T2.2: Scraping. Descarga de titulares de varios medios

Queremos descargar os titulares de noticias de uns cantos medios.

Crea un arquivo chamado: **medios.json** e mete nel os medios, URL e como filtrar os titulares. Tes un exemplo de este arquivo no Anexo II. Engade cinco medios máis.

O código debe descargar os titulares dos medios do arquivo **medios.json**. Debe calcular os pesos en función da popularidade. Indícase un método.

Imos crear un array que teña: Titular, URL completa, peso do titular.

Titulares

Hay que descartar as palabras de tres ou menos caracteres e limpar os signos de puntuación para o conteo (non para amosar o titular).

Cálculo de pesos

Peso de cada palabra: Número de veces que aparece cada palabra en tódolos titulares.
Peso individual de cada titular: O peso de cada palabra dividido entre o número de palabras.
Peso máximo: O peso do titular máis alto (maior) dos titulares.
Peso relativo entre titulares: Peso do titular / Peso máximo.

Exemplo:

Titular	Peso individual	Peso relativo	Tamaño fonte HTML
Este é o titular 1	(3+3+3+3+1)/5=2,6	2,6/2,6=1	1+1 = 2em
Este é o titular 2	(3+3+3+3+1)/5=2,6	2,6/2,6=1	1+1 = 2em
Este é o titular 3	(3+3+3+3+1)/5=2,6	2,6/2,6=1	1+1 = 2em
Unha casa	(1+1)/2=1	1 / 2,6= 0,38	1+0,38=1,38em

Peso Máximo = 2,6

Entrega

Pídese o código nun notebook que faga:

- Amosar un gráfico de barras coas 10 palabras que se repiten máis.
- Escribir a un arquivo HTML: resultado-ano-mes-dia.html (ordeado polo peso, de modo que os titulares que aparecen máis grandes, aparezan ao inicio).
- Amosar por pantalla os 20 titulares e a URL que teñan maior peso relativo (“relevancia”).
- Suxire un método mellor para analizar a relevancia dos titulares.
- Mete os titulares, medio, URL e peso relativo nunha BBDD de SQLite.

Exemplo:

```
<p style="font-size: 2em;"><a href=”http://url-ao-titular”>Este es el titular 1</a></p>
<p style="font-size: 2em;"><a href=”http://url-ao-titular”>Este es el titular 2</a></p>
<p style="font-size: 2em;"><a href=”http://url-ao-titular”>Este es el titular 3</a></p>
<p style="font-size: 1,38em;"><a href=”http://url-ao-titular”>Unha casa</a></p>
```

Anexo I: Código de exemplo

```
import requests
import json
from collections import deque
from bs4 import BeautifulSoup

def dameTitulares(URL, auxTag, auxClase):
  print(f"Filtrando resultados do medio: {URL}")
  _parser = BeautifulSoup(requests.get(URL).content, "html.parser")
  _titulares = _parser.find_all(auxTag, class_=auxClase)
  pilaTitulares = deque()
  for _auxTit in _titulares:
    pilaTitulares.append([_auxTit.text.strip(), _auxTit.find("a").get("href")])
  return (pilaTitulares)
#Cambiar pola lectura do arquivo
URLS = {"La Voz de Galicia": {"url": "https://www.lavozdegalicia.es", "tag": "h4",
"clase": "a-min-headline"},
  "El País": {"url": "https://www.elpais.com", "tag": "h2", "clase": "c_t"},
  "El Mundo": {"url": "https://www.elmundo.es", "tag": "h2", "clase": "ue-c-cover-
content-link"},
  "20 Minutos": {"url": "https://www.20minutos.es", "tag": "h1", "clase": ""},
  "El Espanol": {"url": "https://www.elespanol.com/quincemil", "tag": "div",
"clase": "titulo"}}

allTitulares = deque()
enderezos=json.loads(json.dumps(URLS))
for auxURL in enderezos:
  allTitulares.append(dameTitulares(enderezos[auxURL]['url'], enderezos[auxURL]
['tag'], enderezos[auxURL]['clase']))

for auxTitular in allTitulares:
  print(auxTitular)
```

Anexo II: Medios (medios.json)


```
{
  "La Voz de Galicia": {
    "url": "https://www.lavozdegalicia.es",
    "tag": "h4",
    "clase": "a-min-headline"
  },
  "El País": {
    "url": "https://www.elpais.com",
    "tag": "h2",
    "clase": "c_t"
  },
  "El Mundo": {
    "url": "https://www.elmundo.es",
    "tag": "h2",
    "clase": "ue-c-cover-content-link"
  },
  "20 Minutos": {
    "url": "https://www.20minutos.es",
    "tag": "h1",
    "clase": ""
  },
  "El Espanol": {
    "url": "https://www.elespanol.com/quincemil",
    "tag": "div",
    "clase": "titulo"
  }
}
```