Tipologia i cicle de vida de les dades

Pràctica 1

Nom: Pedro Galán

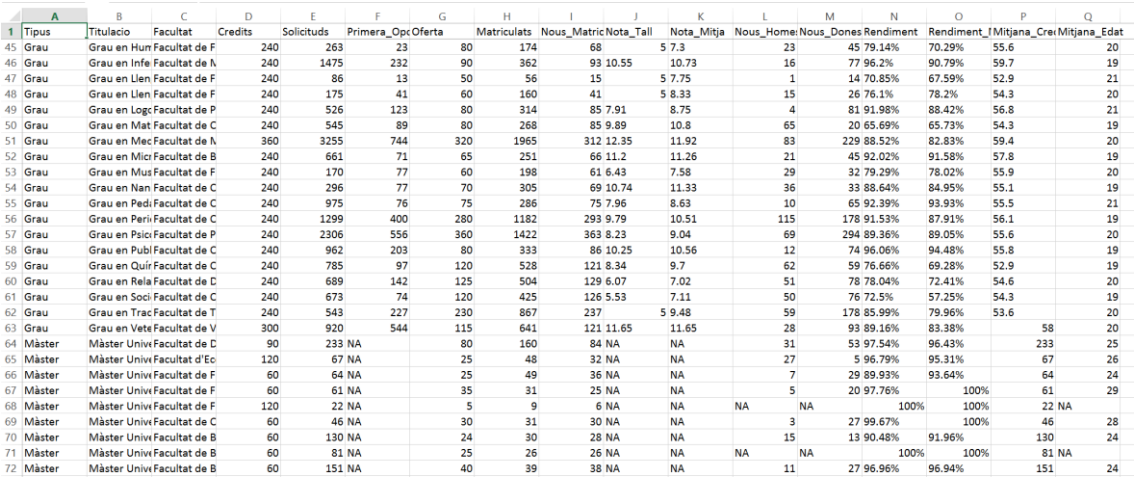
1. Títol del Dataset

El títol escollit pel dataset és: “2016 UAB Degrees”. El nom del fitxer serà “2016-UABDegrees.CSV”

2. Subtítol del Dataset

‘Data coming from quality indicators of all 2016 “grau” degrees in University Autonomous of Barcelona, fetched by scraping techniques’

3. Imatge



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Tipus	Titulacio	Facultat	Credits	Solicituds	Primera_Opc	Oferta	Matriculats	Nous_Matric	Nota_Tall	Nota_Mitja	Nous_Home	Nous_Dones	Rendiment	Rendiment_I	Mitjana_Cre	Mitjana_Edat
45	Grau	Grau en Hum	Facultat de F	240	263	23	80	174	68	5	7.3	23	45	79.14%	70.29%	55.6	20
46	Grau	Grau en Infe	Facultat de N	240	1475	232	90	362	93	10.55	10.73	16	77	96.2%	90.79%	59.7	19
47	Grau	Grau en Llen	Facultat de F	240	86	13	50	56	15	5	7.75	1	14	70.85%	67.59%	52.9	21
48	Grau	Grau en Llen	Facultat de F	240	175	41	60	160	41	5	8.33	15	26	76.1%	78.2%	54.3	20
49	Grau	Grau en Logc	Facultat de P	240	526	123	80	314	85	7.91	8.75	4	81	91.98%	88.42%	56.8	21
50	Grau	Grau en Mat	Facultat de C	240	545	89	80	268	85	9.89	10.8	65	20	65.69%	65.73%	54.3	19
51	Grau	Grau en Mec	Facultat de N	360	3255	744	320	1965	312	12.35	11.92	83	229	88.52%	82.83%	59.4	20
52	Grau	Grau en Micr	Facultat de B	240	661	71	65	251	66	11.2	11.26	21	45	92.02%	91.58%	57.8	19
53	Grau	Grau en Mus	Facultat de F	240	170	77	60	198	61	6.43	7.58	29	32	79.29%	78.02%	55.9	20
54	Grau	Grau en Nan	Facultat de C	240	296	77	70	305	69	10.74	11.33	36	33	88.64%	84.95%	55.1	19
55	Grau	Grau en Ped	Facultat de C	240	975	76	75	286	75	7.96	8.63	10	65	92.39%	93.93%	55.5	21
56	Grau	Grau en Peri	Facultat de C	240	1299	400	280	1182	293	9.79	10.51	115	178	91.53%	87.91%	56.1	19
57	Grau	Grau en Psic	Facultat de P	240	2306	556	360	1422	363	8.23	9.04	69	294	89.36%	89.05%	55.6	20
58	Grau	Grau en Publ	Facultat de C	240	962	203	80	333	86	10.25	10.56	12	74	96.06%	94.48%	55.8	19
59	Grau	Grau en Quír	Facultat de C	240	785	97	120	528	121	8.34	9.7	62	59	76.66%	69.28%	52.9	19
60	Grau	Grau en Rela	Facultat de D	240	689	142	125	504	129	6.07	7.02	51	78	78.04%	72.41%	54.6	20
61	Grau	Grau en Soci	Facultat de C	240	673	74	120	425	126	5.53	7.11	50	76	72.5%	57.25%	54.3	19
62	Grau	Grau en Trac	Facultat de T	240	543	227	230	867	237	5	9.48	59	178	85.99%	79.96%	53.6	20
63	Grau	Grau en Vete	Facultat de V	300	920	544	115	641	121	11.65	11.65	28	93	89.16%	83.38%	58	20
64	Màster	Màster Univ	Facultat de D	90	233	NA	80	160	84	NA	NA	31	53	97.54%	96.43%	233	25
65	Màster	Màster Univ	Facultat d'Ec	120	67	NA	25	48	32	NA	NA	27	5	96.79%	95.31%	67	26
66	Màster	Màster Univ	Facultat de F	60	64	NA	25	49	36	NA	NA	7	29	89.93%	93.64%	64	24
67	Màster	Màster Univ	Facultat de F	60	61	NA	35	31	25	NA	NA	5	20	97.76%	100%	61	29
68	Màster	Màster Univ	Facultat de F	120	22	NA	5	9	6	NA	NA	NA	NA	100%	100%	22	NA
69	Màster	Màster Univ	Facultat de C	60	46	NA	30	31	30	NA	NA	3	27	99.67%	100%	46	28
70	Màster	Màster Univ	Facultat de B	60	130	NA	24	30	28	NA	NA	15	13	90.48%	91.96%	130	24
71	Màster	Màster Univ	Facultat de B	60	81	NA	25	26	26	NA	NA	NA	NA	100%	100%	81	NA
72	Màster	Màster Univ	Facultat de B	60	151	NA	40	39	38	NA	NA	11	27	96.96%	96.94%	151	24

4. Context

Aquest Dataset recull dades publicades pel Sistema Intern de Qualitat de la Universitat Autònoma de Barcelona sobre les seves Titulacions. Les dades corresponen als principals indicadors que recull aquest sistema per tal de realitzar la avaluació de les titulacions.

En concret, són dades que permeten fer-se una idea del volum d’estudiants d’una titulació, de lo fàcil o difícil que es accedir-hi, d’un perfil bàsic del estudiant de la titulació i del que sol·licita l’ingrés a la mateixa i de com és l’evolució dels estudiants dins la titulació tant a nivell de resultats com de ritme a la que es segueix.

5. Contingut

Les dades corresponen a l'any 2016, i per tant a l'any acadèmic 2016-2017. Provenen de la BBDD del Sistema Intern de Qualitat de la UAB i estan publicades a l'apartat públic d'aquest sistema, accessible mitjançant el Portal de Transparència a la URL (http://siq.uab.cat/siq_public/titulacions/). Per construir el dataset s'ha fet servir una tècnica basada en crawling i scraping: Una eina ad-hoc programada en Python ha recorregut el conjunt de pàgines web que mostren informació de cada titulació, seleccionant dintre de cadascuna les dades apropiades i extraient els valors per anar-los introduint al dataset.

El Dataset conté 139 registres.

Dintre de cadascun d'aquests registres trobem els valors que es corresponen als següents camps:

- **Tipus**: És un camp discret que pren valors "Grau" o "Màster" i ens identifica el tipus de titulació que correspon al registre.
- **Titulació**: És un camp de text que conté el nom de la titulació.
- **Facultat**: És un camp de discret que indica la facultat a la que pertany cada titulació. Pren com a valors els noms de les diferents facultats de la UAB.
- **Crèdits**: Indica el numero de crèdits necessaris per assolir la titulació.
- **Solicituds**: Numero de persones que han sol·licitat accés a aquesta titulació pel curs que comença al 2016.
- **Primera_Opcio**: Numero de sol·licituds que corresponen als que l'han demanat com a primera opció .
- **Oferta**: Número de places que s'oferien pel curs que s'inicia al 2016.
- Matriculats: Número total de matriculats en la titulació al 2016.
- **Nous_Matriculats**: Numero total d'estudiants de nou ingrés al 2016.
- **Nota_Tall**: Nota mínima que ha fet falta per poder accedir a la titulació al 2016.
- **Nota_Mitja**: Mitjana de la nota d'entrada dels estudiants de nou accés a la titulació incloent totes les modalitats
- **Nous_Homes**: Número d'estudiants de nou ingrés que són homes.
- **Nous_Dones**: Número d'estudiants de nou ingrés que són dones.
- **Rendiment**: Valor en forma de percentatge que indica la relació entre el nombre de crèdits superats i el nombre de crèdits matriculats pel conjunt d'alumnes de la titulació.
- **Rendiment_Nous**: Valor en forma de percentatge que indica la relació entre el nombre de crèdits superats i el nombre de crèdits matriculats pel conjunt d'alumnes que han ingressat a la titulació per primer cop en 2016.
- **Mitjana_Credits**: Mitjana de crèdits matriculats pels alumnes de la titulació per l'any acadèmic amb inici al 2016.
- **Mitjana_Edat**: Mitjana d'edat dels estudiants de la titulació el 2016.

6. Agraïments

El propietari de les dades és la Universitat Autònoma de Barcelona. Vull agrair-li que posi aquest subset disponible de forma pública a la web i el seu compromís amb la transparència i que m'ha permès fer la recol·lecció del dataset.

7. Inspiració

Les dades originals del SIQ, tot i que tenen una presentació gràfica força acurada, no permeten la descarrega de les dades en format Excel o similar, cosa que si permeten altres apartats del Portal de Transparència. Per tant, dificulta realitzar altres anàlisis diferents a les proposades pel propi sistema. A més, estan contingudes en planes webs separades de forma que només es mostren gràfiques de les dades d'aquella titulació en concret. En canvi, a partir del dataset és possible mostrar informació gràfica sobre el conjunt de les titulacions.

Les dades contingudes al Dataset son interesants des del punt de vista analític perquè permeten una anàlisi comparativa entre les diferents titulacions, així com detectar patrons comuns i/o agrupacions de titulacions.

També és interesant des del punt de vista del valor que poden tenir per realitzar estudis globals sobre els sistema Universitari en conjunt, a partir de la seva incorporació i combinació amb dades similars d'altres institucions.

8. Llicencia

Les dades contingudes en el Dataset han accedides mitjançant el Portal de Transparència de la UAB, una institució pública amb finalitats acadèmiques. Per tant, penso que el més adequat és distribuir el propi dataset amb la llicencia "*Database released under Open Database License, individual contents under Database Contents License*".

Aquesta llicencia permet la manipulació de les dades, combinar-les i crear altres dades i coneixement a partir d'elles i per la natura de les mateixes així és com podrien tenir més valor, segurament en combinació amb altres dades d'altres fonts que permetessin fer un anàlisi més complert i transversal. Però a diferència d'altres llicencies, a més, obliga a distribuir les dades obertes fins i tot encara que permet que existeixi alguna versió restringida a la vegada. Crec que és el més coherent, tenint en compte l'origen públic de la propietat de les dades i el compromís manifest de les Universitats amb l'impuls del Open Data.

9. Codi

El codi està desenvolupat en Python i es troba al directori 'code' dins el repositori de Github *ScraperTCVD*.

El programa realitza tant les funcions de Crawler com de Scraper, de manera que inicialment realitza un escaneig de l'adreça arrel del repositori de pàgines web, genera una llista de les URL que es troba i a partir d'aquí, comença un bucle que per cadascuna d'aquestes adreces realitza la localització i recuperació de les dades que se li demanen.

Com a comentaris a realitzar, els següents:

- El mètode que he triat per realitzar l'scraping ha estat el dels CSSSelectors, un cop avaluat les característiques de cadascun dels tres mètodes explicats a teoria i avaluat igualment la viabilitat d'aplicar aquest mètode en el projecte concret que volia abordar.
- Una de les dificultats principals que he trobat en la fase d'Scraping ha estat el fet que moltes de les dades es trobaven en fileres de taula sense cap tag de classe ni d'id, i per tant dificultava el poder accedir de forma ràpida amb un CSSSelector. A més, moltes de les taules tenien la mateixa classe i tampoc un id diferenciador. Per solucionar-ho, inicialment he partit d'una aproximació que ha sigut seleccionar la classe mare de la taula que contenia la dada que volia recollir, transformar tot el text que contenia en un array i seleccionar l'element que estigues en la posició que m'interessava. Aquest mètode ha funcionat be inicialment, però de seguida s'ha mostrat ineficient davant pàgines que no respectaven estrictament l'estructura, provocant errors en les dades reduïdes.
  Així doncs, he fet una modificació de manera que el que fes el programa fos, un cop tenia la llista d'elements, buscar un element de control conegut i pròxim a la dada que volia recollir (per exemple, el text '*Centres on s'ofereix la titulació'* quan volia recuperar el nom de Facultat), i un cop localitzada, obtenir el seu índex dins la llista i finalment recuperar el valor desitjat a partir d'aquesta posició a mode de referència relativa. Aquesta segona aproximació ha resultat molt més robusta i m'ha permès recuperar informació de totes les titulacions de forma eficient.
- He afegit un control de presència de les dades, perquè he vist que algunes titulacions no disposaven de totes les dades o en alguns apartats les observacions només arribaven fins el 2015 o anys anteriors. En aquests casos, he decidit xequejar si existia la dada per 2016 i en cas que no, en comptes de deixar buit el camp, he inserit en el lloc del valor la cadena "NA". D'aquesta forma, simplificaré la comprensió de les dades i la fase de preparació de dades a partir del dataset.
- Finalment, assenyalar que he definit un retard de forma deliberada al algorisme, de manera que entre l'escaneig d'una pàgina i la següent es produeixi una espera de 5 segon. El motiu d'això es tenir un comportament "amable" amb el servidor que escanejo i mirar d'evitar produir problemes o que em pogués denegar la connexió.

10. Dataset

Disponible amb el nom 2016-UABDataset.csv, utilitzant com a separador el caràcter ";", al directori arrel del repositori de Github *ScraperTCVD*.

Al principi de generar-lo, he observat que tot i fer-ho correctament m'apareixia una línia en blanc entre cadascuna de les línies del fitxer. He descobert que era un problema amb la funció que obria el fitxer per escriptura i amb un paràmetre afegit ho he pogut solucionar.

He definit de forma específica que el separador sigui el ";" en comptes de "," perquè he pogut comprovar que així es pot obrir directament el Excel i recupera els camps bé. Penso que així facilito el veure les dades de manera fàcil inicialment i que si per carregar-ho en qualsevol software d'anàlisi com R, es indiferent un separador o altre perquè disposen de funcions per carregar tots dos.