Escuela Técnica Superior de Ingeniería Informática

Grado en Ingeniería Informática

Tecnologías Informáticas

Acceso Inteligente a la Información

Sistema de búsqueda y recomendación de noticias

Autores:

Daniel Moreno Soto

Francisco Jesús Belmonte Pintre

Curso 2018-2019


UNIVERSIDAD DE SEVILLA

Índice general

1. Resumen

En el presente trabajo se expone el desarrollo de una aplicación web de noticias basada en el framework Django, que a su vez integra los 4 principales contenidos vistos en la asignatura de Acceso Inteligente a la Información, que son: BeautifulSoup para Web Scraping, Whoosh para realizar búsquedas, el propio Django para la aplicación web junto al framework Bootstrap, y un sistema de recomendación basado en contenido.

2. Beautifulsoup

Para la recolección de información en la web, se ha optado por extraer noticias de un medio de comunicación digital como es OkDiario. Este periódico digital tiene sus noticias categorizadas por secciones, además de tener muchas ya que abarca muchos temas como la política, economía, deporte, temas internacionales, etc.

Cada noticia consta del título de la noticia, el autor, la sección a la que pertenece, una pequeña descripción, el contenido completo de la noticia y un enlace fuente a la noticia. Tras la extracción de la información, la misma se escribe en un archivo .csv localizado en la raíz del proyecto, este es okdiario.csv

Este archivo es empleado posteriormente para llevar a cabo un procesado de las noticias y poder poblar la base de datos, así como el índice para el buscador.

3. Procesado

La información condensada en el archivo .csv que se genera del WebScraping es a su vez procesada para poblar la base de datos de la aplicación web de Django y para poblar el índice de Whoosh usado en el buscador del que dispone la aplicación.

A través de los archivos populate y populateIndex.py, ambas operaciones se llevan a cabo a través de sus respectivos métodos: populateDatabase y populateIndexer.

4. Whoosh

El presente proyecto cuenta con un buscador de contenidos implementado en Whoosh. Se basa en un Schema con 7 campos, uno númerico, el pk, y los 6 restantes de tipo TEXT, que son: sección, título, autor, descripción, link, y contenido.

Desde el archivo populateIndex.py, y mediante el método populateIndexer, se prepara el buscador de contenidos para que esté operativo y que pueda ser accesible por el usuario. El proceso consiste en crear un índice en /main, en caso de que no exista y sino repoblarlo mediante el uso de un writer que escriba en el índice todos los contenidos de la noticia.

Para la búsqueda se ha empleado un multiparser que busque los términos de la búsqueda en todos y cada uno de los 6 campos principales, y OrGroup para que devuelva como Hit cualquier noticia para la que encuentra al menos 1 de los términos de la cadena completa introducida en el buscador. El usuario accede al buscador e introduce los términos de la búsqueda, es entonces cuando se abre un buscador y se lleva a cabo el proceso de búsqueda.

5. Django

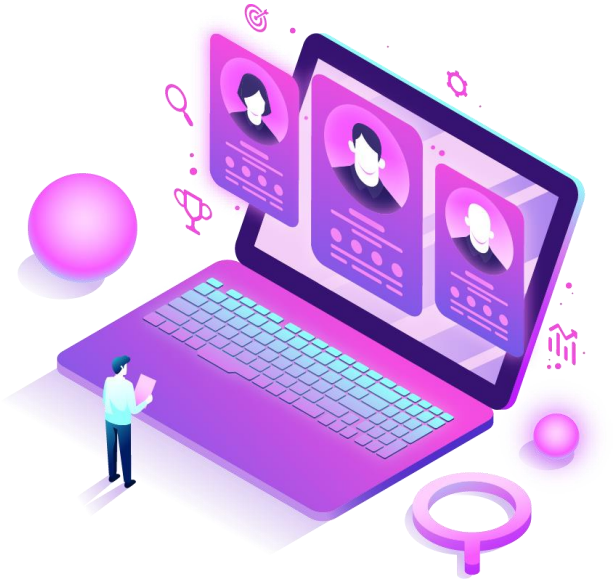
La aplicación web desplegada está desarrollada en este framework. El modelo relacional empleado en la misma consta de las siguientes tablas:

- Sección
- Autor
- User
- Usuario
- DictSection (Ver en Sistema de Recomendación)
- DictAutor (Ver en Sistema de Recomendación)
- Noticia

Para el aspecto visual de la página web, se ha implementado BootStrap en los archivos .html de representación. La aplicación web cuenta con un formulario de autentificación y un formulario de registro.

Existen 2 vistas, la del administrador y la del usuario normal. Es el administrador quien tiene acceso a los métodos de población de la base de datos y el índice de Whoosh, mientras que desde la vista del usuario normal sólo se puede acceder al buscador de Whoosh y al sistema de recomendación de noticias.

6. Sistema de recomendación basado en contenido



Para definir el perfil del usuario se han tomado como características la sección y el autor que más suele leer el usuario. Para saber en todo momento qué es aquello que más lee el usuario, hemos empleado la frecuencia/número de apariciones.

Se han empleado 2 modelos que funcionan a modo de diccionario, uno de ellos para llevar la cuenta de las secciones que más visita el usuario, y otro para llevar la cuenta de los autores que más visita el usuario.

Cada uno de ellos cuenta con 3 campos: user, key y value. El user referencia al usuario al que pertenece, key es la sección o autor y por último value es un valor numérico que se incrementa en una unidad cada vez que se visita una noticia.

Hemos supuesto que la sección de la noticia tiene más peso que el autor al que más suele leer el usuario. Se le ha dado un 70% de peso a la sección y un 30% al autor.

Cuando el sistema de recomendación va a recomendar una noticia a un usuario, se extraen las preferencias del usuario a partir de los 2 modelos anteriormente descritos. A continuación, se itera sobre el conjunto total de noticias y para cada una de ellas se hace:

- Se extrae la sección y se calcula la frecuencia/número de apariciones de noticias con esa sección a las que ha accedido el usuario entre la suma de todas las frecuencias de secciones. El cálculo anterior se multiplica por w = 70/100

- Se extrae el autor y se calcula la frecuencia/número de apariciones de noticias con ese autor a las que ha accedido el usuario entre la suma de todas las frecuencias de autores. El cálculo anterior se multiplica por w = 30/100

- Se suman ambas operaciones, y el resultado es el peso final que tiene la noticia para recomendársela al usuario.

Una vez tenemos una lista con cada noticia y su peso correspondiente, se reordena de manera inversa para disponer al principio de aquellas con más peso, y se toman las “n” primeras para ser devueltas como las recomendadas por el sistema.

7. Conclusiones

El framework Django facilita el diseño de aplicaciones web y acerca la programación web a aquellos programadores no tan orientados a la misma, gracias al empleo de un lenguaje sencillo como es Python.

En lo que respecta a la asignatura y los contenidos de la misma, es interesante su enfoque totalmente práctico y la cantidad de nuevos recursos actuales que se aprenden de la misma, y que pueden ser aplicados desde ya en el mundo laboral.