Universitat Oberta de Catalunya
Màster Universitari en Ciència de Dades

Tipologia i Cicle de Vida de les Dades

PRÀCTICA 1. Creació d'un dataset a partir de les dades contingudes a un lloc web (web scraping)

Salvador Sanchis Beneseit
Joan Manuel López Ruiz

Abril 2022

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

Durant els primers mesos del confinament provocat per la pandèmia de la Covid-19, moltes professions considerades 'no essencials' van veure completament aturada la seva activitat. A Catalunya, com a la resta d'Espanya i molts altres països del sud d'Europa, el sector cultural va patir les conseqüències de l'aturada de forma molt intensa. A les arts escèniques en particular, el teletreball no oferia una alternativa a la impossibilitat d'assistir als teatres: sense públic, l'accés de l'artista al teatre era un sense sentit. Aquesta impossibilitat d'accedir als teatres va situar les arts escèniques en una situació paral·lela a l'oci nocturn; malgrat que el teatre és cultura, esdevé tan prescindible com els restaurants i les discoteques. El tancament de teatres va perdurar ben passat la finalització del confinament total, i l'activitat escènica no es va recuperar del tot fins a la temporada 2020/21, un any després de l'inici de la pandèmia.

La dansa contemporània, una de les baules febles d'un ja molt precari paisatge cultural, no disposa d'estructures de suport estatal per contrarestar la inseguretat laboral associada a la professió, una inseguretat que és alta fins i tot en temps de normalitat. La docència, que és sovint una de les fonts alternatives d'ingressos de les professionals de la dansa, si no la principal, també es va interrompre de forma abrupta, i moltes artistes van provar de continuar la seva activitat docent a distància, en un intent desesperat de lluitar contra l'evidència que ensenyar a ballar sense estar present en un espai compartit és una mena de contrasentit.

En aquest context, el Mercat de les Flors de Barcelona, com altres teatres arreu del món, va iniciar una sèrie d'activitats i projectes accessibles a través de la seva pàgina web: streamings de conferències, mostra de gravacions íntegres d'espectacles de temporades anteriors… Un dels projectes va consistir a encarregar un text a un seguit d'artistes de l'àmbit de la dansa. El text en qüestió era una reflexió sobre les conseqüències del confinament. Les artistes que hi van participar eren convidades a respondre a la següent pregunta:

> *La pandèmia de la COVID-19 ha suposat un abans i un després en les nostres vides i ha col·locat el cos en un espai de fragilitat i d'amenaça constant, de manera que les relacions entre les persones i l'espai que ocupen cobren un relleu especial. Des d'aquesta consciència, de quina manera podem llegir avui el teu treball (la teva obra artística o de pensament) i quines possibilitats de desenvolupament veus per a la dansa?*

En total, 22 artistes van respondre a la invitació, i els textos que van escriure es van publicar al blog del Mercat de les Flors en els mesos de maig i juny del 2020.

La pàgina que dóna accés a aquests textos és la que hem triat pel nostre projecte de web scraping:

> https://mercatflors.cat/blog/reflexions-entorn-dun-confinament/

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

“La dansa confinada.”

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.

El nostre dataset recull tots els articles publicats al blog del Mercat de les Flors sota la categoria “Reflexions entorn d’un confinament”, incloent-hi la data de publicació, el títol de l’article, el text complet, una frase destacada del text, i la imatge que il·lustra l’article.

Aquest dataset podria ser un punt de partida per estudiar com van viure el confinament causat per la pandèmia de la covid-19 els professionals de la dansa, que es duria a terme mitjançant tècniques de Processament del Llenguatge Natural a partir de l’anàlisi dels textos escrits i publicats els mesos de maig i juny de 2020. A més del text complet de l’article, la frase destacada a la pàgina principal per a cadascun d’ells resumeix ja el caire de cada article.

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

Donat que estem recopilant textos d’articles, hem optat per crear un núvol d’etiquetes que destaqui les paraules més freqüents a cada article per a la representació gràfica del dataset resultant.

Si bé existeixen moltes aplicacions en línia gratuïtes que generen el núvol d’etiquetes a partir d’un text, com per exemple wordclouds https://www.wordclouds.com/, nosaltres hem optat per fer una mica de recerca i explorar com generar aquesta visualització amb python. Dins del script per a fer el scrap dels articles hem creat la funció nuvol_tags() que, donat un text que li passem per paràmetre, crea una imatge amb el núvol d’etiquetes. A la imatge adjunta mostrem el resultat de la funció nuvol_tags() sobre el text de la Lia Rodrigues.


vides
que
crec
cossos
més
lloc
han
Moltes
sempre
fragilitat
constant
temps
diferent
amenaça
sistema
femenines
sent
del
treball
consciència
essent
continua
considereen
hi
No
mai
persona
continuarà
lectura
d
neoliberal
faig
descartables
capitalista
tots
actiu
negres
hagi
pobres
estaran
trans

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

A la taula següent es descriuen els camps que s'inclouen al dataset generat. El període de temps de les dades comprèn des del 7 de maig de 2020, data de publicació del primer article, fins el 14 de juny de 2020, data de publicació dels darrers articles d'aquesta secció.

Quant a la recol·lecció de les dades, dins del tag <body> de la pàgina principal de la secció hi ha 22 tags <article> que contenen la informació de cadascun dels 22 articles publicats.

CAMPS DEL DATASET GENERAT	
ID	Codi identificador de l'article, número seqüencial des de 1 generat al bucle for que recull tots els articles
Títol	Títol de l'article, recollit del primer tag <a> fill del tag <article>
Data	Data de publicació de l'article, recollit del tag <time> fill del tag <article>
Text	Text íntegre de l'article, recollit de la pàgina especificada a l'enllaç a l'atribut 'href' dins del primer tag <a> fill del tag <article>. Dins de la segona pàgina, la pàgina de l'article, el text íntegre es recull dels strings dels tags <p> fills del tag <article>.
ImageURL	Enllaç a la imatge associada a l'article. Recollit de l'atribut 'src' del tag <img> fill del tag <article>
Frase	Frase resum de l'article publicada a la portada, d'interès per a conèixer què se'n destaca d'aquest article. Recollida del primer tag <p> fill del tag <article>

La pàgina principal de la secció "Reflexions entorn d'un confinament" del blog del Mercat de les Flors fa servir navegació amb scroll infinit, i per tant ha sigut necessari utilitzar les funcionalitats de la llibreria Selenium per a simular la navegació en la totalitat de la pàgina i poder obtenir tot el seu contingut.

Al script creat per a recopilar el contingut de la pàgina principal s'ha creat una funció desa_img() que recull i desa la imatge associada a cada article a partir de l'URL de la imatge. En aquest cas ha sigut suficient la llibreria request per a obtenir el contingut gràfic enllaçat. Hem decidit separar aquesta funcionalitat per facilitar la comprensió del codi. Dins d'aquest script s'ha desenvolupat també la funció extreu_text() que retorna el text complet de l'article a partir de l'URL de l'article. Com en el cas anterior, hem decidit separar aquesta funcionalitat per facilitar la comprensió del codi.

Pel que fa a les pàgines amb cada article, aquestes s'enllacen des de la pàgina principal de la secció. Aquestes pàgines no fan servir scroll infinit, i per tant s'ha pogut fer servir la llibreria

requests per a obtenir el seu contingut. El text complet de l'article es troba dins del tag <article> en etiquetes filles <p>, per tant ha calgut recuperar i concatenar els continguts dins d'aquest tag per a obtenir el text complet.

Amb l'objectiu de facilitar la comprensió del procés de recollida de la informació, s'inclouen a continuació les estructures imbricades entre els tag <article> tant de la pàgina principal com d'una de les pàgines d'articles. Per a la pàgina principal només s'ha agafat, a tall d'exemple, el contingut imbricat d'un dels tags <article>.

Estructura imbricada a la pàgina principal. Destacades en groc les etiquetes de les quals s'ha extret la informació:

```
<article class="even">
	<div class="inner">
	<h2><a href="https://mercatflors.cat/blog/sonia-gomez/" title="SÒNIA
GÓMEZ">SÒNIA GÓMEZ</a></h2>
	<img width="245" height="125"
src="https://mercatflors.cat/blog/wp-content/uploads/2020/06/retrat-So`nia-Go´me
z-245x125.jpg" class="attachment-thumbnail size-thumbnail wp-post-image"
alt="retrat So`nia Go´mez" loading="lazy" />
	<div class="metadata">
		<time datetime="2020-06-14T13:53:18+00:00">14 de juny de
2020</time> - <span><a
href="https://mercatflors.cat/blog/sonia-gomez/#respond">Cap
comentari</a></span>
	</div>
	<p>&#8220;El que sé és que la situació em donarà ales per millorar la
feble implicació que tenia amb la terra&#8221;</p>
	<div id="share">
		<!-- AddThis Button BEGIN -->
		<div class="addthis_toolbox"></div>
		<p><a class="addthis_button_compact">Compartir</a></p>
		<script type="text/javascript"
src="http://s7.addthis.com/js/250/addthis_widget.js#pubid=xa-5012b87952e78f17">
</script>
		<!-- AddThis Button END -->
	</div>
	</div>
</article>
```

Estructura imbricada a la pàgina d'un dels articles. Destacades en groc les etiquetes de les quals s'ha extret la informació:

```
<article class="main">
    <div class="inner">
	<h1>LIA RODRIGUES</h1>
	<img width="385" height="140"
src="https://mercatflors.cat/blog/wp-content/uploads/2020/06/lia-rodrigues-385x
140.jpg" class="attachment-medium size-medium wp-post-image" alt=""
loading="lazy" />
	<div class="metadata">
```

```
            <time datetime="2020-06-09T09:16:22+00:00">9 de juny de
2020</time> - <span><a
href="https://mercatflors.cat/blog/reflexions-entorn-dun-confinament-lia-rodrig
ues/#respond">Cap comentari</a></span>
      </div>
      <p><a
href="https://mercatflors.cat/blog/reflexions-entorn-dun-confinament/"><span
style="color: #0000ff;">&lt; Torna enrere</span></a></p>
      <p>Moltes vides i cossos han estat, estan i estaran sempre en lloc de
fragilitat i amenaça constant. Els cossos i vides d'aquells que es consideren
descartables en tots els temps i llocs: vides negres, vides trans, vides
femenines, vides pobres.</p>
      <p>No crec que hi hagi més consciència. El sistema capitalista neoliberal
continua essent més actiu que mai. Crec que la lectura del treball que faig
continuarà sent diferent per a cada persona.</p>
      <p style="text-align: left;">Resposta de la creadora Lia Rodrigues a la
pregunta:</p>
      <p style="text-align: left;"><em>La pandèmia de la COVID-19 ha suposat un
abans i un després en les nostres vides i ha col·locat el cos en un espai de
fragilitat i d&#8217;amenaça constant, de manera que les relacions entre les
persones i l&#8217;espai que ocupen cobren un relleu especial. Des
d&#8217;aquesta consciència, de quina manera podem llegir avui el teu treball
(la teva obra artística o de pensament) i quines possibilitats de
desenvolupament veus per a la dansa?</em></p>

      <div id="share">
            <!-- AddThis Button BEGIN -->
            <div class="addthis_toolbox"></div>
            <p><a class="addthis_button_compact">Compartir</a></p>
            <script type="text/javascript"
src="http://s7.addthis.com/js/250/addthis_widget.js#pubid=xa-5012b87952e78f17">
</script>
            <!-- AddThis Button END -->
      </div>
      <div id="post-comments">
            <!-- You can start editing here. -->
            <!-- If comments are open, but there are no comments. -->
            <div id="respond" class="comment-respond">
            <h3 id="reply-title" class="comment-reply-title">Deixa un
comentari <small><a rel="nofollow" id="cancel-comment-reply-link"
href="/blog/reflexions-entorn-dun-confinament-lia-rodrigues/#respond"
style="display:none;">Cancel·la les respostes</a></small></h3>
            <p class="must-log-in">Heu d'<a
href="https://mercatflors.cat/blog/wp-login.php?redirect_to=https%3A%2F%2Fmerca
tflors.cat%2Fblog%2Freflexions-entorn-dun-confinament-lia-rodrigues%2F">iniciar
la sessió</a> per escriure un comentari.</p>
            </div><!-- #respond -->
      </div>
      </div>
</article>
```

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.

El propietari del conjunt de dades és el **Mercat de les Flors**, donat que els textos que extraiem s'han publicat al seu blog. Existeix un fitxer robots.txt que permet l'accés complet a tots els robots.

Contingut del fitxer robots. txt – *https://mercatflors.cat/robots.txt*

```
User-agent: *
Disallow:

User-agent: msnbot
Crawl-delay: 10

User-Agent: bingbot
Crawl-delay: 10
```

Existeix un interès per conèixer com es va viure la situació excepcional provocada pel confinament a causa de la pandèmia de Covid19, com va afectar les persones i al seu estat anímic. Una cerca a Google per *sentiment analysis lockdown covid 19* dóna com a resultat diversos estudis en què s'ha copsat l'estat anímic de les persones durant el confinament de 2020 a partir de l'anàlisi dels textos publicats a les xarxes socials, especialment twitter, mitjançant tècniques de Processament del Llenguatge Natural. A continuació destaquem alguns d'aquests estudis:

> **COVID-19 Related Sentiment Analysis Using State-of-the-Art Machine Learning and Deep Learning Techniques**. Zunera Jalil, et at.
>
> https://www.frontiersin.org/articles/10.3389/fpubh.2021.812735/full

Sentiment Analysis of COVID-19 Nationwide Lockdown effect in India. N. Afroz, et al.

https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/en/covidwho-1218867

Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. Gopalkrishna Barkur, et al.

https://pubmed.ncbi.nlm.nih.gov/32305035/

Twitter Sentiment Analysis during the Lockdown on New Zealand. Smah Doeban Almotiri

https://publications.waset.org/10012359/twitter-sentiment-analysis-during-the-lockdown-on-new-zealand

Per tal d'actuar amb els principis ètics i legals a l'hora de fer scraping al blog del Mercat de les Flors, i prevenir el bloqueig de la navegació automàtica hem pres les següents mesures:

- Consulta del fitxer robots.txt, disponible a la url *https://mercatflors.cat/robots.txt*. El contingut d'aquest fitxer, inclòs anteriorment en aquesta pregunta, no desaconsella el rastreig automàtic de la pàgina per a extreure'n els continguts.
- Consulta de drets o llicències. S'ha consultat a la web del Mercat de les Flors l'existència de drets reservats (copyright) o altres llicències que limitessin o impedissin la publicació del material recopilat.
- Canvi del user agent. Com a mesura per a prevenir bloquejos durant el procés de scrap, hem configurat l'user agent de les peticions per tal que coincidís amb el d'un navegador web, en el cas del nostre script fem servir Mozilla Firefox sobre un sistema operatiu Ubuntu.
- Retard entre peticions. Com a mesura per a prevenir bloquejos durant el procés de scrap, hem afegit un delay de 2 segons entre les peticions per tal d'evitar col·lapsar el servidor.

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.

En relació a la temàtica de la pàgina web que explorem, i partint del context que hem presentat en el primer apartat, volem destacar l'interès del nostre objecte d'estudi com a testimoni dels efectes socials de la pandèmia. La particularitat d'aquest testimoni rau en dos factors principals, d'una banda, l'impacte concret del confinament sobre el sector de la cultura, i de l'altra banda, el fet que la dansa, entre totes les arts escèniques, estigui tan lligada a la presencialitat física,

als cossos compartint un espai comú, i en depengui. Podem dir que l'impacte del distanciament físic sobre l'activitat de la dansa anticipa i revela la importància de l'aspecte físic de les nostres interaccions socials, i en aquest sentit ens hauria d'interpel·lar ja no com a potencials consumidors de cultura, sinó simplement com a individus socials que som.

Més enllà, considerem que l'estructura del mateix testimoni, 22 textos agrupats sobre la mateixa temàtica, ens convida a convertir el material en un dataset. No es tracta d'una agrupació arbitrària d'elements, sinó que entenem que el fitxer final té una coherència temàtica i unicitat que el fa identificable i li dóna consistència.

Els 22 textos als quals accedim ja estan disponibles a través d'una sola pàgina, però cal accedir a cada text d'un en un a partir de la pàgina principal del blog del Mercat de les Flors. Així doncs, una primera funció del nostre dataset, que no per ser evident deixa de ser important, és que ens permet emmagatzemar i disposar de tots els textos en un sol fitxer, classificats i amb el seu contingut íntegre.

A partir d'aquesta funció d'emmagatzematge, entenem que el dataset ofereix un punt de partida ideal per extreure coneixement de la col·lecció d'articles. Concretament, imaginem operacions d'anàlisi textual com l'anàlisi de sentiments, l'anàlisi de temàtiques, classificació de text, registre de freqüència de paraules o la detecció de paraules clau. Totes aquestes operacions es veurien facilitades pel fet de disposar d'aquests textos en el format que resulta del nostre projecte de web scraping, classificats i accessibles a partir d'un sol document.

En l'apartat previ hem vist alguns exemples d'anàlisi de sentiments a partir de dades relacionades amb l'epidèmia de la Covid. En aquests casos, les dades sorgeixen de textos publicats a les xarxes socials per individus que provenen de col·lectius heterogenis. En relació amb aquests estudis, la particularitat del nostre projecte és, d'una banda, el fet que els textos provenen d'articles escrits per ser publicats en un blog, i de l'altra banda, que no es tracta de reflexions sobre l'impacte de la padèmia sobre la societat en general, sinó sobre un sector professional concret.

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

Ni a la web del Mercat de les Flors ni tampoc al seu blog no s'especifica cap tipus de llicència sobre els seus continguts o cap copyright explícit, per tant entenem que no hi ha cap limitació quant a l'ús dels textos.

De totes maneres volem proposar la llicència **CC BY-NC-SA 4.0 License** per al dataset resultant mencionant com a font original de les dades el Mercat de les Flors, de manera que hi hagi un reconeixement als creadors del contingut inclòs i no es permeti l'ús comercial d'aquestes dades, almenys sense el consentiment explícit del seu propietari.

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

El codi font amb el qual s'ha generat el dataset està disponible al següent repositori de Github:

https://github.com/salvasky/tcvd_pra1

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

L'enllaç del DOI obtingut a Zenodo en publicar el dataset és el **10.5281/zenodo.6448767**

L'enllaç per accedir-hi directament sobre la plataforma Zenodo és:
https://doi.org/10.5281/zenodo.6448767

El nom que hem donat al dataset és "La Dansa Confinada" i hem publicat el dataset real sense restriccions. A l'hora de publicar el dataset a Zenodo ens hem trobat amb què no ens oferia l'opció d'assignar-li la llicència escollida a l'apartat 8.

(*) Si existeix qualsevol impediment per publicar el dataset real, s'haurà de justificar aquesta situació i publicar a Zenodo un dataset simulat. En aquest cas, el dataset real es comunicarà al professor de forma privada (p.ex., enllaç de Google Drive).

Contribucions	Signatura
Investigació prèvia	Joan Manuel López Ruiz / Salvador Sanchis Beneseit
Redacció de les respostes	Joan Manuel López Ruiz / Salvador Sanchis Beneseit
Desenvolupament del codi	Joan Manuel López Ruiz / Salvador Sanchis Beneseit