Python

Procesamiento de Lenguaje Natural

Rogelio Ferreira Escutia

Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

xumarhu.net

“Tokenizar” un texto

NLTK

- **Instalación de “NLTK” y “stopwords” (en consola):**

```
(base) MacBook-Pro-de-Rogelio-2:~ rogelioferreiraescutia$ python3
Python 3.8.5 (default, Sep  4 2020, 02:22:02)
[Clang 10.0.0 ] :: Anaconda, Inc. on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> nltk.download('stopwords')
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/rogelioferreiraescutia/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
True
>>>
(base) MacBook-Pro-de-Rogelio-2:~ rogelioferreiraescutia$
```

Tokenizar

- **Abrir un editor de texto y empezar a crear el código, lo primero es cargar las bibliotecas que ocuparemos:**

```
import re                                   # Biblioteca para manejo de expresiones regulares
from nltk.tokenize import word_tokenize     # importamos de NLTK para dividir el texto en tokens
from nltk.corpus import stopwords           # importamos de NLTK las "stopwords"
```

Tokenizar

- **Se selecciona un texto a analizar, en este caso seleccioné una frase corta de Gabriela Mistral:**

```
frase_gabriela_mistral = "Como soy reina y fui mendiga, ahora vivo en puro temblor de que me dejes, y te pregunto, pálida, a cada hora: -¿Estás conmigo aún? ¡Ay, no te alejes!- Quisiera hacer las marchas sonriendo y confiando ahora que has venido; pero hasta en el dormir estoy temiendo y pregunto entre sueños: -¿No te has ido?-"
```

Tokenizar

- **Ahora eliminamos los signos de puntuación que por el momento no ocupamos, así como cualquier otro caracter no deseado que se encuentre en el texto:**

```
# Eliminar puntuaciones del texto
puntuacion = '()[],."“”¿?:;_-¡!'                    # Agregar los que falten
texto_sin_puntuacion = ""
for caracter in texto:
    if(caracter not in puntuacion):
        texto_sin_puntuacion = texto_sin_puntuacion + caracter
print("\nTexto Completo Sin Puntuacion:\n")
print(texto_sin_puntuacion)
```

```
Texto Completo Sin Puntuacion:

Como soy reina y fui mendiga ahora vivo en puro temblor de que me dejes y te pregunto pálida a cada hora Estás
 conmigo aún Ay no te alejes Quisiera hacer las marchas sonriendo y confiando ahora que has venido pero hasta
en el dormir estoy temiendo y pregunto entre sueños No te has ido
```

Tokenizar

- **Eliminamos espacios en blanco utilizando expresiones regulares:**

```
# Eliminar espacios
texto_sin_puntuacion_ni_espacios = re.sub(r'[^\w\s]','', texto_sin_puntuacion)
print("\nTexto Completo Sin Puntuacion ni Espacios:\n")
print(texto_sin_puntuacion_ni_espacios)
```

Tokenizar

- **Convertimos a tokens todo el texto:**

```
# Convertimos a tokens todo mi texto y lo imprimimos en pantalla
tokens_de_mi_texto = word_tokenize(texto_sin_puntuacion_ni_espacios)
print('\nImpresión de todos los tokens de mi texto:\n')
print (tokens_de_mi_texto)
print('\n Tokens Totales:',len(tokens_de_mi_texto))
```

```
Impresión de todos los tokens de mi texto:

['Como', 'soy', 'reina', 'y', 'fui', 'mendiga', 'ahora', 'vivo', 'en', 'puro', 'temblor', 'de', 'que', 'me', '
dejes', 'y', 'te', 'pregunto', 'pálida', 'a', 'cada', 'hora', 'Estás', 'conmigo', 'aún', 'Ay', 'no', 'te', 'al
ejes', 'Quisiera', 'hacer', 'las', 'marchas', 'sonriendo', 'y', 'confiando', 'ahora', 'que', 'has', 'venido',
'pero', 'hasta', 'en', 'el', 'dormir', 'estoy', 'temiendo', 'y', 'pregunto', 'entre', 'sueños', 'No', 'te', 'h
as', 'ido']

 Tokens Totales: 55
```

Tokenizar

- **Cargamos las “stopwords” (palabras sin relevancia para nuestro análisis) del lenguaje español:**

```
# Cargamos las "stopwords" del español (las palabras que no nos aportan información)
palabras_no_relevantes = set(stopwords.words('spanish'))
```

Tokenizar

- **Por último, eliminamos las “stopwords” de nuestro texto y nos queda finalmente nuestra lista final de palabras relevantes en nuestro texto (queda “tokenizado”):**

```
# Filtramos los tokens eliminando las "stopwords"
lista_final = []
for palabra in tokens_de_mi_texto:
    if palabra not in palabras_no_relevantes:
        lista_final.append(palabra)
print('\nLista Final eliminando las palabras no relevantes:\n')
print(lista_final)
print('Total:',len(lista_final))
```

```
Lista Final eliminando las palabras no relevantes:

['Como', 'reina', 'mendiga', 'ahora', 'vivo', 'puro', 'temblor', 'dejes', 'pregunto', 'pálida', 'cada', 'hora'
, 'Estás', 'conmigo', 'aún', 'Ay', 'alejes', 'Quisiera', 'hacer', 'marchas', 'sonriendo', 'confiando', 'ahora'
, 'venido', 'dormir', 'temiendo', 'pregunto', 'sueños', 'No', 'ido']
Total:  30
```

Rogelio Ferreira Escutia

Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

rogelio.fe@morelia.tecnm.mx

rogeplus@gmail.com

xumarhu.net

@rogeplus

https://www.youtube.com/channel/UC0on88n3LwTKxJb8T09sGjg

rogelioferreiraescutia