Nombre: Gabriel Muñoz Marcelo Callisaya
CI: 9873103

DAT 261 - Procesamiento del lenguaje natural

Tarea 3.

1. Considerar el texto de la tarea nro. 2.- Rosa Blanca: Los alemanes que intentaron derrocar a Hitler disponible en: https://www.bbc.com/news/magazine-21521060
2. Normalizar las palabras con NLTK
3. Normalizar las palabras con spaCy
4. Determinar la diversidad léxica con NLTK y SpaCy, usando técnicas de normalización.
5. Comparar los resultados con la tarea nro. 2.

Normalización con NLTK:

Se aplicaron conversión a minúsculas, eliminación de tildes, signos de puntuación, números, espacios redundantes, stop words, stemming y lematización.

```
Tokens (Stemming):
['seventi', 'year', 'ago', 'today', 'three', 'german', 'student',
'execut', 'munich', 'lead', 'resist', 'movement', 'hitler', 'sinc',
'member', 'white', 'rose', 'group', 'becom', 'german', 'nation', 'hero',
'lilo', 'furstramdohr', 'one', 'world', 'war', 'ii', 'height', 'munich',
'centr', 'nazi', 'power', 'group', 'student', 'start', 'campaign',
'passiv', 'resist', 'liselott', 'furstramdohr', 'alreadi', 'widow',
'age', 'follow', 'husband', 'death', 'russian', 'front', 'introduc',
'...']
Tokens (Lematización):
['seventy', 'year', 'ago', 'today', 'three', 'german', 'student',
'executed', 'munich', 'leading', 'resistance', 'movement', 'hitler',
'since', 'member', 'white', 'rose', 'group', 'become', 'german',
'national', 'hero', 'lilo', 'furstramdohr', 'one', 'world', 'war',
'ii', 'height', 'munich', 'centre', 'nazi', 'power', 'group',
'student', 'started', 'campaign', 'passive', 'resistance', 'liselotte',
'furstramdohr', 'already', 'widow', 'age', 'following', 'husband',
'death', 'russian', 'front', 'introduced', '...']
```

Resultados:

```
Tipos (Stemming): 308
Tokens (Stemming): 488
Diversidad léxica (Stemming): 0.6311
Tipos (Lematización): 319
Tokens (Lematización): 488
Diversidad léxica (Lematización): 0.6537
```

Normalización con spaCy (Diversidad léxica):

Se aplicaron las mismas técnicas, usando lematización nativa de spaCy y stemming de NLTK.

```
Tokens (Stemming):
['seventi', 'year', 'ago', 'today', 'three', 'german', 'student',
'execut', 'munich', 'lead', 'resist', 'movement', 'hitler', 'sinc',
'member', 'white', 'rose', 'group', 'becom', 'german', 'nation', 'hero',
'lilo', 'furstramdohr', 'one', 'world', 'war', 'ii', 'height', 'munich',
'centr', 'nazi', 'power', 'group', 'student', 'start', 'campaign',
'passiv', 'resist', 'liselott', 'furstramdohr', 'alreadi', 'widow',
'age', 'follow', 'husband', 'death', 'russian', 'front', 'introduc',
'...']
Tokens (Lematización):
seventy
year
ago
today
three
german
...
campaign
passive
resistance
liselotte
furstramdohr
already
widow
age
follow
husband
death
russian
front
introduce
...
```

Resultados:

```
Tipos (Stemming): 303
Tokens (Stemming): 486
Diversidad léxica (Stemming): 0.6235
Tipos (Lematización): 296
Tokens (Lematización): 486
Diversidad léxica (Lematización): 0.6091
```

Comparación con Tarea 2:

Comparación con Tarea 2:

En la Tarea 2, los resultados fueron:

```
NLTK: Tipos = 410, Tokens = 1062, Diversidad léxica = 0.3861 (38.61%)
spaCy: Tipos = 412, Tokens = 1098, Diversidad léxica = 0.3752 (37.52%)
```

En la Tarea 3, la normalización cambió bastante las cosas. Al usar técnicas como convertir todo a minúsculas, quitar tildes, signos de puntuación, números, espacios de más y stop words, se redujeron mucho los tokens (488 en NLTK, 486 en spaCy) y los tipos (308-319 en NLTK, 296-303 en spaCy). Esto pasó porque se eliminaron palabras vacías (como «the», «and») y elementos que no aportan al significado, como comas o números.

El stemming y la lematización también ayudaron a unificar palabras. Por ejemplo, «students» se volvió «student» (lematización) o «studen» (stemming), lo que bajó los tipos más que los tokens, aumentando la diversidad léxica. Los resultados fueron:

```
NLTK (Stemming): Tipos = 308, Tokens = 488, Diversidad léxica = 0.6311
NLTK (Lematización): Tipos = 319, Tokens = 488, Diversidad léxica =
0.6537
spaCy (Stemming): Tipos = 303, Tokens = 486, Diversidad léxica = 0.6235
spaCy (Lematización): Tipos = 296, Tokens = 486, Diversidad léxica
= 0.6091
```

Las diferencias entre NLTK y spaCy siguen ahí, como en la Tarea 2, por cómo tokenizan. Por ejemplo, NLTK toma «furstramdohr» como un solo token, pero spaCy lo separa. Sin embargo, al quitar guiones y otros signos en la normalización, estas diferencias son menos marcadas. La lematización de spaCy es más conservadora, dando menos tipos (296 vs. 319 de NLTK) y una diversidad léxica un poco menor.

En resumen, la normalización limpió el texto de ruido y unificó palabras, haciendo que la diversidad léxica suba bastante comparada con la Tarea 2. Ambas bibliotecas dieron resultados parecidos, mostrando que son efectivas para este tipo de análisis.