Vocabulary Challenge

Carlos Oswaldo Rodriguez Salas

Introducción

La generación de un vocabulario en donde se incluya la frecuencia con la que aparece cada palabra en un texto es un paso fundamental en el preprocesamiento de lenguaje natural para su posterior codificación y uso en modelos de Machine Learning e Inteligencia Artificial. En este reporte presentó de manera breve como realizar este proceso en Python. Elegí este lenguaje ya que el texto que procesaremos en esta ocasión es relativamente corto por lo que no necesitamos mucha eficiencia, además, Python cuenta con las herramientas necesarias para realizar este proceso de manera sencilla y facil de entender.

1. Estandarización del texto

Para este ejercicio utilizaremos la primera parte de Los Miserables de Victor Hugo. El texto se descargo desde https://aprende.org/pruebat?sectionId=6 en formato PDF. Después, se selecciono todo el contenido del PDF y se pegó en un archivo de texto, el cuál puede ser leido facilmente en Python.

El codigo para leer el archivo es el siguiente:

```
path='../data/Los_Miserables.txt'

with open(path, 'r') as file:
    file_content = file.read()
```

De manera que el texto termina contenido en la variable *file_content* como una string.

Para la estandarización del texto se realizaron las siguientes operaciones:

```
#Se convierte todo el texto a minusculas
texto=file_content.lower()

#Se eliminan los números
texto=re.sub(r'\d+','',texto)

#Se eliminan los signos de puntuación
texto=re.sub(r'[^\w\s]','',texto)

#Se eliminan los espacios en blanco al inicio y al final del texto
texto=texto.strip()

#Se eliminan los saltos de línea
texto=texto.replace('\n',' ').replace('\r',' ')

#Se eliminan los espacios en blanco múltiples
texto=re.sub(' +',' ',texto)

#Se eliminan los acentos
def standardize(s):
    replacements = (
        ("á", "a"),
        ("é", "e"),
        ("i", "i"),
        ("ó", "o"),
        ("ú", "u"),
    )
    for a, b in replacements:
        s = s.replace(a, b)
    return s

texto=standardize(texto)

#Se dividen las palabras
palabras=texto.split(' ')
```

Esto nos da como resultado una lista con todas la palabras en el texto estandarizadas: en minusculas, sin numeros, sin signos de puntuacion y sin acentos.

2. Estadística

Con esta lista ya podemos hacer un primer acercamiento estadistico:

```
len(palabras)
✓ 0.0s

109222
```

```
len(set(palabras))
✓ 0.0s

13114
```

Tenemos 109,222 palabras en el texto 13,114 palabras unicas.

Para generar el vocabulario se utiliza el siguiente codigo:

```
vocab=dict(zip(set(palabras),[0]*len(set(palabras))))
0.0s
```

```
print(vocab)
0.0s

letra': 0, 'imponer': 0, 'mitad': 0, 'circulos': 0, 'mu
```

```
for word in palabras:
    vocab[word]+=1
```

Y para ordenar las palabras por mayor y menor frecuencia se generan los siguientes diccionarios:

```
most_frequent_vocab={k: v for k, v in sorted(vocab.items(), key=lambda item: item[1],reverse=True)}
least_frequent_vocab={k: v for k, v in sorted(vocab.items(), key=lambda item: item[1])}
```