
T1.2: Nube de palabras

Sistemas de Big Data

26/11/23 – IES Fernando Wirtz

David Fernández Reboredo

Índice

Tarefa 1.2..... 3

 Exercicio Principal..... 3

 Ampliación 1..... 6

 Ampliación 2..... 7

Tarefa 1.2

Exercicio Principal

Para comezar a realizar a tarefa deberemos realizar un pull sobre o repositorio proporcionado, para elo e tan fácil como facer click na icona da barra lateral esquerda como podemos observar na seguinte imaxe:



Posteriormente a realizar a seguinte acción deberemos ingresar o link que se nos proporcionó, na tarefa.

A partires de ahí xa poderemos traballar co repositorio proporcionado.

Unha vez realizado esto, deberemos a todas as librerías facer un conda install.

Deberemos facelo coas librerías seguintes:

```
- conda install -c conda-forge wordcloud
```

```
- conda install numpy
```

```
- conda install PIL
```

```
- conda install wordclud
```

```
- conda install nltk
```

```
- conda install urllib3
```

```
- conda install bs4
```

Posteriormente a esto simplemente deberemos de correr o código, para unha mellor utilización agrupámoslos en funcións

```

def limpiarcodigo(texto):
    clean_texto=''
    punctuation=[]
    for s in string.punctuation:
        punctuation.append(str(s))
    sp_punctuation = ["¿", "¡", "“", "”", "...", ":", "]", "»", "«", "->", '(', ')']

    punctuation += sp_punctuation

    punctuation[:10]
    stop_words = stopwords.words('spanish')

    stop_words += ["\u200b", "\xa0", 'de', 'y', 'es', 'se', 'en', 'con', 'para', 'como', 'puede', 'cómo', 'hacer', 'forma', 'parte', 'hace', 'además', ""]

    stop_words[:10]

    for sw in stop_words:
        clean_texto = texto.lower().replace(sw, "")

    for p in punctuation:
        clean_texto = clean_texto.replace(p, "")

    for stop in stop_words:
        clean_texto_list = clean_texto.split()
        clean_texto_list = [i.strip() for i in clean_texto_list]
        try:
            while stop in clean_texto_list: clean_texto_list.remove(stop)

```

O cal empregaremos para facer todo o filtrado de palabras

Por outro lado o xerador da nube de palabras

```

def generarNube(clean_texto):
    word_cloud = WordCloud(height=800, width=800, background_color='white', max_words=150, min_font_size=5).generate(clean_texto)
    plt.figure(figsize=(10,8))
    plt.imshow(word_cloud)
    plt.axis('off')
    plt.tight_layout(pad=0)
    plt.show()

```

Usamos o noso ficheiro txt prefieto e realizaremos o seu parseo:

```

def parsear_txt():
    texto=''
    with open('defino.txt', 'r') as fichero:
        for linea in fichero:
            texto=texto+linea
    return texto+''
    fichero.close()

```

✓ 0.0s

```

texto=parsear_txt()
● texto=limparcodigo(texto)
print(texto)
generarNube(texto)

```

Finalmente como podemos observar na parte inferior correremos o código xerandonos o seguinte resultado:

relacional obtener convencional referencia
almacenar selección técnicas dependientes procesa
definir big data volumen volumen estadísticas permite
uso manera especialmente inteligente aplicaciones procesamiento combinación variedad
limpia análisis analiza tamaño aplicaciones conjuntos vista procesar
manipulación red bigdata principales análisis tratamiento
grande funcionalidades principal analizan difíciles escala transmisión
conclusiones situaciones propósito tecnologías
diferente manipular tecnologías tratamiento
demasiado cantidad fuentes entiende gestión almacenamiento procesamiento
analise aplicaciones finalidad
entendemos tratarlos obtiene cantidad
complejos masivos volúmenes estructura
gestionado términos tipo formación datos
dependiendo tipo formación datos
trabajar hardware tipo formación datos
consulta software grandes cantidades
precisan elevados informáticas especializado nuevas tradicionales tramitar

Ampliación 1

Para a ampliación deberemos de facelo mediante unha URL

O seguinte código proporcionaranos a obtención do texto na URL ao obter nel os h4 da html:

```
def def_urls(urlprop):
    ua = "Mozilla/5.0 (Linux; U; Android 2.2; en-us; Nexus One Build/FRF91) AppleWebKit/533.1 (KHTML, like Gecko) Version/4.0 Mobile Safari/533.1"
    h = {"User-Agent": ua}
    resultados=''
    http = urllib3.PoolManager()
    medioDigital= urlprop
    r = http.request('GET', medioDigital, fields=None, headers=h)
    sopa = BeautifulSoup(r.data, "html.parser")
    web_solotexto = sopa.get_text()

    salida = ''

    for linea in web_solotexto.split('\n'):
        aux=linea.strip()
        if aux and len(aux) > 50:
            salida += aux + '\n'

    print (salida)

    #Otra forma, quizás más elegante:
    titulares = sopa.find_all('h4')

    for titular in titulares:
        resultados=titular.get_text().strip()
    return resultados + salida +''
```

```
url="https://www.xataka.com/energia/reactor-jet-ha-completado-exito-sus-pruebas-finales-deuterio-tritio-hito-crucial-para-fusion-nuclear"
url=def_urls(url)
```

✓ 1.0s

El reactor JET ha completado con éxito sus pruebas finales con deuterio y tritio. Es un hito crucial para la fusión nuclear

El reactor JET ha completado con éxito sus pruebas finales con deuterio y tritio. Es un hito crucial para la fusión nuclear

Los técnicos de ITER iniciarán las pruebas de alta potencia con deuterio y tritio en 2035

Las pruebas con plasma ionizado llevadas a cabo en JET son cruciales para que ITER y DEMO salgan bien

Si todo sigue su curso como está previsto ITER (International Thermonuclear Experimental Reactor), el reactor experimental de fusión nuclear de

No obstante, antes de iniciar las pruebas de alta potencia con el combustible final ITER deberá superar otros tests que también son cruciales.

El reactor experimental JT-60SA reside en Naka, una pequeña ciudad no muy alejada de Tokio (Japón). Su propósito muy a grandes rasgos es llevar

JET ha probado la fusión con el mismo combustible que utilizará ITER

El último gran hito de JET vio la luz pública el 9 de febrero de 2022. Ese día los científicos que lo operan anunciaron que habían logrado bat

La mejor arma de EEUU frente a China es una máquina europea: el equipo de litografía 'EUV-High NA' de ASML

Los técnicos de JET han completado el programa DTE3, que es la tercera y última campaña de pruebas con plasma ionizado que contiene núcleos de

Es que hace apenas unas horas los responsables de este reactor experimental de fusión nuclear han anunciado que sus científicos han completa

Hay muchas razones por las que esta campaña de pruebas con el mismo combustible que usará ITER tiene muchísima importancia. Una de las más re

Además han permitido refinar la administración del tritio, que es un isótopo radiactivo, y también han ayudado a los técnicos a comprender con

En Xataka: Estos son los plazos que maneja ITER actualmente para demostrar la viabilidad de la fusión nuclear

Streaming Análisis Energía Espacio Móviles Xataka Movilidad Apple Samsung Inteligencia artificial China Empleo Windows 11 Ver más temas

XatakaXataka MóvilXataka AndroidXataka Smart HomeApplesferaGenbetaMundo Xiaomi

tecotec Conga 12090, análisis: el mejor fregado que he probado en un robot aspirador merecía más

Si tienes fotos viejas o borrosas, esta nueva herramienta IA de expertos españoles hace un upscaling milagroso (y hasta creativo)

O resultado de este mostraranos unha serie de palabras separadas por un espacio

Posteriormente poderemos realizar as operacións de limpeza e creación da nube de palabras.

```
url=limparcodigo(url)
print(url)
```

```
generarNube(url)
```

Nos dará como resultado la siguiente nube de palabras

