INSTITUTO BELTRAN

Alumno: Ariel Colatto

2do Año de IA

Materia: Técnicas de procesamiento del habla

Informe Canalización

El objetivo de este informe es describir y analizar el proceso de procesamiento de lenguaje natural (NLP) aplicado a un corpus de oraciones. Este análisis se realizó mediante una **canalización (pipeline)** que permite limpiar, transformar y extraer información útil del texto, utilizando herramientas de la biblioteca NLTK y técnicas estadísticas como TF-IDF.

El análisis se llevó a cabo en varias etapas secuenciales, que se describen a continuación:

1. Tokenización

Se dividió el texto en unidades mínimas significativas (palabras), eliminando signos de puntuación y separando las oraciones en tokens individuales.

2. Eliminación de Stopwords

Se eliminaron las palabras vacías del idioma inglés (como “is”, “and”, “the”), que no aportan significado relevante al análisis. Esto permitió enfocarse en los términos que aportan información contextual.

3. Lematización

Se redujeron las palabras a su forma base (por ejemplo, “compiled” se transformó en “compile”), para evitar duplicidad en el conteo de significados equivalentes.

4. Preparación del Corpus

Cada oración fue limpiada y lematizada individualmente, generando un corpus listo para análisis vectorial. Esto incluyó la reconstrucción del corpus con palabras significativas en su forma base, sin redundancias ni palabras irrelevantes.

5. Vectorización TF-IDF

Se aplicó la técnica de TF-IDF (Term Frequency - Inverse Document Frequency) para cuantificar la importancia de cada término en el corpus. Esto permitió identificar las palabras más relevantes en relación con la totalidad del texto, evitando que palabras frecuentes en todas las oraciones se sobredimensionen.

6. Análisis de Frecuencia

Se realizó un análisis estadístico de la frecuencia de las palabras más comunes en el corpus, así como la detección de palabras repetidas en cada oración y aquellas con menor frecuencia.

7. Visualización de Resultados

Se graficó la distribución de frecuencias de palabras tanto con stopwords como sin ellas y lematizadas, permitiendo observar visualmente cuáles son los términos predominantes**.**

El análisis generó un vocabulario de 51 palabras clave relevantes, eliminando redundancias semánticas gracias a la lematización. Entre estas palabras se destacan**:**

- **Python, javascript, compile, language, use, rust, web, data, science, application.**

Palabra menos usada:

La palabra menos frecuente fue compile, apareciendo solo una vez.

Palabra más repetida en cada oración:

Se determinó la palabra más repetida dentro de cada oración. En varias ocasiones, se repitieron términos como language, type, python y javascript, lo cual refuerza su protagonismo temático.

Palabra mas repetida en todo el texto: Las 6 palabras más usadas fueron: [('python', 7), ('javascript', 7), ('cplus', 5), ('rust', 5), ('interpret', 3), ('language', 3)]

Analisis td-idf

"python interpret language cplus compile language'

```
[[0.        0.        0.        0.        0.        0.
  0.        0.        0.        0.        0.44570103 0.26466657
  0.        0.        0.        0.        0.        0.
  0.        0.        0.        0.        0.        0.
  0.        0.        0.331481  0.        0.        0.
  0.7577732(compile) 0.        0.        0.        0.        0.
  0.        0.21726097(python) 0.        0.        0.        0.
  0.        0.        0.        0.        0.        0.
  0.        0.        0.        0.        0.        0.       ]
Vocabulario:
['application' 'artificial' 'available' 'backend' 'beginner' 'browser'
 'cloud' 'code' 'community' 'compilation' 'compile' 'cplus' 'data'
 'development' 'due' 'dynamically' 'ecosystem' 'ensures' 'execution'
 'experienced' 'extensive' 'go' 'great' 'ideal' 'include' 'intelligence'
 'interpret' 'java' 'javascript' 'js' 'language' 'large' 'library'
 'nature' 'node' 'number' 'programmer' 'python' 'require' 'run' 'rust'
 'science' 'security' 'server' 'slow' 'statically' 'strong' 'suitable'
 'typed' 'use' 'various' 'weakly' 'web' 'widely']
```

El de mayor peso es compile: tiene mucho peso en la oración, pero no en el resto del texto, al contrario de python que aparece muchas veces en todo el texto y por eso acá no tiene tanto peso

Grafico

Al final del texto hice gráficos de frecuencia con y sin stopwords. Los gráficos mostraron claramente el dominio de ciertas palabras clave, especialmente python y javascript, lo que confirma su alta representación en el corpus.