

Repositori d'articles publicats a la revista Quaderns de l'ICA

Tipologia i cicle de vida de les dades

Uoc

Universitat Oberta
de Catalunya

Màster de Ciència de Dades

La tècnica del web scraping tracta d'automatitzar l'exploració de llocs web per tal d'extreure'n les dades que contenen, informació que pot resultar d'interès per a molts diversos àmbits.

L'ICA és l'Institut Català de l'Antropologia i té com a principal objectiu treballar per difondre els coneixements i la pràctica de la disciplina antropològica i per crear un espai de discussió, investigació i aprofundiment científic de la societat i la cultura a Catalunya.

La nostra tasca ha consistit a apropar aquests dos elements.

Miquel Àngel Fraire Ferrer
Carla Manzanares Calvo

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

El nostre treball extreu la informació de la pàgina web de l'Institut Català d'Antropologia o ICA (www.antropologia.cat). Aquesta pàgina ofereix informació molt variada, però nosaltres ens hem centrat a explorar l'apartat "Quaderns" per tal d'arribar als articles publicats a cada número de la revista que edita aquest institució.. És a dir, la primera vista del lloc web ofereix un resum d'informació com ara l'agenda, les notícies, els grups de treball, les publicacions, etc. És doncs en aquest última vista on nosaltres hi hem posat el focus perquè dins d'aquest i, en un segon nivell, es troben els diferents quaderns publicats al llarg del temps per la institució. A més a més, dins de cada quadern, i en un tercer nivell, és on es troben els diferents articles que a nosaltres ens interessien.

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

El títol pel dataset és "RepositoriArticlesICA", per tal com el dataset resultant d'haver aplicat la tècnica de web scraping és un recull de tots els articles publicats als números que han anat sortint de la revista Quaderns editada per l'ICA (Institut Català d'Antropologia), i que són lliurement accessibles (i en la majoria de casos fins i tot descarregables a través d'un enllaç a la seva versió en PDF) des del web d'aquesta institució <<https://www.antropologia.cat/>>.

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.

Aquest dataset aplega informació relativa a tots els números de la revista Quaderns que edita l'Institut Català d'Antropologia i els articles que en formen part, juntament amb els enllaços descarregables de les seves versions en format PDF quan aquests eren disponibles, informació que és de lliure consulta a la pàgina web de la institució.

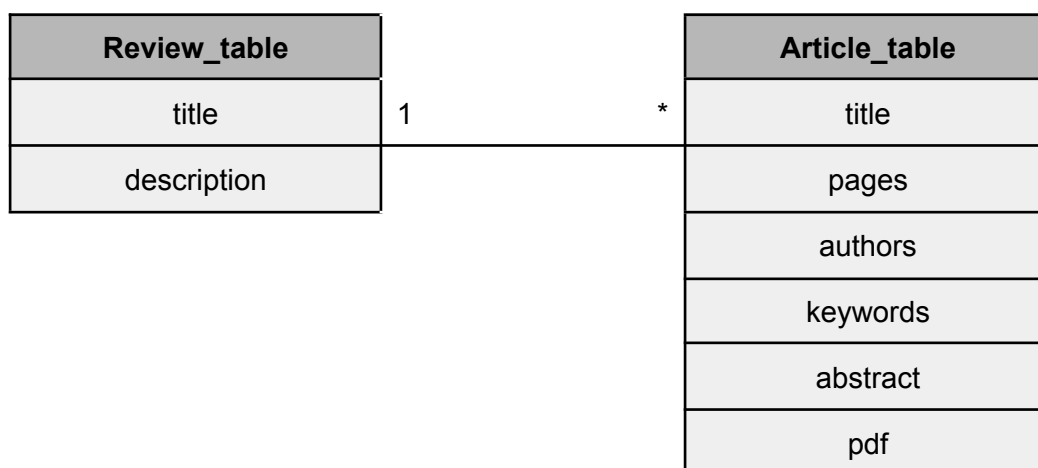
Val a dir que el nostre dataset es presenta contingut en un arxiu de format .csv o *comma separated value* tal com demanava l'enunciat de la pràctica, però que ens hem permès la llibertat d'establir com a criteri de separació dels camps un caràcter diferent de la tradicional coma (o, de vegades, el menys freqüent punt i coma), atès que la majoria dels camps amb què s'estructura el conjunt de dades resultant és de tipus textual i en molts casos inclou comes (per exemple, quan un article té diversos co-autors, o bé en el cas de les paraules clau). Aquesta circumstància desaconsella de mantenir la coma com a criteri de separació, perquè aleshores s'estaria separant en nous i imprevistos camps les cadenes de caràcters amb comes incloses que, no obstant això, constitueixen un únic i singular valor. Com que el punt i coma i qualsevol altre signe lingüístic tampoc no eren procedents, hem optat doncs per escollir el caràcter | com a criteri de separació.

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

El dataset resultant es presenta en una sola taula com la que es pot veure a sota:

ATRIBUTS	TIPUS DE DADES
review_title	string
review_description	string
article_title	string
article_pages	string
article_authors	list
article_keywords	list
article_abstract	memo
article_pdf	url

Ara bé, la idea que hi ha al darrere és el d'una base de dades relacional amb dues taules que mantenen una relació del tipus *un a diversos registres*. En el cas que ens ocupa, la informació relativa a cada número o edició de la revista estaria representada per la taula *ReviewTable*, i aquests podrien contenir diversos articles, la informació relativa als quals estaria representada per la taula *Article_table*. A sota hi podeu veure l'esquema que exemplifica aquest tipus d'estructura:



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

Aquests són els camps que inclou el dataset resultant i que tipifiquen tant el número de revista com els articles que contenen, així com la relació de contingut i continent que s'estableix entre aquests dos elements:

- **Review Title:** títol del número de la revista.
- **Review Description:** breu descripció del número, no sempre hi és present, però quan hi consta se sol emprar per proporcionar el noms dels coordinadors del número.
- **Article title:** títol de l'article.
- **Article pages:** pàgines que ocupa l'article dins el seu número corresponent.
- **Article authors:** autors de l'article.
- **Article keywords:** paraules clau.
- **Article abstract:** resum de l'article.
- **Article pdf:** enllaç o url d'accés i descàrrega de la versió en PDF de l'article.

Les dades recollides pertanyen a publicacions que abasten un període temporal que va des de l'any 1980 fins a l'actualitat. Per descomptat, tots aquells exemplars que van veure la llum abans de l'aparició d'internet o, més concretament, abans que l'ICA tingués el seu propi lloc web, es varen incorporar digitalment amb posterioritat a la seva publicació en format paper. Tot i que la data de publicació del darrer exemplar és el 2020, ens consta que l'ICA segueix mantenint tant la revista *Quaderns* com la pàgina web, per tant cal suposar seguiran digitalitzant-ne i publicant al lloc web els propers exemplars. Val a dir que, tret que l'estructura de les pàgines que contenen aquestes publicacions quedés alterada a causa de futures modificacions del lloc web, el codi del nostre web scraper estaria preparat per seguir extraient informació dels números de la revista *Quadern* i els articles que en formaran part en properes edicions, a mesura que des de l'ICA els vagin incorporant al web.

Tal com ha estat dit abans, les dades s'han extret mitjançant la tècnica de web scraping, la qual automatitza la tasca de navegar i extreure les dades contingudes en un lloc web. A tal efecte hem programat un senzill web scraper mitjançant el llenguatge de programació Python tot fent ús de la biblioteca BeautifulSoup, dedicada al web scraping, així com de la biblioteca Pandas, dedicada a la gestió de conjunts estructurats de dades.

Per tal d'extreure amb èxit les dades pertanyents als camps que hem enumerat més amunt, la implementació del nostre web scraper ha hagut de superar tres dificultats principals: (a) l'exploració del web a diversos nivells de profunditat, (b) la paginació de la vista que dona accés als diversos números de la revista, i (c) la doble estructura divergent en el codi html de les pàgines que contenen els articles.

- (a) **Exploració del web a diversos nivells de profunditat.** La informació que volíem obtenir està disposada en tres nivells de profunditat, els quals es corresponen a l'anterior esquematització del dataset en taules relacionades per un vincle d'*un a diversos registres*. Un primer nivell és la pàgina inicial d'exploració, que consisteix en una vista on hi ha disposats els enllaços a les pàgines que contenen la informació de cada número publicat de la revista Quaderns. El segon nivell de profunditat es correspon a aquestes pàgines relatives a cada número, les quals al seu torn exposen en una vista pròpia els enllaços a cada article que forma part del número de la revista que representen. Finalment, el tercer nivell de profunditat és la pàgina relativa a cada article, des d'on es té accés a l'enllaç de descàrrega de l'article en format PDF. Com que el segon i el tercer nivells de profunditat contenen una sèrie d'enllaços cap a les pàgines dels nivells subsegüents, hem hagut de programar el web scraper per tal que efectués dues iteracions niuades, la primera per accedir, explorar i extreure dades de totes les pàgines relatives als números de la revista, i la segona per accedir, explorar i extreure dades de totes les pàgines relatives als articles continguts a cada número de la revista.
- (b) **Paginació de la vista que dóna accés als diversos números de la revista.** El primer nivell de profunditat que conté la vista amb els enllaços a cada número de la revista manté una estructura paginada, això és que aquesta vista té una limitació de fins a 20 enllaços a números de revista, a partir dels quals els enllaços que superen aquest nombre queden continguts en una nova pàgina del mateix nivell, és a dir una pàgina germana. Per tant, prèviament a les iteracions de primer i segon nivell explicades al punt anterior, hem hagut de preveure que tot el codi iteri l'exploració sencera per nivells de manera que aquesta pugui abastar totes les pàgines germanes del primer nivell. Tot i que es dóna el cas que actualment només n'hi ha dues, el nostre web scraper està preparat per seguir l'exploració de la primera pàgina fins a la darrera d'aquest sistema de paginació, en cas que, com és de preveure, se amb el temps se n'hi vagin afegint a mesura que el web incorpora nous números de la revista.
- (c) **Doble estructura divergent en el codi html de les pàgines que contenen els articles.** Una característica que fa possible l'exploració i extracció automàtica de dades contingudes en pàgines web és la regularitat en l'estructura d'aquestes. En el nostre cas però, ha resultat que a partir del número 30 el patró html de les pàgines de segon i tercer nivells canvia completament. Per tant, el nostre codi es veu obligat, d'una banda a reconèixer quines pàgines segueixen una o altra estructura (per sort, aquest canvi de format ve acompanyat d'un canvi en la url de les pàgines, fet que n'ha facilitat implementar el reconeixement), i d'altra banda a aplicar estratègies d'exploració diferents en cada cas. No cal dir que hem hagut estudiar i identificar l'estructura d'etiquetes que contenen les dades a extreure en cada cas.

Pel que fa a la tria de la informació a extreure, hem seguit dos criteris: d'una banda, les dades que millor poguessin tipificar els dos elements en què s'estructura el nostre dataset, que són els números de la revista i els articles de cada número; de l'altra, evitar la redundància. En el primer cas cas, hem descartat dades que no aportaven informació rellevant o no directament relacionada amb els esmentats elements, com ara la data d'incorporació digital o bé enllaços cap a d'altres apartats del lloc web. En el segon cas, hem descartat dades que apareixien repetides en més d'un nivell de profunditat, com per exemple els autors de cada article, presents al llistat de tots els articles a nivell de número de revista al qual pertanyen, però també presents a dins de la pàgina de tercer nivell que se centra en l'article en qüestió.

Acabarem aquest apartat indicant que els camps que hem seleccionat per extreure'n les dades, alguns cops no contenien la informació esperada sinó que estaven buits. Per evitar llistes de camps amb quantitats dispars d'elements, que a l'hora de construir el dataset ocasionarien problemes, ens hem vist obligats a incorporar en el codi condicionals que, en cas de no detectar les esmentades dades, omplissin igualment l'espai reservat a la llista amb un valor nul.



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.

Per una banda, els propietaris de les dades són l'ICA i els diferents autors que han redactat els articles. El recull d'aquests autors es poden trobar en el dataset resultant. Per altra banda, a l'hora de treballar amb el codi, ens hem guiat per l'exemple de dos treballs anteriors similars i dos manuals, a més a més, d'algun videotutorial i webs comunitàries de resolució de problemes informàtics.

BIBLIOGRAFIA

- Laia Subirats, Mireia Calvo (2109): Web scrapping. Barcelona, FUOC
- Richard Lawson (2015): Web Scrapping with Python. Birmingham, Packt Publishing

WEBGRAFIA

-  Web Scrapping with Python - BeautifulSoup Crash Course - videotutorial sobre com fer webscapping amb Python publicat per freeCodeCamp.org
-  Python Tutorial: Web Scrapping with BeautifulSoup and Requests - videotutorial sobre com fer webscapping amb Python publicat per Corey Schafer el 8/11/2017.

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.

Com a científics socials que som, ens interessava enfocar-nos en la nostra branca i mirar d'assolir una interdisciplinarietat entre la informàtica i les ciències socials. Se sap que el món de les socials és el més abandonat, el germà mitjà de les disciplines; i hem pensat que és interessant començar a fer reculls de BBDD del que les diferents pàgines web ofereixen al respecte. Per això, i particularment perquè una de nosaltres és antropòloga, hem escollit l'ICA.

Cal afegir que, així com en altres formacions es familiaritzen amb conceptes d'estadística o algunes altres inclús treballen conceptes d'informàtica; l'antropologia no fa ni una aproximació al que podrien ser una mica de matemàtiques, per això la relació informàtica-socials és tan nul·la, i per això volíem començar a donar-los valor des d'aquest enfocament.

Finalment, les preguntes que es podrien formular arrel del nostre dataset són moltes (com poden sorgir de qualsevol dataset), però alguns exemples són: "Quins autors treballen els conceptes de 'salut'?", "En quins quaderns es parla de migració?", "Quin és el tema més treballat per l'antropologia? (el més popular)", "I el segon?", "Quins temes són més treballats segons diferents períodes de temps?", "Quins investigadors han treballat junts?", etc.

8. **Llicència.** Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció:

Per al nostre dataset no hi han drets reservats - Released Under CC0: Public Domain License. Ens interessa que la base de dades sigui de domini públic de manera que altres científics i investigadors puguin basar-s'hi lliurement sobre, millorar i reutilitzar aquest contingut per a qualsevol propòsit sense restriccions segons la llei de drets d'autor o bases de dades.

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

Trobareu els arxius amb el codi del web scraper als següents repositoris de GitHub:

https://github.com/frairefm/UOC_DataScience_TipologiaCicleDades
<https://github.com/cmanzaca/PR1Tipologia>

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

Havent publicat el dataset obtingut al repositori Zenodo, l'identificador que us hi donarà accés és el 10.5281/zenodo.5654913.

Contribucions	Signatura
Investigació prèvia	MFF i CMC
Redacció de les respostes	MFF i CMC
Desenvolupament del codi	MFF i CMC