DOCUMENTACIÓ DE LA PRÀCTICA

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 events passats. Per tant, l'usuari sempre trobarà informació actual o per events futurs.

El conjunt de dades que es pot extreure són les URLs 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 de ***ArticlesTimeOutCat.csv***.

3. Descripció del dataset.

Les dades que extreiem 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 per la ciutat de Barcelona, segons els apartats de búsqueda que té la web.4. Representació gràfica.

5. Contingut.

Per a cada article, register 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 Titol en la web.

Tipus: Defineix el tipus de l'article. Dif

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 búsqueda dins de la web.

Es fa una búsqueda 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'events presents i futurs, els articles passats no són visibles. Tot i així, si ho haguéssim considerat rellevant, ho hauriem fet condicionant la búsqueda per el camp del DataSet *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's de totes les categories.
- Per a cadascuna de les categories, hem fet una búsqueda 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 URLs 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 → event) i defineix les metadades de cada article de forma estandarditzada i fent ús del vocabulari schema.org (https://schema.org/).

A més, amb aquest dataset, 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 events.

8. Llicència.

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.