FUNDAMENTOS DE PROGRAMACIÓN

Ejercicio de Laboratorio: Stack Overflow

Autor: José A. Troyano. **Revisor:** Fermín Cruz. **Última modificación:** 1/12/2020

En este proyecto trabajaremos con preguntas sobre el lenguaje Python. Los datos están extraídos de Stack Overflow (www.stackoverflow.com), y se corresponden con una colección de preguntas desde 2008 hasta 2016 relacionadas con Python. Los datos con los que trabajaremos incluyen distintas informaciones sobre las preguntas. A partir de ellos, generaremos una serie de informes y gráficas que resumirán aspectos relevantes de las temáticas más consultadas.

El formato de entrada es CSV. Cada registro del fichero de entrada ocupa una línea y contiene cuatro informaciones sobre las preguntas (puntuación, título, año y etiqueta principal). Estas son las primeras líneas del fichero:

	A	B	C	D
1	score	title	year	tag
2	21	How can I find the full path to a font from its display name on a Mac?	2008	photoshop
3	27	Get a preview JPEG of a PDF on Windows?	2008	pdf
4	40	Continuous Integration System for a Python Codebase	2008	extreme-programming
5	25	cx_Oracle: How do I iterate over a result set?	2008	cx-oracle
6	28	Using 'in' to match an attribute of Python objects in an array	2008	iteration
7	30	Class views in Django	2008	oop
8	20	Python and MySQL	2008	bpgsql
9	256	How do I use Python's itertools.groupby()?	2008	iteration
10	364	Adding a Method to an Existing Object Instance	2008	monkeypatching
11	251	How do you express binary literals in Python?	2008	literals

Figura 1: fichero de datos

Además de distintos indicadores, generaremos un par de salidas gráficas que mostrarán, respectivamente, la evolución (Figura 2) y la distribución (Figura 3) de uso de ciertas etiquetas.

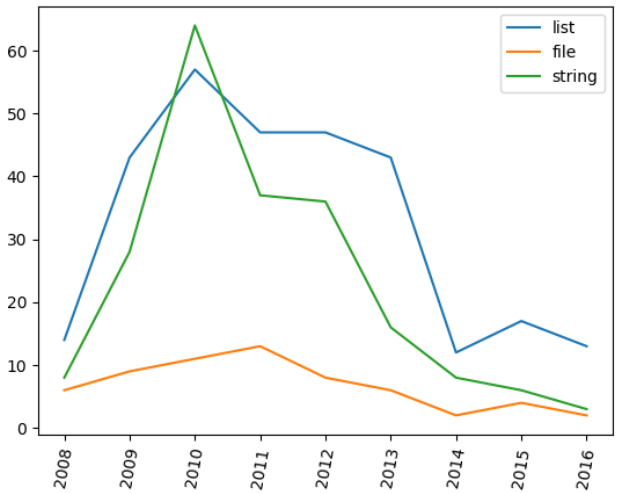

Figura 2: Evolución de etiquetas

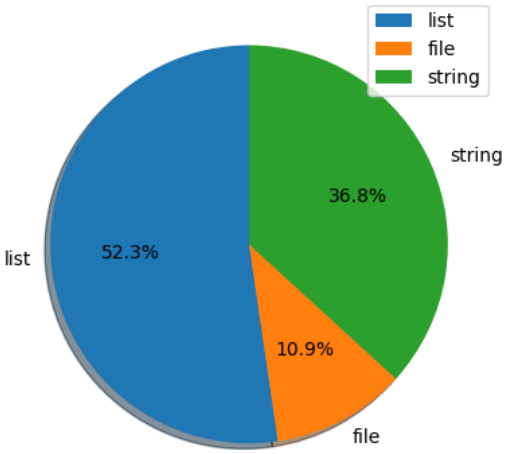

Figura 3: Distribución de uso de etiquetas

Para almacenar en Python la información de cada una de las líneas se usará la siguiente definición de namedtuple:

```
Pregunta = namedtuple('Pregunta', 'puntuacion, titulo, año, etiqueta')
```

Cree un fichero **stackoverflow.py** e incluya en él la definición del namedtuple anterior (recuerde que debe importar namedtuple del módulo collections para poder utilizarlo). A continuación, implemente las funciones que se le piden.

1. **leer_preguntas**: recibe la ruta de un fichero CSV codificado en UTF-8, y devuelve una lista de tuplas de tipo Pregunta conteniendo todos los datos almacenados en el fichero.
2. **filtrar_por_año**: recibe una lista de tuplas de tipo Pregunta y un año de tipo int, y devuelve una lista con aquellas tuplas de tipo Pregunta cuyo año coincida con el valor recibido por parámetro.
3. **calcular_etiquetas**: recibe una lista de tuplas de tipo Pregunta, y devuelve un conjunto {str} con etiquetas que aparecen en las preguntas recibidas por parámetro.
4. **calcular_preguntas_mejor_valoradas**: recibe una lista de tuplas de tipo Pregunta y un parámetro *n* de tipo int, y devuelve una lista de *n* tuplas (str, int) con los títulos y puntuaciones de las preguntas con las puntuaciones más altas. La lista devuelta debe estar ordenada de mayor a menor puntuación. Si hubiera menos de *n* preguntas en la lista de entrada, devolverá tantas tuplas como preguntas haya.
5. **contar_etiquetas**: recibe una lista de tuplas de tipo Pregunta, y devuelve un diccionario {str: int} en el que las claves son las etiquetas de las preguntas y los valores indican cuántas preguntas hay para cada etiqueta.
6. **mostrar_distribucion_etiquetas:** recibe una lista de tuplas de tipo Pregunta y un parámetro *etiquetas* de tipo lista [str], y muestra un diagrama de tarta con la distribución de uso de las etiquetas (Figura 2). Se usarán las siguientes instrucciones para dibujar el diagrama:

   ```
   plt.pie(tamaños, labels=etiquetas, autopct='%1.1f%%', shadow=True, startangle=90)
   plt.legend()
   plt.show()
   ```

   Donde *tamaños* es una lista con el número de preguntas para cada etiqueta del parámetro *etiquetas*.
7. **calcular_palabras_clave**: recibe el título de una pregunta y un parámetro *stopwords* de tipo lista [str], y devuelve una lista [str] con las palabras clave obtenidas del título. El parámetro *stopwords* tendrá como valor por defecto una lista vacía. El procedimiento para obtener las palabras claves es el siguiente:
   a) Convertir el título a minúsculas.
   b) Descomponer el título en una lista de términos separados por espacios.
   c) Eliminar los siguientes símbolos de los términos: '¿?[](){}¡!-+/*,;.<>='.
   d) Dejar en la lista de términos solo aquellos que estén compuestos por letras.
   e) Eliminar de la lista los términos que aparezcan en la lista de *stopwords*.
8. **contar_palabras_clave:** recibe una lista de tuplas de tipo Pregunta y un parámetro *stopwords* de tipo lista [str], y devuelve una lista de tuplas de tipo (str, int) con las palabras clave que aparecen en los títulos de las preguntas y el número de ocurrencias de cada una. La lista devuelta estará ordenada de mayor a menor frecuencia de aparición de las palabras clave. Utilice la función anterior para obtener las palabras clave, excluyendo las *stopwords*.

9. **agrupar_preguntas_por_año:** recibe una lista de tuplas de tipo Pregunta, y devuelve un diccionario {int: [Pregunta]} cuyas claves son los años y los valores son listas con las preguntas de cada año.
10. **mostrar_evolucion_etiquetas**: recibe una lista de tuplas de tipo Pregunta y un parámetro *etiquetas* de tipo lista [str], y muestra una gráfica con la evolución del uso de las *etiquetas* a lo largo de los años (Figura 3). Se usarán las siguientes instrucciones para dibujar la gráfica:

```
for etiqueta, evolucion in zip(etiquetas, evoluciones):
    plt.plot(evolucion, label=etiqueta)
plt.xticks(range(len(años)), años, rotation=80, fontsize=10)
plt.legend()
plt.show()
```

Donde *años* y *evoluciones* son dos listas con la siguiente información:

- *años*: lista de los años incluidos en la colección de preguntas, ordenados de menor a mayor.
- *evoluciones*: lista con la evolución de uso de cada etiqueta, alineada con la lista de *etiquetas*. Cada evolución consiste en una lista de frecuencias, alineada con la lista de años, correspondientes con el número de veces que la etiqueta ha sido usada cada año.

Cree un fichero **stackoverflow_TEST.py**. Importe todas las funciones del módulo stackoverflow. Cargue los datos del fichero CSV y muestre en consola los datos leídos. Incluya llamadas a todas las funciones implementadas, mostrando los resultados en la consola.