Marco de Referencia

1. Expresiones regulares: Las expresiones regulares se usa como alternativa para la limpieza de datos combinándolo con la función regexp_replace, en esta parte lo que se busca es hacer uso de expresiones regulares para detectar los caracteres diferentes a [^a-zA-Z] y eliminarlos.
   En ese primer caso se buscaba eliminar todos los caracteres especiales dejando solo las letras contenidas en el alfabeto de la 'a' a la 'z' incluyendo mayúsculas o minúsculas.
   Y nuevamente teniendo en cuenta que se nota que en el sistema se detectan palabras separadas por mas de 2 o 3 espacios se usa una siguiente expresión regular [\s+]{2,} para detectar ese error y corregirlo eliminando esos espacios y dejando solo 1, para evitar que al tokenizar la columna que contiene el texto queden tokens con espacios vacios.

https://platzi.com/blog/expresiones-regulares-python/

https://docs.aws.amazon.com/es_es/redshift/latest/dg/REGEXP_REPLACE.html

2. Tokenizer: después de completar los anteriores pasos de limpieza se busca como remover las stop words, pero para este paso se requiere tener tokenizado el texto por lo cual se procede a hacer lo siguiente

```
In [7]: tokenization=Tokenizer(inputCol='clean1',outputCol='tokens')

In [8]: tokenized_df=tokenization.transform(df_2)

In [9]: tokenized_df.select('tokens').show(2,False)
|[content]
|
|[washington, congressional, republicans, have, a, new, fear, when, it, come
s, to, their, health, care, lawsuit, against, the, obama, administration, the
y, might, win, the, incoming, trump, administration, could, choose, to, no, l
onger, defend, the, executive, branch, against, the, suit, which, challenges,
the, administrations, authority, to, spend, billions, of, dollars, on, healt
h, insurance, subsidies, for, and, americans, handing, house, republicans, a,
big, victory, on, issues, but, a, sudden, loss, of, the, disputed, subsidies,
could, conceivably, cause, the, health, care, program, to, implode, leaving,
millions, of, people, without, access, to, health, insurance, before, republi
cans, have, prepared, a, replacement, that, could, lead, to, chaos, in, the,
insurance, market, and, spur, a, political, backlash, just, as, republicans,
gain, full, control, of, the, government, to, stave, off, that, outcome, repu
blicans, could, find, themselves, in, the, awkward, position, of, appropriati
ng, huge, sums, to, temporarily, prop, up, the, obama, health, care, law, ang
ering, conservative, voters, who, have, been, demanding, an, end, to, the, la
w, for, years, in, another, twist, donald, j, trumps, administration, worrie
```

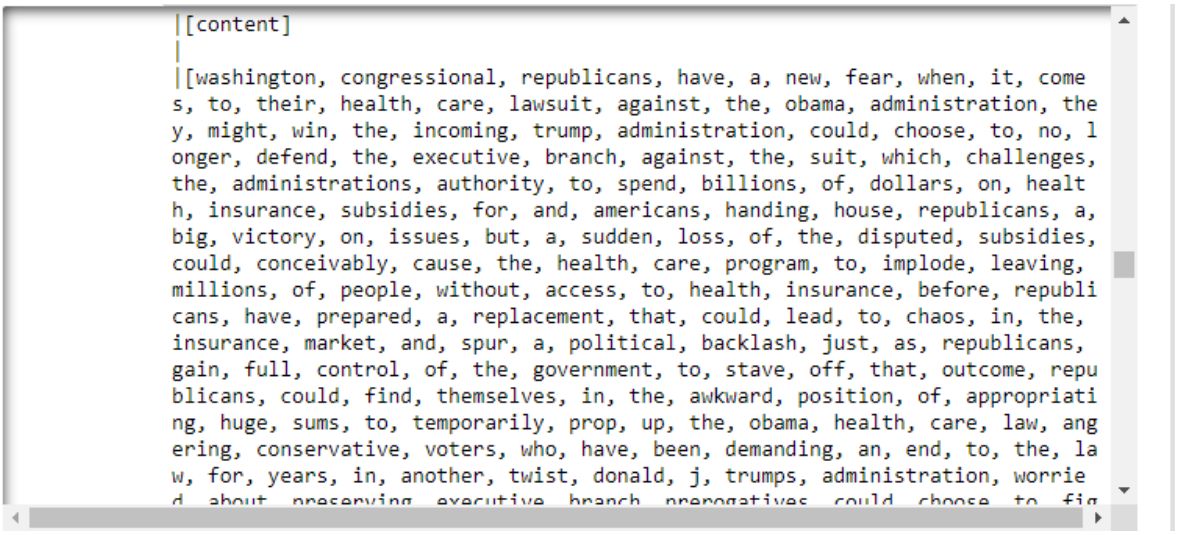
Usando una librería de pyspark v2.2.0 (pyspark.ml.feature import Tokenizer)

https://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html?highlight=import%20tokenizer#pyspark.ml.feature.Tokenizer

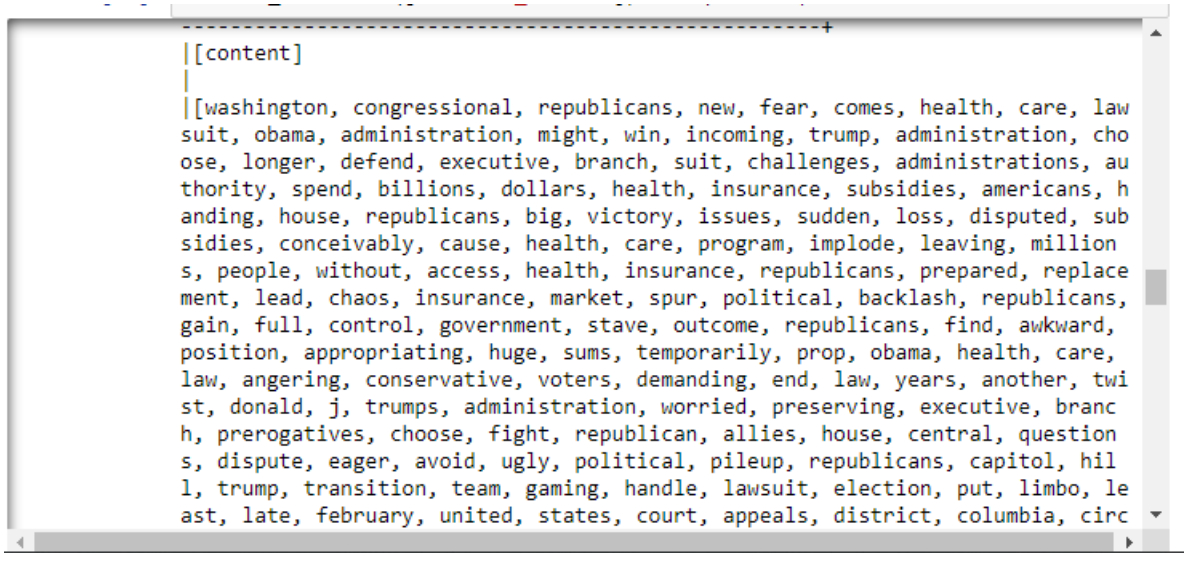
3. StopWordsRemover: esta librería se encarga de remover algunas palabras del dataset ya tokenizado como se muestra una aproximación en el siguiente ejemplo

id	raw	filtered
0	[I, saw, the, red, baloon]	[saw, red, baloon]
1	[Mary, had, a, little, lamb]	[Mary, little, lamb]

Esto es en el dataset antes de correr el comando de StopWordsRemover



Después de correr el StopWordsRemover se evidencia como algunas de las palabras son eliminadas



https://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html?highlight=import%20tokenizer#pyspark.ml.feature.StopWordsRemover

4. Stemming y Lemmatization: para esta ultima parte de la limpieza de los datasets se requiere hacer una limpieza de los datos convirtiendo algunos de los datos, en la siguiente se muestra un ejemplo aproximado de como debería ser

Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Se adjunta esta imagen para mostrar que a pesar de que se instalo la librería necesaria y se cumplieron con los requisitos para usar el 'import nltk' se sigue presentando el erro como si no se tuvieran instaladas esas librerias en el sistema.

```
In [16]: pip install nltk
```

```
Requirement already satisfied: nltk in /opt/anaconda3/lib/python3.7/site-packag
es (3.4.4)
Requirement already satisfied: six in /opt/anaconda3/lib/python3.7/site-package
s (from nltk) (1.12.0)
Note: you may need to restart the kernel to use updated packages.
```

```
In [17]: import nltk
         from nltk.stem import WordNetLemmatize
```

```
---------------------------------------------------------------------------
ImportError                               Traceback (most recent call last)
<ipython-input-17-26c8c0d5dcfa> in <module>
      1 import nltk
----> 2 from nltk.stem import WordNetLemmatize

ImportError: cannot import name 'WordNetLemmatize' from 'nltk.stem' (/opt/anaco
nda3/lib/python3.7/site-packages/nltk/stem/__init__.py)
```

```
In [ ]: # Init the Wordnet Lemmatizer
        lemmatizer = WordNetLemmatizer()

        # Lemmatize Single Word
        print(lemmatizer.lemmatize("bats"))
```

https://www.machinelearningplus.com/nlp/lemmatization-examples-python/