Tipología y ciclo de vida del dato

Practica1

Nombre: Leandro Alonso Von Semasco

Titulo del dataSet:

Análisis de publicaciones diarias de páginas de actualidad en el mundo del videojuego/anime/manga

Subtitulo del dataSet:

Análisis de una página de noticias sobre temas de videojuegos, anime, manga y cultura japonesa. Como introducción a la extracción de noticias de una página de actualidad, con lo que obtener la principal temática de un determinado día, así como poder clasificarlo de forma sencilla por la autoría de la noticia con lo que poder ver la tendencia o evolución de dicha persona en sus escritos.

Imagen del dataSet:

Contexto.

El contexto serian las noticias de actualidad, dentro de una temática concreta como seria de la cultura japonesa y sus importaciones a occidente, el ver qué campo tiene más relevancia o mas actualizaciones en el día a día.

Contenido.

Los campos a manejar en el dataSet son únicamente la categorización de la noticia, el titulo, el autor y la fecha. El origen de los datos pertenece exclusivamente a la portada de la página web http://ramenparados.com/ para recogerlos se ha usado el programa en python que se encuentra en el repositorio de Github: https://github.com/leovs21/Prac1-Tipolog-a-y-ciclo-de-vida-de-los-datos

En el codigo se han realizado las importaciones de BeautifulSoup, request para poder trabajar con la pagina web y de csv para poder realizar el almacenamiento del dataSet.

En la función para enlazar el archivo .csv con el que vamos a trabajar, se ha tenido que dar un encoding especial “ utf-8-sig” ya que al tratar de temática japonesa, algunas palabras estaban adaptadas de la escritura original, por lo que tenían caracteres anormales y había que realizar una codificación distinta básica.

A continuación realizamos la extracción de información de la página tal y como lo hemos estudiado, examinando la pagina, los bloques que nos interesan y dentro de los cuales la forma de extraer los campos para los que estamos construyendo el dataSet. Almacenamos esta información por medio de la escritura en el .csv y finalizamos el programa.

La mejora para dicho programa, podría ser la comprobación de los artículos que vayan surgiendo si están en el dataSet, y si no es así, abriendo el archivo en modo escritura pero sin borrar la información, solo para añadir, e ir ampliando dicho conjunto de datos únicamente realizando la ejecución de forma periódica.

Agradecimientos.

Todo los datos obtenidos por el programa han sido obtenidos de la pagina web http://ramenparados.com/

Inspiración.

Aparte de la cultura japonesa que me apasiona, la labor de aquellas personas que publican de forma constante noticias sobre la actualidad me parece un duro trabajo. Dentro de este trabajo hay una evolución, en el desarrollo de cada uno de los autores de los mismos, pudiendo ver quien se centra en que temas en particular. Incluso entrando en cada uno de ellos llegar a ver como de extenso es el contenido, si estas deacuerdo con el matiz de una noticia o critica determinada.

Por medio de este programa espero realizar una primera aproximación a este estudio y forma de recopilar información al respecto.

Licencia.

Código.

El código desarrollado se encuentra en el repositorio de gitHub con el nombre de webScraping.py

DataSet.

El dataSet resultante del programa a sido subido al mismo repositorio que el código con el nombre de datos.csv