SPYDER

Rafael Fernández Ortiz

Se denomina ***Spyder*** o ***Web Spyder*** al bot/script que se utiliza para analizar o rasrear el código html de una dirección url. De este modo, puedes identificar las palabras y por tanto hacer un análisis de frecuencia de ciertas palabras, o identificar los distintos hipervínculos que existen en dicha url.

Este es un primer proyecto de clase, así repaso el concepto y la utilidad de las clases, de un objeto *spyder*.

Las distintas **funciones** o **métodos** del *spyder* permiten, o bien extraer los links de direcciones *url* y *pdfs*, y mostrarlo en pantalla mediante una lista; o bien, escribirlo en un archivo *.txt*

Comentarios: Este es una segunda versión del código, he añadido el método para identificar los enlaces a las distintas imagenes que hay en la url. Aún hay que afinarlo muchísimo.

```
import urllib.request as url
class Spyder(object):
    def __init__(self,url):
        self.url = str(url)
    def __getHtml(self):
        dir = url.urlopen(self.url)
        html = str(dir.read())
        return html
    def __getIMG(self):
        img = []
        html = self.__getHtml()
        while True:
            pos = html.find('src="')
            if pos == -1:
                break
            html = html[(pos+5):]
            img.append(html[:html.find('"')].lower())
        return img
    def __getUrls(self):
        urls = []
        html = self.__getHtml()
        while True:
            pos = html.find('href="')
            if pos == -1:
                break
            html = html[(pos + 6):]
            urls.append(html[:html.find('"')].lower())
        return urls
    def __buscaEnUrls(self,cosa):
        return [x for x in self.__getUrls() if str(cosa) in x]
    def __buscaEnImg(self,cosa):
        return [x for x in self.__getIMG() if str(cosa) in x]

    def getLink(self):
```

```
        return self.__buscaEnUrls("http")
    def getPdf(self):
        return self.__buscaEnUrls(".pdf")
    def getImg(self):
        img = []
        extensiones = [".jpg", ".gif", ".png"]
        img += self.__buscaEnImg(".jpg")
        img += self.__buscaEnImg(".gif")
        img += self.__buscaEnImg(".png")
        return img
    def getLinktoTxt(self):
        fich = open("links.txt",'w')
        link = self.getLink()
        for i in link:
            fich.write(str(i)+"\n")
        fich.close()
        return fich
    def getPdftoTxt(self):
        fich = open("pdfs.txt",'w')
        link = self.getPdf()
        for i in link:
            fich.write(str(i)+"\n")
        fich.close()
        return fich

# Ejemplo
# sp1 = Spyder("https://es.wikipedia.org/wiki/Vincent_van_Gogh")
# print(sp1.getImg())
# sp1.getLinktoTxt()
```