Python

Procesamiento de Lenguaje Natural

Palabras Importantes

Rogelio Ferreira Escutia

Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

Extracción de las palabras mas importantes

Palabras importantes

- **El primer paso para tratar de encontrar el tema (ó los temas principales) de los que trata un texto es encontrar las "palabras importantes" que nos indiquen su contenido.**

- **La forma mas sencilla de encontrar esto es buscando las palabras que mas se repiten en un mismo texto (quitando las "stopwords"), esto nos da la idea del contenido del mismo (aunque no siempre puede ser así, la misma ambigüedad de las palabras y del texto nos pueden llevar a conclusiones erróneas, pero es un primer paso).**

Extraer Palabras (1)

- **Bibliotecas a utilizar y cargar archivo a procesar:**

```
# Bibliotecas a utilizar
import re                                  # Manejo de expresiones regulares
import nltk                                # Para procesar lenguaje natural
from nltk.tokenize import word_tokenize # Para Tokenizar un texto
from nltk.corpus import stopwords          # Cargar las "Stopwords" del español
from collections import Counter            # importamos para poder contar las palabras repetidas
from collections import OrderedDict        # importamos para ordenar el conteo de palabras repetidas
```

```
# Cargar archivo de texto a memoria
with open('discurso_steve_jobs_stanford_12_junio_2005.txt','r') as archivo_en_memoria:
    texto = archivo_en_memoria.read()
```

Extraer Palabras (2)

- **Limpiar texto y Tokenizar:**

```
# Eliminar simbolos y caracteres especiales usando expresiones regulares
texto_sin_simbolos = re.sub(r'[^\w\s]', '', texto)

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

Extraer Palabras (3)

- **Cargar “StopWords” del español y hacemos el filtrado:**

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

# Filtramos los tokens eliminando las "stopwords"
lista_final = []
for palabra in tokens_de_mi_texto:
    if palabra not in palabras_vacias:
        lista_final.append(palabra)
```

Extraer Palabras (4)

- **Imprimir texto “tokenizado”:**

```
# Impresión final de las palabras relevantes del texto (Tokenizar)
print('\nLista Final eliminando las palabras vacías (no relevantes):\n\n', lista_final)
print("\nTotal de Tokens sin Stopwords: ", len(lista_final))
```

- **Contar y ordenar las palabras repetidas:**

```
# Contar y ordenar las palabras repetidas
contador = Counter(lista_final)
print('\nLista de palabras y cuántas veces se repiten (en orden descendente por cantidad de repeticiones)
:\n', contador)
print('Total:',len(contador))
```

Extraer Palabras (5)

- **Contar la palabras repetidas:**

```
# Contar las palabras repetidas
contador_ordenado = OrderedDict(contador)
print('\nLista de palabras y cuántas veces se repiten (conforme van apareciendo en el texto):\n',
contador_ordenado)
print('Total:',len(contador))
```

Extraer Palabras (6)

- **Definir texto a analizar:**

```
# Extraemos sólo las 5 palabras más repetidas de todo el texto
las_mas_repetidas = OrderedDict(contador.most_common(5))
print('\nImpresión de las 5 palabras mas repetidas de todo el texto:\n', las_mas_repetidas)
print('\nTotal:', len(las_mas_repetidas), "\n")
```

- **Salida:**

```
Impresión de las 5 palabras mas repetidas de todo el texto:
 OrderedDict([('vida', 14), ('Y', 14), ('años', 13), ('No', 10), ('Apple', 9)])

Total: 5
```

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