Cecilia Gómez Castañeda
En conjunto: con Marcos Cortes

Este reporte contiene la descripción de los pasos realizados para preprocesar el vocabulario del libro Los Miserables. El procesamiento se realizó utilizando Python.

Primero se cargó el archivo con el contenido del libro en txt utilizando las funciones propias del lenguaje python para lectura de documentos.

Una vez cargado el contenido se aplicó una función de limpieza para eliminar, acentos, caracteres especiale y dejar las palabras en minúsculas, para esto se utilizó la librería re.

```python
# Función para limpiar el texto
def limpiar_texto(texto):
    # Convertir a minúsculas
    texto = texto.lower()

    # Eliminar acentos
    texto = re.sub(r'[áàäâ]', 'a', texto)
    texto = re.sub(r'[éèëê]', 'e', texto)
    texto = re.sub(r'[íìïî]', 'i', texto)
    texto = re.sub(r'[óòöô]', 'o', texto)
    texto = re.sub(r'[úùüû]', 'u', texto)
    texto = re.sub(r'[ý]', 'y', texto)
    texto = re.sub(r'[ñ]', 'n', texto)

    # Eliminar caracteres especiales y números, dejando solo letras y espacios
    texto = re.sub(r'[^a-z\s]', '', texto)

    return texto
```

Se recorrió la totalidad de las palabras en el texto, agregando a la lista vocabulario cada palabra nueva encontrada, y cuando existe la palabra en la lista incrementando el contador para conocer la frecuencia de cada palabra. Una vez teniendo la frecuencia de las palabra se realizó el ordenamiento en función de la frecuencia d para poder conocer las palabras más y menos frecuentes:

```python
# Ordena por frecuencia
valores_ord = sorted(vocabulario.items(), key=operator.itemgetter(1), reverse=True)
```

Finalmente se utilizo la libreria de python para convertir de un dataframe a un formato parquet

Resultados

Total de palabras:109261

Total de palabras no repetidas:13105

100 palabras son utilizadas con mayor frecuencia:

de	5325	pero	372	he	182	sido	124
la	3918	hombre	363	ser	181	casa	123
que	3818	si	358	muy	178	son	121
el	3394	sus	344	javert	175	aqui	120
y	3123	todo	327	nada	174	noche	119
en	2836	me	326	mismo	173	hecho	118
a	2489	sin	311	o	164	tres	115
se	1681	obispo	286	os	163	dia	114
un	1601	dijo	281	poco	158	luego	113
no	1498	cuando	274	tan	158	cabeza	113
los	1353	estaba	273	bien	157	decir	112
una	1319	sobre	269	ni	156	voz	111
su	1245	dos	264	ella	155	alli	107
por	936	este	261	quien	151	ojos	107
las	935	aquel	253	alcalde	149	monsenor	105
con	924	mi	244	vez	148	aun	105
habia	858	ya	229	despues	146		
del	813	hacia	219	fue	145		
al	756	yo	218	todos	141		
es	749	esto	218	puerta	137		
lo	719	madeleine	214	anos	136		
le	667	tenia	212	hubiera	133		
era	650	jean	200	cual	133		
como	572	ha	199	donde	131		
mas	513	fantine	194	dios	130		
para	504	valjean	192	mujer	127		
senor	447	aquella	190	momento	125		
esta	414	hay	186	tiempo	124		

A continuación se listan 100 palabra utilizadas únicamente 1
reprobo, emocionantes, florecer, palidos, fulgor, sepultura, recluyo, lugares, conversaciones, bejean, bojean, boujean, almibarado, rehusado, perillanes, sucia, abundaron, abonada, drapeau, blanc, ensenara, partidarios, despavorida, reflexionando, puestos, velaban, colgo, esperara, inconscientemente, ensimismamiento, candela, boquiabierta, retenido, embargada, barrote, guardaria, maestra, lateral, registrado, conducian, peldanos, deshecha, huella, penultima, obtuvo, envolvio, embalaba, mordiendo, comprobado, migas, encontradas, pesquisas, enrojecidos, violencias, integros, entranas, obligan, doblar, leerlo, servira, sonidos, inarticulados, persiguiendome, turbaria, alboroto, murmullos, protestas, ambiente, respirable, integramente, objecion, correcto, amuralladas, retirarse, quedarse, aventurar, desfallecer, insista, evadido, mintio, seguidas, holocausto, valga, reparo, singularidad, bujia, marchando, brumas, alejaba, devuelta, posiblemente, reservar, simplifico, estricto, enterrada, gratuito, cementerio, encontrados, sufrio, promiscuidad

Conclusiones

El ejercicio permitió practicar sobre limpiez de datos en formato texto, también se nota que este proceso puede ser engorroso, sin embargo, no presenta mucha dificultad.

El hecho de que existan bastantes librerías para hacer funciones de ordenamiento, lectura de archivos y conversión de datos a formato parquet facilita mucho el desarrollo de estos tipos de trabajo y permite avanzar más rápido al siguiente paso de análisis.