Pràctica 1

M2.951 – Tipologia i cicle de vida de les dades

Alumne: Miquel Ribó i Pal

1. Títol del data set

Oferta formativa de graus de la Universitat Politècnica de Catalunya.

2. Subtítol del data set

Oferta formativa de graus de la Universitat Politècnica de Catalunya, amb informació complementària (lloc web del grau i càrrega lectiva total).

Si el lloc web del grau dóna informació en format html del pla d'estudis del grau, es consigna, per a cada assignatura, a més a més de les dades anteriors, el nom, semestre, tipus i càrrega lectiva de l'assignatura. Si existeix, es dóna també l'enllaç al document de descripció de l'assignatura i la menció a què pertany.

3. Imatge

Fig. 1. Imatge per al data set. [Font: Fragment modificat d'una captura de pantalla de https://www.upc.edu/ca realitzada el 27-10-2018]

4. Context

El conjunt de dades conté informació acadèmica sobre els graus que ha ofert durant el curs 2018-2019 la Universitat Politècnica de Catalunya.

La informació continguda al data set no és de gaire interès per a potencials alumnes que estiguin interessats en cursar un grau a la UPC. Aquesta informació pot ser més aviat útil per a membres de la comunitat universitària que vulguin treure conclusions sobre l'estructura de l'oferta formativa de graus d'una universitat de caire tècnic de cara a comparar-la amb la pròpia, reformar plans d'estudis, proposar-ne de nous, etc..

5. Contingut

Les dades recollides tenen una estructura natural d'agregat, que s'adiria més bé a un format JSON que no pas csv. Aquest fet ocasiona que hi hagi força redundància a alguns camps de les dades recollides.

Totes les dades s'han desat com a cadenes alfanumèriques delimitades per cometes dobles ("). Les dades no disponibles s'han desat com cadenes buides ("").

Els camps del data set són:

- **Nom grau**. Nom del grau
- **URL grau**. URL del lloc web on hi ha la informació acadèmica del grau
- **Crèdits grau**. Càrrega docent total del grau, mesurada en crèdits ECTS (un curs acadèmic complet té una càrrega de 60 crèdits ECTS). Per a alguns graus aquesta informació no està directament disponible

(Si un grau no té disponible informació detallada de la seva estructura acadèmica en format html al seu lloc web, les dades anteriors són les úniques que es consignen per al grau, en una sola fila del data set que conté valors buits per a la resta de camps que s'enuncien a continuació.)

- **Nom assig**. Nom d'una assignatura del grau
- **URL assig**. URL des d'on es pot obtenir informació detallada sobre l'assignatura. Aquesta informació, bé apunta a un document en format pdf, bé no està disponible
- **Crèdits assig**. Càrrega de treball de l'assignatura expressada en crèdits ECTS
- **Tipus assig**. Tipus d'assignatura (obligatòria, optativa o projecte (treball final de grau))
- **Semestre assig**. Semestre del grau en què s'imparteix l'assignatura. Les assignatures optatives sovint es troben assignades al web del grau a un semestre arbitrari o fictici
- **Menció assig**. En plans d'estudis que tenen mencions o especialitats, indica aquelles assignatures que són pròpies d'una menció determinada (és a dir, que no són cursades per totes les mencions). Per a les assignatures comunes, o aquelles de plans d'estudis sense mencions, conté cadenes buides.

(Si un grau té disponible informació detallada de la seva estructura acadèmica en format html al seu lloc web, tots els camps anteriors es consignen per a cada assignatura. Per tant, els camps **Nom grau**, **URL grau** i **Crèdits grau** es repliquen redundantment per a cada assignatura.)

Com es comenta més avall, a la descripció del codi, el data set no conté informació detallada d'assignatures de graus impartits per centres adscrits o de dobles titulacions amb centres d'altres universitats perquè els seus llocs web tenien una estructura molt divergent que n'impossibilitava un procés de crawling/scraping sistemàtic. L'absència al data set de la informació d'assignatures corresponent a graus impartits per centres adscrits és fins i tot desitjable perquè en la seva pràctica totalitat corresponen a titulacions en l'àmbit de la gestió, i no pas en l'àmbit desitjat, que és el tècnic. Quant als graus de doble titulació, la seva estructura acadèmica no té perquè reflectir la política acadèmica de la UPC, puix que la seva estructura ha de ser de consens amb la de les altres institucions involucrades. Per tant, la seva absència no es greu si el que es vol analitzar són les titulacions que han estat concebudes íntegrament segons la política acadèmica de la universitat.

6. Agraïments

Les dades són dades públiques extretes del web propietat de la Universitat Politècnica de Catalunya. El web crawler/scraper emprat ha respectat en tot moment les indicacions del propietari manifestades al fitxer `robots.txt`.

La selecció, extracció i presentació de les dades en format csv és pròpia d'aquesta pràctica.

7. Inspiració

Soc professor de Propagació electromagnètica i de Circuits de microones als graus de La Salle Campus Barcelona – URL. Al llarg dels anys, ja sigui en claustres departamentals, en grups de treball per a la definició de nous estudis o la reforma dels existents, o bé en converses de cafè, he sentit parlar recurrentment de com s'havien d'organitzar els graus que s'hi imparteixen (tècnics també, i alguns coincidents amb els de la UPC), és a dir, de

- quin era el nombre de crèdits ECSTS ideal que havien de tenir les assignatures,
- si aquest havia de ser fix o adaptar-se a la seva complexitat,
- si les assignatures optatives podien tenir creditatges menors i ser més abundants,
- si l'oferta d'assignatures optatives era adequada,
- si calia fer estudis amb molta optativitat (per tal que els alumnes se'ls adaptessin a les seves preferències) o estudis amb mencions o especialitats pre-definides (i poca o gens optativitat),
- quina havia de ser la política de reciclatge o compartició d'assignatures entre graus,
- si calia limitar el nombre d'assignatures (i augmentar-ne el creditatge) durant el primer curs per tal que els alumnes poguessin fer més fàcilment la transició del batxillerat a la universitat,
- quina havia de ser la càrrega/durada del treball/projecte final de grau,
- etc..

Tenir dades estadístiques rigoroses de quines opcions han pres altres universitats del mateix àmbit educatiu pot ser un factor a tenir en compte la propera vegada que calgui tornar a reflexionar sobre les qüestions anteriors, o altres de similars. El data set obtingut, a partir del qual es poden respondre moltes de (sinó totes) les qüestions anteriors (i moltes altres) per a la UPC, fóra un primer pas (caldria replicar el procés amb altres universitats de l'àmbit tècnic) per a prendre decisions (o si més no per a fer argumentacions) basades en dades estadístiques rigoroses.

8. Llicència

El data set obtingut tindria en principi una circulació molt restringida a l'interior de la meva empresa. Pel fet de dipositar-lo a https://github.com/ podria ser, no obstant això, accedit de manera més general per un nombre potencialment major d'usuaris per a fer-ne usos diferents al proposat.

Totes les llicències proposades a l'enunciat permeten als futurs usuaris l'ús i transformació dels continguts del data set. Donada la poca rellevància de les dades obtingudes (no corresponen per exemple a un estudi complet de tota l'oferta educativa tècnica d'un país) i la seva disponibilitat (són totes elles dades accessibles per al públic general al lloc web de la UPC), no sembla massa raonable imposar-hi restriccions d'ús o de cita.

Per tant, semblaria adequat publicar-les sota llicència CC0 1.0 Universal (Public Domain Dedication), amb què

- el data set i les dades que conté poden ser emprades lliurement (copiades, modificades i distribuïdes lliurement, fins i tot per a fins comercials, sense citar la font o demanar permís) per tothom,
- qui pugui fer ús de les dades no pot donar a entendre de cap de les maneres que jo, com a compilador del data set, dono suport a l'ús que en fa, i
- les dades s'ofereixen al domini públic sense cap garantia sobre la seva exactitud, i es declina qualsevol responsabilitat quant als usos que es puguin fer de les dades.

Fonts:

- https://creativecommons.org/publicdomain/zero/1.0/
- https://creativecommons.org/licenses/by-nc-sa/4.0/
- https://creativecommons.org/licenses/by-sa/4.0/
- https://opendatacommons.org/licenses/odbl/summary/

9. Codi

El programari que recull les dades (`M2_951_Practica1__Web_scrapper.py`) primerament fa un procés de web crawling/scraping del lloc web https://www.upc.edu/ca/graus/ (Fig. 2) (funció `crawlscrape_url_principal()`)per a obtenir les URL dels llocs web on la UPC consigna la informació acadèmica pública de cada grau que oferta.

Fig. 2. Aparença del lloc web d'on cal obtenir les URL dels llocs web dels diferents graus. [Font: Fragment de captura de pantalla de https://www.upc.edu/ca/graus/ realitzada el 27-10-2018]

A continuació realitza un procés de web crawling/scraping (funció `crawlscrape_url_grau()`) de cadascun dels llocs web anteriors per a obtenir la informació rellevant sobre cada grau. Els llocs web dedicats a cada grau (per exemple el de la Fig. 3) poden tenir estructures lleugerament heterogènies. Els que donen informació sobre dobles titulacions impartides amb altres centres o per centres adscrits estan enllaçats a altres llocs web (els propis dels altres centres), i no s'han seguit per tenir estructures web molt divergents les unes de les altres.

Fig. 3. Aparença d'un lloc web (del Grau en Enginyeria Informàtica) on cal fer el procés de web crawlling/scraping. [Font: Fragment de captura de pantalla https://www.upc.edu/ca/graus/enginyeria-informatica-barcelona-fib realitzada el 27-10-2018]

Els dels graus definits únicament per la UPC tenen per contra un format web molt estructurat, que n'ha permès extreure fàcilment les dades d'assignatures. Les úniques variacions de format que presenten aquests llocs web son la presència o no de mencions als plans d'estudi, la disponibilitat o no d'una URL d'informació addicional (en format pdf) per a les assignatures llistades, i la disponibilitat o no de la càrrega total en crèdits ECTS del grau. Finalment cal dir que hi ha força estudis amb el mateix nom, oferts en seus diferents i amb programes d'estudi diferents, que es diferencien al data set per la tupla (**Nom grau**, **URL grau**).

Per a les assignatures que tenen com a atribut la URL del document pdf de descripció de l'assignatura, el procés de web crawling/scraping pot continuar, opcionalment, seguint l'enllaç i descarregant i

desant al disc dur aquest arxiu binari (funció `descarrega_pdf()`). Al programa principal desenvolupat, aquesta opció està desactivada per defecte per tal que el procés de generació del fitxer csv acabi en un temps raonable, però es pot activar fàcilment canviant un paràmetre de la funció `crawlscrape_url_grau()`, o modificant el programa principal per tal que desi només els documents de graus que siguin d'especial interès per a nosaltres, o de grups d'assignatures (per exemple totes les que contenen "càlcul" i són de primer curs) que vulguem analitzar després amb detall, potser fent un procés d'scraping dels documents pdf desats. Tot això pot fer-se també al programa d'explotació posterior de les dades desades al fitxer csv perquè aquest darrer conserva tant les URL dels llocs web dels graus com les dels documents pdf de les assignatures.

Per tant, el procés de crawling/scraping que es realitza per a l'obtenció de totes les dades desitjades visita i obté dades (dades finals o noves URL a visitar) de llocs web seguint una estructura en arbre com la que es mostra esquemàticament a la Fig. 4.

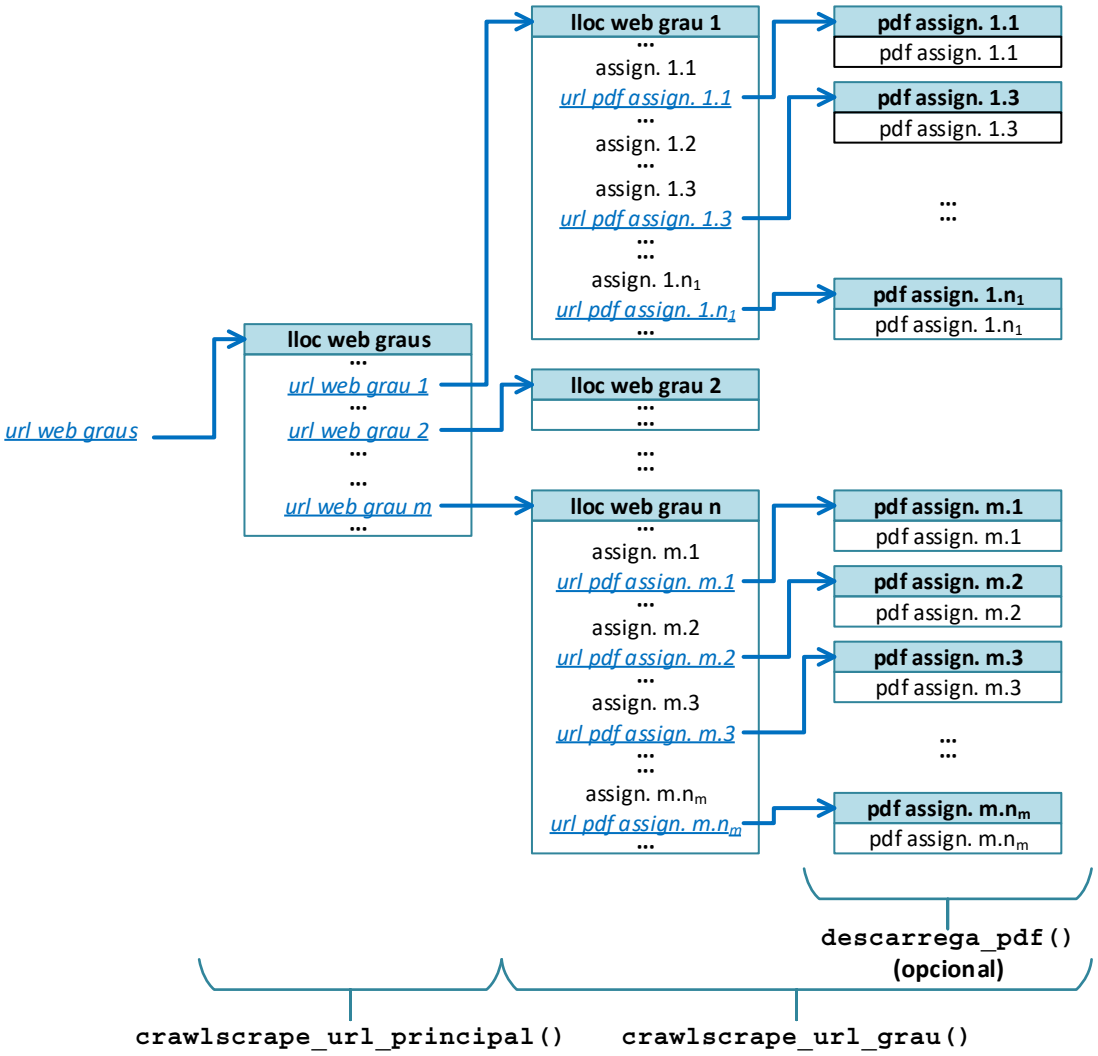

Fig. 4. Procés de web crawling/scraping complet que es realitza per a l'obtenció de les dades, amb les funcions involucrades. Notes: (1) Hi ha webs de graus que no tenen estructura d'assignatures. (2) Hi ha assignatures que no tenen URL del document pdf associat. (3) L'obtenció dels documents pdf és opcional.

El codi s'ha generat amb Python 3.6. Consta dels elements següents:

a. **Classe `Temporitzador`**. Classe que implementa un temporitzador que pot funcionar segons dos esquemes de retard:
   i. Relatiu: s'assegura que hagi transcorregut un interval de temps mínim des de la darrera crida al temporitzador. A la primera crida no s'origina cap retard
   ii. Absolut: origina un retard fix en ser invocat el seu mètode `espera()`.

   S'ha programat amb dos esquemes de retard per tal de tenir més versatilitat a l'hora de programar els espais temporals entre peticions web. La versió actual del codi empra ambdós esquemes de retard.

b. **Funció `descarrega_url()`**. Funció molt general que obté els continguts d'una URL donada, realitzant un control d'errors (davant un error o situació no desitjada, la funció acaba de manera controlada i retorna un codi d'error). Per a errors associats amb problemes al servidor, permet re-intents, però no per a problemes de connexió menys específics. La funció respecta les restriccions d'accés imposades al fitxer `robots.txt`.
   La funció pot retornar tant dades en format text com binari. Per tant, és igualment adequada per a descarregar codi html que imatges o arxius en format pdf si s'especifica que retorni dades en format binari.
c. **Funció `descarrega_pdf()`**. Funció que descarrega un arxiu en format pdf (o, de fet, qualsevol altre arxiu binari com ara una imatge) de la URL indicada i el desa al directori (existent) amb el nom especificat.
d. **Funció `crawlscrape_url_principal()`**. Funció que realitza un procés de web clawling/scraping del lloc web https://www.upc.edu/ca/graus/ per tal d'obtenir-ne, resseguint el codi html, les URL dels llocs web de tots els graus oferts per la UPC.
e. **Funció `crawlscrape_url_grau()`**. Funció que realitza un procés de crawling/scraping de cadascun dels llocs web obtinguts per la funció anterior. La funció s'adapta a una certa heterogeneïtat en els formats dels llocs web. Retorna un diccionari amb la informació acadèmica rellevant obtinguda de cada lloc. Si es desitja, pot desar també, al directori especificat, els arxius pdf de les assignatures que en tenen i que les descriuen. Per a graus realitzats per centres adscrits, o corresponents a dobles titulacions interuniversitàries, en què el lloc web de la UPC bàsicament apunta a un altre lloc web amb format molt heterogeni o a un document pdf, no segueix aquests enllaços i només recupera la informació bàsica que proveeix el lloc web de graus de la UPC.
f. **Programa principal**. Empra les funcions `crawlscrape_url_principal()` i `crawlscrape_url_grau()`, així com la classe `Temporitzador` per a obtenir les dades rellevants de l'oferta de graus de la UPC, que desa a un fitxer de text amb format csv, `dades_graus_upc.csv`. L'opció programada no desa els arxius pdf associats a les assignatures per tal de mantenir un temps d'execució moderat (entre descàrrega i descàrrega d'arxiu pdf la funció `crawlscrape_url_grau()` deixa intervals de guarda per tal de no tenir problemes amb el servidor). S'ha comprovat que la descàrrega dels documents pdf de les assignatures funciona correctament activant-la per a graus concrets en la fase de proves del programari.

10.Data set

El data set obtingut, `dades_graus_upc.csv`, recopila la informació disponible al lloc web de graus de la UPC, per als 77 graus que ofereix.

Emprant aquest data set es poden respondre preguntes com les que es mencionaven més amunt. Només a tall d'exemple, la resposta a dues d'aquestes preguntes, per a la UPC, és (l'arxiu `M2_951_Practica1__Web_scrapper__exemples_d_us.py` conté el codi per a respondre-les):

a. El percentatge d'assignatures optatives que ofereix la UPC és aproximadament del 45% (respecte del total d'assignatures).
b. Les distribucions de crèdits ECTS per a assignatures obligatòries, optatives i treballs finals de grau es poden veure a la Fig. 5.

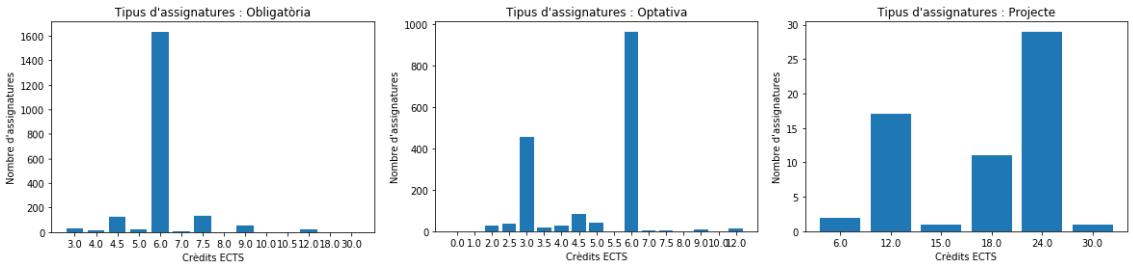

Fig. 5. Distribucions de crèdits ECTS per a assignatures obligatòries, optatives i projectes de fi de grau per als graus de la UPC.

Per tant, es pot observar com s'opta majoritàriament per assignatures de 6 crèdits ECTS tant a les assignatures obligatòries com a les optatives. A les optatives, però, hi ha molta oferta d'assignatures de creditatge meitat, que permeten molta versatilitat als alumnes per a configurar el seu itinerari d'optativitat emprant, si més no parcialment, assignatures de curta durada/continguts molt concrets. Per als treballs de fi de grau, s'opta per treballs de força dedicació a per a la majoria d'estudis.

Bibliografia emprada i fonts consultades

A part de les fonts citades més amunt, s'han consultat les següents:

- R. Lawson, *Web Scraping withPython, Packt Publishing*, 2015
- L. Subirats i M. Calvo, *Web Scraping*, Universitat Oberta de Catalunya, 2018
- Llocs web dels diferents mòduls emprats al programari
- https://stackoverflow.com/