació de l'objecte
ws.scrape()  # Obtenció de les dades
ws.write_csv()  # Emmagatzematge de les dades
df = ws.get_data()  # Obtenció d'un dataframe
```

Els fitxers que formen el codi font són:

- **main.py -** Mètode principal, inicia el procés de *web scraping* donat un curs.
- **scraper.py -** Conté la implementació de la classe *WebScraper* i els mètodes encarregats de descarregar, recopilar i generar el dataset.

- **columnFilter.py -** Conté diferents classes que actuen de filtres per a la neteja de dades dels diferents camps del dataset.

10. Dataset

Totes les dades recopilades es troben disponibles en format CSV a l'enllaç següent: https://github.com/jmari/BorsaDensenyamentCat/tree/master/data

Es disposa d'un fitxer (*dadesyyYY.csv*) per a cada curs entre 1314 i 1819. El caràcter ';' s'utilitza com a separador i la primera fila conté el nom de tots els camps.

```
curs;sstt;especialitat;inicials;bloc;n_interi;data_ini;especialitat_dest;codi_centre;centre;tipus_jornada;data_fi
1819;1;133;DR,A;1;77;17/07/2018;193;;EOI Barcelona-Vall d'Hebron;1.0;31/08/2019
1819;1;133;FE,D;1;562;17/07/2018;133;;EOI Santa Coloma;1.0;31/08/2019
1819;1;133;G,I;1;2556;17/07/2018;133;;EOI Barcelona IV;0.5;31/08/2019
1819;1;133;LA,MR;1;10013;17/07/2018;133;;EOI de Martorell;1.0;31/08/2019
1819;1;133;MP,MA;1;10453;17/07/2018;AN;;Institut Arnau Cadell;1.0;31/08/2019
1819;1;133;MB,L;1;12481;17/07/2018;133;;EOI de Viladecans;1.0;31/08/2019
1819;1;133;GG,M;1;14737;17/07/2018;AL;;Institut La Pineda;1.0;31/08/2019
1819;1;133;VC,J;1;16030;17/07/2018;AN;;Institut de Castellar;1.0;31/08/2019
1819;1;133;MODL,I;1;16567;17/07/2018;AL;;Institut Icària;1.0;31/08/2019
1819;1;133;VG,E;1;19312;17/07/2018;AL;;Institut Ausiàs March;0.5;31/08/2019
1819;1;133;PS,T;1;22075;17/07/2018;133;;EOI Guinardó;0.5;31/08/2019
1819;1;133;PP,J;1;23341;;;;;;
1819;1;133;SS,M;1;25299;;;;;;
1819;1;133;B,A;1;25347;17/07/2018;133;;EOI d'Olot;0.5;31/08/2019
1819;1;133;BSJ,H;1;36238;;;;;;
1819;1;133;TY,E;1;36304;;;;;;
1819;1;133;LL,S;1;36353;;;;;;
1819;1;133;BG,A;2;47955;;;;;;
1819;1;133;SJ,M;2;53935;14/09/2018;133;;EOI Barcelona III;0.5;31/08/2019
1819;1;133;PC,M;2;71340;17/09/2018;133;;EOI de Manresa;;07/12/2018
1819;1;133;GQ,G;2;73165;06/11/2018;133;;EOI Martorell;;21/12/2018
```

Primers registres del fitxer dades1819.csv

	A	B	C	D	E	F	G	H	I	J	K	L
1	curs	sstt	especialitat	inicials	bloc	n_interi	data_ini	especialitat_dest	codi_centre	centre	tipus_jornada	data_fi
2	1819	1	133	DR,A	1	77	17/7/18	193		EOI Barcelona-Vall d'Hebron	1.0	31/8/19
3	1819	1	133	FE,D	1	562	17/7/18	133		EOI Santa Coloma	1.0	31/8/19
4	1819	1	133	G,I	1	2556	17/7/18	133		EOI Barcelona IV	0.5	31/8/19
5	1819	1	133	LA,MR	1	10013	17/7/18	133		EOI de Martorell	1.0	31/8/19
6	1819	1	133	MP,MA	1	10453	17/7/18	AN		Institut Arnau Cadell	1.0	31/8/19
7	1819	1	133	MB,L	1	12481	17/7/18	133		EOI de Viladecans	1.0	31/8/19
8	1819	1	133	GG,M	1	14737	17/7/18	AL		Institut La Pineda	1.0	31/8/19
9	1819	1	133	VC,J	1	16030	17/7/18	AN		Institut de Castellar	1.0	31/8/19
10	1819	1	133	MODL,I	1	16567	17/7/18	AL		Institut Icria	1.0	31/8/19
11	1819	1	133	VG,E	1	19312	17/7/18	AL		Institut Ausis March	0.5	31/8/19
12	1819	1	133	PS,T	1	22075	17/7/18	133		EOI Guinard	0.5	31/8/19
13	1819	1	133	PP,J	1	23341						
14	1819	1	133	SS,M	1	25299						
15	1819	1	133	B,A	1	25347	17/7/18	133		EOI d'Olot	0.5	31/8/19
16	1819	1	133	BSJ,H	1	36238						
17	1819	1	133	TY,E	1	36304						
18	1819	1	133	LL,S	1	36353						
19	1819	1	133	BG,A	2	47955						
20	1819	1	133	SJ,M	2	53935	14/9/18	133		EOI Barcelona III	0.5	31/8/19
21	1819	1	133	PC,M	2	71340	17/9/18	133		EOI de Manresa		7/12/18
22	1819	1	133	GQ,G	2	73165	6/11/18	133		EOI Martorell		21/12/18

Primers registres del fitxer dades1819.csv visualitzats a través d'un full de càlcul

11. Recursos

SUBIRATS, L.; CALVO, M. (2018). *Web Scraping*. Editorial UOC.

LAWSON, R. (2015). *Web Scraping with Python*. Packt Publishing Ltd.

12. Taula de contribucions al treball

Contribucions	Signatura
Recerca prèvia	Jesús Marí, Víctor Boix
Redacció de les respostes	Jesús Marí, Víctor Boix
Desenvolupament del codi	Jesús Marí, Víctor Boix

13. Annex: Codis dels serveis territorials

Codi	Servei territorial
1	Barcelona Consorci
2	Barcelona Comarques
3	Baix Llobregat
4	Vallès Occidental
5	Maresme i Vallès Oriental
6	Catalunya Central
17	Girona
25	Lleida
43	Tarragona
44	Terres de l'Ebre