Vocabulary challenge
Alumnos: Mar Bazúa y Néstor Medina

1. Obtención y procesamiento del libro

Se descargó el libro *Los Miserables* de Victor Hugo desde la plataforma de la Fundación Carlos Slim. Este libro está en formato EPUB, el cual es un formato electrónico estructurado en HTML. Se tuvo que convertir el archivo en CSV, para ello se utilizó la librería ebooklib.Se recorrieron los elementos del libro, extrayendo los párrafos de texto y almacenándolos en un archivo CSV usando pandas. Esto permitió facilitar el procesamiento posterior.

2. Limpieza del texto

El archivo CSV fue procesado para eliminar caracteres innecesarios, espacios en blanco adicionales y cualquier contenido no textual relevante. También se eliminaron los signos de puntuación y se convirtieron todas las palabras a minúsculas para estandarizar el procesamiento.

```
### Funcion para limpiar el texto
def clean_text(text):
   if not isinstance(text, str):  # Si el valor no es string,
convertirlo
       return ""
   text = text.lower()  # Convertir a minúsculas
   text = unidecode.unidecode(text)  # Eliminar acentos
   text = re.sub(r"\s+", " ", text)  # Remover espacios extra
   text = text.translate(str.maketrans("", "", string.punctuation))  #
Eliminar puntuación
   text = text.strip()  # Eliminar espacios al inicio y al final

   return text
```

3. Creación del vocabulario

Siguiendo las instrucciones, se generó un vocabulario con todas las palabras únicas del libro. Este vocabulario se almacenó en formato parquet para optimizar el almacenamiento y la rapidez de acceso.

```
# Guardar vocabulario en Parquet
vocab_df = pd.DataFrame(list(vocabulary.items()), columns=["Word",
"Index"])
vocab_parquet_path = "/content/drive/My Drive/Colab
Notebooks/datos/vocabulary.parquet"
vocab_df.to_parquet(vocab_parquet_path, index=False)
```

4. Estadísticas del texto

Se realizaron diferentes análisis sobre el texto procesado:

- **Cantidad total de palabras en el texto original**: Se contó el total de palabras después de la limpieza. Considerando stopwords y también eliminándolas.

Vocabulary challenge
Alumnos: Mar Bazúa y Néstor Medina

- **Cantidad de palabras únicas en el vocabulario**: Se determinó la cantidad de palabras distintas en el texto.
- **100 palabras más frecuentes**: Se generó una lista con las palabras más utilizadas en el libro.
- **100 palabras menos frecuentes**: Se identificaron las palabras que aparecen menos veces.

Esto se realizó aprovechando las bondades de la librería collections y tokenizando el libro de la siguiente forma:

```
# Contador de palabras basado en el vocabulario
word_counts = collections.Counter()
word_counts_nostop_words = collections.Counter()

# Contar frecuencia solo para palabras en el vocabulario
for tokens in df["Tokens"]:
   for token in tokens:
       if token in vocabulary:  # Solo contar palabras del vocabulario
           word_counts[token] += 1

# Contar frecuencia solo para palabras en el vocabulario
for tokens in df["Tokens_no_stopwords"]:
   for token in tokens:
       if token in vocabulary:  # Solo contar palabras del vocabulario
           word_counts_nostop_words[token] += 1

#Calcular estadísticas considerando todas las palabras
total_words = sum(word_counts.values())  # Total de palabras en el
texto
unique_words = len(word_counts)  # Palabras únicas en el vocabulario
most_common_words = word_counts.most_common(100)  # 100 más frecuentes
least_common_words = word_counts.most_common()[-100:]  # 100 menos
frecuentes

#Calcular estadísticas quitando stop_words
total_words_no_sw = sum(word_counts_nostop_words.values())  # Total de
palabras en el texto
unique_words_no_sw = len(word_counts_nostop_words)  # Palabras únicas
en el vocabulario
most_common_words_no_sw = word_counts_nostop_words.most_common(100)  #
100 más frecuentes
least_common_words_no_sw =
word_counts_nostop_words.most_common()[-100:]  # 100 menos frecuentes
```

Estos análisis proporcionan una idea del uso del lenguaje en *Los Miserables* y pueden ser de utilidad en aplicaciones de procesamiento de lenguaje natural.

5. Conclusión

El procesamiento del libro permitió transformar un texto en bruto en un conjunto estructurado de datos listos para su análisis. Se logró estandarizar, almacenar y extraer información valiosa sobre la distribución de palabras en la obra. Sería interesante realizar este ejercicio para libros del mismo autor, y comparar las coincidencias en el vocabulario, o también obras del mismo género.