Facultad de Ciencias Físico Matemáticas - Universidad Autónoma de Nuevo León

Maestría en Ciencia de Datos

Procesamiento y Clasificación de Datos

Tarea 1: Procesamiento de Datos

José Alberto López Alvarez

Matricula:1553133

jose.lopezalv@uanl.edu.mx

15 de mayo de 2022, Monterrey Nuevo León, México

Introducción

Para esta tarea trabajaremos en el procesamiento de datos de texto, lo cual implica la extracción, limpieza y análisis de un conjunto de palabras, con el fin de conocer las características e interpretación de un determinado texto.

En este ejercicio utilizaremos un texto de la enciclopedia de Wikipedia que habla sobre el holocausto. La mayoría conocemos los antecedentes de la época del holocausto y lo que represento para la historia de la humanidad, sin embargo, elegí trabajar sobre este texto ya que tiene mucha literatura y creo que seria interesante hacer un análisis estadístico de la gráfica cloud words y de frecuencia para tratar de adivinar de que trata el texto suponiendo que no supiéramos su procedencia y contenido.

Procedimiento

Lo primero que tenemos que hacer es descargar las librerías para el procesamiento de datos y que nos permitan importar contenidos desde Wikipedia y después seleccionamos el tema a consultar dentro de esta plataforma:

```
import nltk
import string
import re

!pip install Wikipedia

import wikipedia as wiki

wiki_info = wiki.summary('The_Holocaust')
wiki_info


nltk.download('stopwords')
nltk.download('punkt')
```

```
nltk.download('wordnet')

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

from nltk.stem.porter import PorterStemmer
from nltk.tokenize import word_tokenize
stemmer = PorterStemmer()

from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
lemmatizer = WordNetLemmatizer()
```

Después procedemos a la limpieza y procesamiento de los datos, en esta parte tenemos que seguir los siguientes puntos:

1. Minúsculas: convertiremos todas las letras de texto en minúsculas para estandarizar cada palabra a una misma escala y evitar repeticiones.

```
def text_lowercase(text):
    return text.lower()
Holocaust_lowcase = text_lowercase(wiki_info)
Holocaust_lowcase
```

2. Números: dentro de la limpieza de datos de texto es importante quitar los números para trabajar solo con datos de texto y analizar las palabras.

```
def remove_numbers(text):
    result = re.sub(r'\d+', '', text)
    return result
Holocaust_numbers = remove_numbers(Holocaust_lowcase)
Holocaust_numbers
```

3. Signos de puntuación: otro proceso dentro de la limpieza de datos de texto que es muy importante es la eliminación de los signos de puntuación.

```
def remove_punctuation(text):
    translator = str.maketrans('', '', string.punctuation)
    return text.translate(translator)
Holocaust_punct = remove_punctuation(Holocaust_numbers)
Holocaust_punct
```

4. Stopwords: las stopwords son palabras que comúnmente conocemos como conectores o palabras vacías (ejemplo: is, the, it, them, has, my, she, etc.), en el procesamiento de datos de texto es importante eliminar estas palabras ya que son muy repetitivas, pero tienen poca relevancia y dentro de un análisis estadístico la frecuencia de estas palabras podría sesgar la información de las palabras que si son importantes.

```
def remove_stopwords(text):
    stop_words = set(stopwords.words("english"))
    word_tokens = word_tokenize(text)
    filtered_text = [word for word in word_tokens if word not in stop_wor
ds]
    return filtered_text
Holocaust_stop = remove_stopwords(Holocaust_punct)
Holocaust_stop
```

5. Tokenización y Lematización: otra herramienta importante en el procesamiento de datos de texto es la tokenización y Lematización de las palabras, esto se refiere a homologar palabras que pudieran ser verbos, sustantivos o adjetivos provenientes de una misma palabra y poderlas clasificar como una sola para medir su frecuencia.

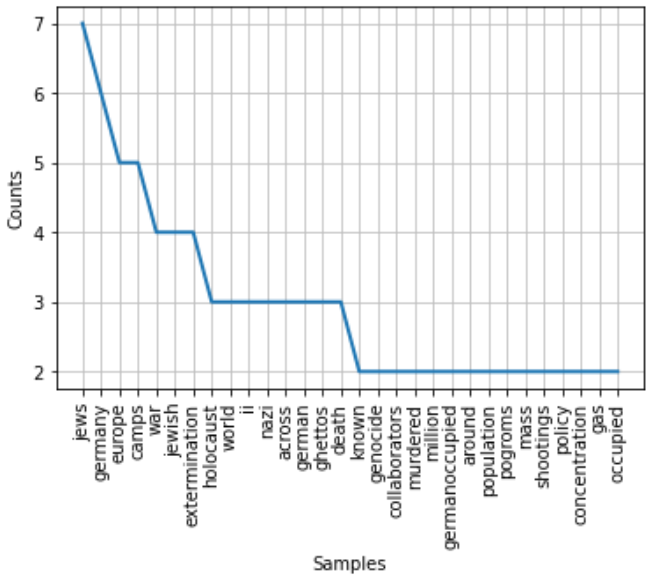
```
def stem_words(text):
    word_tokens = word_tokenize(text)
    stems = [stemmer.stem(word) for word in word_tokens]
    return stems
stem_words(Holocaust_punct)

def lemmatize_word(text):
    word_tokens = word_tokenize(text)
    # provide context i.e. part-of-speech
    lemmas = [lemmatizer.lemmatize(word, pos ='v') for word in word_token
s]
    return lemmas
lemmatize_word(Holocaust_punct)
```

Resultados y Conclusión

Después del procesamiento y limpieza de los datos podemos hacer un análisis estadístico de los datos, como en cualquier otro tipo de análisis estadísticos podemos observar el comportamiento y la interpretación del texto.

Gráfica 1



Gráfica 2

En la ***gráfica 1 y 2*** podemos observar las palabras con mayor frecuencia y relevancia en el texto, a partir de esta información ya podríamos deducir de que trata el texto sin necesidad de que nos hubieran dicho que habla sobre el holocausto ya que vemos palabras como judíos, alemanes, Europa, guerra, nazis, muertes, genocidio, campos, etc.

Referencias

[1] ***https://github.com/AlbertoLopezAlvz/ProcesamientodeDatos/blob/main/README.md***

[2]https://en.wikipedia.org/wiki/The_Holocaust