

Context.

Per tal de generar el conjunt de dades, la informació s'ha recol·lectat a partir de la pàgina web <https://www.normacomics.com/>

La web és una tenda digital especialitzada en còmics, motiu pel qual podem veure una gran varietat de productes a la web.

Norma Comics és considerada una de les llibreries més important d'Europa en l'àmbit dels còmics. Es diferencien tres àrees de còmics: europeu, americà i manga. També disposen de més àrees com per exemple, seccions de còmic infantils, figures, regals... que per al nostre objectiu resulten menys interessants.

La secció que més ens interessa és la secció de novetats. La web considera novetat tot el que s'ha publicat els darrers dos mesos. Per al nostre conjunt de dades, volem obtenir les novetats de la web. La tenda és tot un referent, la llibreria ha guanyat el premi a la millor tenda de còmics de 2018. Per tant les dades obtingudes venen d'una font de dades de confiança.

Títol del dataset.

Novedades Cómic

Descripció del dataset.

La web Norma Comics considera com a novetat tot el que s'ha publicat els dos últims mesos. Per tant el conjunt de dades, ens mostra tots els còmics nous, desde fa dos mesos, que han sigut publicats a la web de Norma Comics.

A banda, es genera una carpeta amb les imatges de cada comic. D'aquesta manera es pot relacionar la imatge amb el còmic, ja que al fitxer csv guardem el nom de la imatge per poder fer aquesta relació.

Com que la web Norma Comics està en espanyol, hem considerat generar el conjunt de dades també en espanyol.

Representació gràfica.



La imatge l'hem dissenyada nosaltres mateix. Hem seleccionat 9 còmics de la secció de novetats de la web Norma Comics, i a més els còmics triats formen part de les àrees més representatives en les que ells treballen (europeu, americà i manga). Hem inclòs dues paraules, Novedades Comics.

Contingut.

Per cada còmic, que correspon a un registre del conjunt de dades, tenim:

ISBN	Número Estándar Internacional de Libres (International Standard Book Number en anglès). <ul style="list-style-type: none"> - Clau primària - Cadena de caràcters
Título	Títol del còmic. <ul style="list-style-type: none"> - Cadena de caràcters
Precio	Preu del còmic en divisa euros. <ul style="list-style-type: none"> - Numèric Nota: Amb dos decimals
Autores	Autor o autors del còmic. <ul style="list-style-type: none"> - Cadena de caràcters Nota: Un còmic pot tenir més d'un autor. Si té més d'un autor van separats per comes.
Colecciones	Col·lecció o col·leccions a la que pertany el còmic. <ul style="list-style-type: none"> - Cadena de caràcters Nota: Un còmic pot pertànyer a més d'una col·lecció. Si té més d'una col·lecció van separades per comes. Les col·leccions principals són Americà, Europeu i Manga.
Editoriales	Editorial o editorials del còmic. <ul style="list-style-type: none"> - Cadena de caràcters Nota: Un còmic pot pertànyer a més d'una editorial. Si té més d'una editorial van separades per comes.
Series	Sèrie o sèries a la que pertany el còmic. <ul style="list-style-type: none"> - Cadena de caràcters Nota: Un còmic pot pertànyer a més d'una sèrie. Si té més d'una sèrie van separades per comes.
Fecha de venta	Data de venda del còmic (DD/MM/YYYY). <ul style="list-style-type: none"> - Date
Formato	Format del còmic. <ul style="list-style-type: none"> - Cadena de caràcters
Páginas	Nombre de pàgines del còmic. <ul style="list-style-type: none"> - Numèric.
Nombre Imagen	Nom de la imatge del còmic. <ul style="list-style-type: none"> - Cadena de caràcters Nota: El scraper genera una carpeta "pictures_comics" on es troben totes les imatges dels còmics que formen part del conjunt de dades. D'aquesta manera, les podem relacionar per nom, ja que la nomenclatura de les imatges a la carpeta és la mateixa que a la variable "Nombre Imagen".

El període de temps de les dades és bimensual, perquè el que volem mostrar són els còmics novetat. Per tant, el conjunt de dades s'actualitzarà per exemple el dia 1 dels mesos impars a la matinada. Les dades es recullen de la web Norma Comics.

Agraïments.

Per a poder generar el conjunt de dades, s'ha accedit a la pàgina web de Norma Comics (<https://www.normacomics.com/>)

La web és una tenda digital especialitzada en còmics on va ser completament renovada l'any 2010. Norma Comics no sols té tenda digital, té tenda física en Barcelona i en altres parts d'Espanya. La tenda de Barcelona és considerada una de les llibreries més importants d'Europa en l'àmbit dels còmics, on podem destacar que l'any 2018 va guanyar el premi a la millor tenda de còmics.

Inspiració.

El conjunt de dades presentat respon a la intenció de mantenir actualitzada una base de dades de còmics. Cada dos mesos, amb el csv obtingut a partir de les dades de la web de Norma Comics, serà possible afegir totes les novetats.

En l'àmbit de mineria de dades, es podria crear un model capaç de respondre preguntes com quines editorials tenen més probabilitat de fer nous llançaments, per d'aquesta manera apostar per una editorial o una altra. També es poden construir models per fer un seguiment dels nostres autors o sèries preferits, o inclús predir el preu d'un còmic en el suposat cas que ens falti aquesta informació.

Llicència.



La llicència de Creative Commons BY-NC-SA permet ajustar i construir a partir de l'obra original amb finalitats no comercials, sempre i quan sigui reconeguda l'autoria i qualsevol nova creació es mantingui sota una llicència amb els mateixos termes.

Sobretot és important el fet de mantenir la no-comercialitat de l'eina donada la procedència comercial de les dades.

Codi.

Es pot consultar el codi font a través del següent enllaç o bé accedint a la carpeta codi.

<https://github.com/csanchisp/Web-Scraping/tree/master/Codi>

Dataset

Es pot consultar el dataset a través del següent enllaç.

<https://github.com/csanchisp/Web-Scraping>

Taula de contribucions

Contribucions	Signa
Recerca prèvia	Ivan Borrego García, Cristina Sanchis Puerto
Redacció de les respostes	Ivan Borrego García, Cristina Sanchis Puerto
Desenvolupament codi	Ivan Borrego García, Cristina Sanchis Puerto