1. Context.

TimeOut és una web (i també una revista) on per diverses ciutats mundials, entre elles Barcelona, conté articles sobre què fer, on menjar, quines activitats realitzar, etc. Aquests articles són actualitzats cada cert període de temps per tal que no continguin esdeveniments passats. Per tant, l'usuari sempre trobarà informació actual o per esdeveniments futurs.

El conjunt de dades que es pot extreure són les URL de tots els articles de TimeOut per la ciutat de Barcelona, en funció del tipus d'activitat (cultura, teatres, menjar i beure, arts i museus, viatges, entre d'altres), la data de publicació i les paraules clau per a poder fer una cerca més acurada del contingut de l'article.

2. Títol.

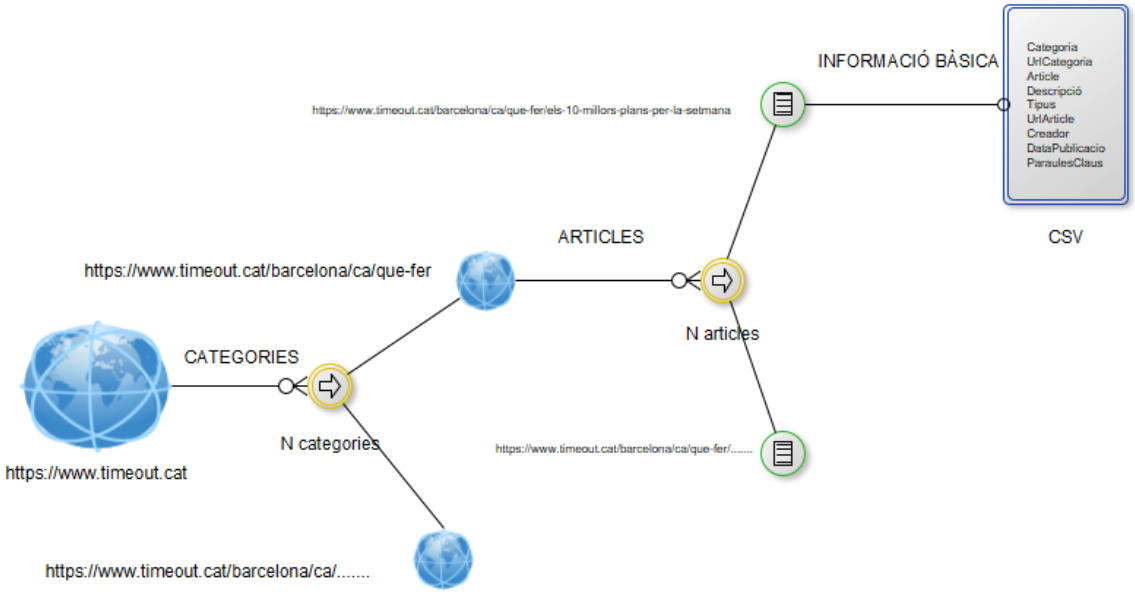
El títol definit per el *dataset* que generem, aplicant tècniques de web scraping a la URL https://www.timeout.cat, rep el nom "d'Oci a la ciutat de Barcelona", i el nom del fitxer en format CSV és "***ArticlesTimeOutCat.csv***".

3. Descripció del dataset.

Les dades que extraiem de la URL https://www.timeout.cat i que deixem en el *dataset* ***ArticlesTimeOutCat.csv*** contenen la informació bàsica dels articles que s'han publicat en la web del TimeOut sobre l'oci en la ciutat de Barcelona, segons la categorització que té la web.

4. Representació gràfica.

A continuació incloem una representació gràfica del projecte on es mostra com anem buscant la informació que inserim en el *dataset*:



5. Contingut.

Per a cada article, un registre en el conjunt de dades, guardem la següent informació:

- ***Categoria:*** Especifica en quina categoria de la web es troba l'article.
  Alguns dels exemples d'apartats actuals són *Cine*, *Llibres* o *Teatre i Dansa.*
- ***UrlCategoria:*** URL de l'apartat on es poden consultar tots els articles relacionats.
- ***Article:*** Nom de l'article.
- ***Descripció***: Descripció breu de l'article que surt com a Títol en la web.
- ***Tipus***: Defineix el tipus de de URL, en aquest cas tot són del tipus "Article".
- ***UrlArticle:*** URL de l'article.
- ***Creador:*** Enumera els creadors de l'article.
- ***DataPublicacio:*** Data de publicació de l'article amb el format següent:
  yyyy-MM-ddTHH:mm:ssZ
- ***ParaulesClaus***: Paraules Claus de l'article per la seva cerca dins de la web.

Es fa una cerca per tota la web, creada per *Time Out Spain Media SLU* el 14/09/2006.

No hem tingut en compte cap període de temps ja que la web conté informació d'esdeveniments presents i futurs, els articles passats no són visibles. Tot i així, si ho haguéssim considerat rellevant, ho hauríem fet condicionant la cerca per el camp del d*ataset DataPublicació*.

Les dades s'han recollit de forma recursiva mitjançant la implementació d'un script utilitzant el llenguatge de programació Python:

- Primer de tot hem descarregat la web de https://www.timeout.cat i obtingut les URL de totes les categories.

- Per a cadascuna de les categories, hem fet una cerca dels articles publicats.

- Per a cadascun dels articles, hem obtingut les dades que ens han semblat més rellevants per guardar en el *dataset*.

- Finalment, les dades s'han emmagatzemat en el CSV i per a cada article, s'ha creat un nou registre amb aquestes dades.

6. Agraïments.

Les dades del *dataset* que han estat obtingudes són propietat de *Time Out Spain Media SLU* (https://www.timeout.cat/barcelona/ca).

Per tal de seguir els principis ètica i legals de TimeOut, l'script implementat s'assegura que no està accedint a cap de les URL no permeses en el fitxer de robots.txt.

7. Inspiració.

Aquest conjunt de dades ens permet descobrir l'oci d'una ciutat. La web és interessant ja que estructura les dades en forma d'arbre (categoria → article → esdeveniments) i defineix les metadades de cada article de forma estandarditzada i fent ús del vocabulari controlat schema.org (https://schema.org/).

A més, amb aquest *dataset*, es pot permetre realitzar estadístiques de quines són les activitats d'oci que més es realitzen a la ciutat en un cert període de temps i així poder diversificar esdeveniments futurs.

8. Llicència.

La llicència escollida és "*Released Under CC BY-NC-SA 4.0 License*", els principals motius són els següents:

- No es permet l'ús comercial, ja que les condicions d'ús de TimeOut indica que l'única autorització que es té és la de visualitzar el contingut per ús personal i no comercial.

- Es permet compartir i adaptar el data set però sempre sota la mateixa llicència, per tal que les restriccions proporcionades pel propietari de les dades se segueixin complint.

9. Codi.

El codi per a generar el *dataset* es troba en el següent repositori del GitHub:

- https://github.com/mfontsanc/TIP-PRA1

10. Dataset.

L'enllaç DOI del *dataset* és el següent:

- https://zenodo.org/record/5648528#.YYVunWDMLIU

Contribucions	Signatura
Investigació prèvia	Isabel Barrera Benavent, Maria Font Sánchez
Redacció de respostes	Isabel Barrera Benavent, Maria Font Sánchez
Desenvolupament del codi	Isabel Barrera Benavent, Maria Font Sánchez