Documentación

Se llaman las librerías que serán utilizadas. Entre ellas una libreria propia:

-utilitools: Libreria de varios usos, se usa para configurar el arbol de rutas particularmente.

In [ ]:
```
import pandas as pd
import numpy as np
import os
from utilitools import *
```

Se procede a leer el arbol con las rutas

In [ ]:
```
data = leer_paths('data')
recomendador = leer_paths('recomendador')
scripts = leer_paths('scripts')

# Esta función carga el archivo .csv que se encuentra en la ruta path con el nombre csv_
def leer_csv(path,csv_file):
    '''
    Inserta ruta de carpeta donde se encuentra el archivo y nombre de archivo para hacer
    '''
    for file in os.listdir(path):
        file_path = os.path.join(path,file)

        if csv_file in file_path:
            df = pd.read_csv(file_path)
    return df

df_clientes = leer_csv(data,'clientes.csv')
df_noticias = leer_csv(data,'noticias.csv')
df_cn = leer_csv(data,'clientes_noticias.csv')
df_clientes = leer_csv(data,'clientes.csv')
```

A continuación, creamos una función que toma en cuenta las expresiones regulares, además, se define una lista con las abreviaciones de tipos de sociedades comerciales en Colombia. Esto se realiza para homogeneizar los nombres con lo que se espera encontrar en las noticias y poder relacionar los datos de una mejor manera

In [ ]:
```
# Se crea una función que realiza limpieza a la columna nombre, a través de expresiones
def limpieza_clientes(df_clientes):
    df_clientes['nombre'] = df_clientes['nombre'].str.lower()
    df_clientes['nombre'] = df_clientes['nombre'].str.replace(r'[^a-z 0-9]', '',regex=Tr
    df_clientes['nombre'] = df_clientes['nombre'].str.replace(r' s$', '',regex=True)
    df_clientes['nombre'] = df_clientes['nombre'].str.replace(r' e$', '',regex=True)
    df_clientes['nombre'] = df_clientes['nombre'].str.replace(r' d$', '',regex=True)

    df_clientes['nombre'] = df_clientes['nombre'].str.replace('cia','compania')

    tipi = [
        'sas',
        'sa',
        's a',
        's a s',
        'ltda',
        'limited',
        'limitada',
        'limit',
```

```
        'limitad',
        'sca',
        's c a',
        'esp',
        'e s p',
        'inc',
        's en c',
        'ci',
        'c i',
        'ltd',
        ]

    regstr = ''

    for term in tipi:
        regstr+=r'\s' + term +r'\b' + '|'

    df_clientes['nombre'] = df_clientes['nombre'].str.replace(regstr,'',regex=True)
    df_clientes['nombre'].str.strip()

    return df_clientes

df_clientes = limpieza_clientes(df_clientes)

#df_clientes['nombre'].to_excel('nombres.xlsx',index=False)
```

Out[ ]:

	nit	nombre	desc_ciiu_division	desc_ciuu_grupo	desc_ciiuu_clase	subsec
0	805027024	supermercado la gran colombia	ELABORACION DE PRODUCTOS ALIMENTICIOS	PROCESAMIENTO Y CONSERVACION DE CARNE, PESCADO...	PROCESAMIENTO Y CONSERVACION DE CARNE Y PRODUC...	CARNES FRIAS
1	890100026	camaguey	ELABORACION DE PRODUCTOS ALIMENTICIOS	PROCESAMIENTO Y CONSERVACION DE CARNE, PESCADO...	PROCESAMIENTO Y CONSERVACION DE CARNE Y PRODUC...	CARNES FRIAS
2	801004045	don pollo	ELABORACION DE PRODUCTOS ALIMENTICIOS	PROCESAMIENTO Y CONSERVACION DE CARNE, PESCADO...	PROCESAMIENTO Y CONSERVACION DE CARNE Y PRODUC...	CARNES FRIAS
3	900319372	red carnica	ELABORACION DE PRODUCTOS ALIMENTICIOS	PROCESAMIENTO Y CONSERVACION DE CARNE, PESCADO...	PROCESAMIENTO Y CONSERVACION DE CARNE Y PRODUC...	CARNES FRIAS
4	800198020	carnes casablanca	ELABORACION DE PRODUCTOS ALIMENTICIOS	PROCESAMIENTO Y CONSERVACION DE CARNE, PESCADO...	PROCESAMIENTO Y CONSERVACION DE CARNE Y PRODUC...	CARNES FRIAS
...	...	...	...	...	...	...
1502	890209174	ismocol	ACTIVIDADES DE SERVICIOS DE APOYO PARA LA EXPL...	ACTIVIDADES DE APOYO PARA LA EXTRACCION DE PET...	ACTIVIDADES DE APOYO PARA LA EXTRACCION DE PET...	EXTRACCION DE PETROLEO
1503	830069311	nabors drilling international bermuda	ACTIVIDADES DE SERVICIOS DE APOYO PARA LA EXPL...	ACTIVIDADES DE APOYO PARA LA EXTRACCION DE PET...	ACTIVIDADES DE APOYO PARA LA EXTRACCION DE PET...	EXTRACCION DE PETROLEO
1504	830130106	soenergy international colombia	ACTIVIDADES DE SERVICIOS DE APOYO PARA LA EXPL...	ACTIVIDADES DE APOYO PARA LA EXTRACCION DE PET...	ACTIVIDADES DE APOYO PARA LA EXTRACCION DE PET...	GAS
1505	890110188	independence drilling	ACTIVIDADES DE SERVICIOS DE	ACTIVIDADES DE APOYO PARA LA	ACTIVIDADES DE APOYO PARA LA	EXTRACCION DE PETROLEO

			APOYO PARA LA EXPL...	EXTRACCION DE PET...	EXTRACCION DE PET...	
1506	891102723	mecanicos asocompaniados	ACTIVIDADES DE SERVICIOS DE APOYO PARA LA EXPL...	ACTIVIDADES DE APOYO PARA LA EXTRACCION DE PET...	ACTIVIDADES DE APOYO PARA LA EXTRACCION DE PET...	EXTRACCION DE PETROLEO

1507 rows × 6 columns

In [ ]:
```
df_cn
```

Out[ ]:

	nit	news_id	news_url_absolute	news_init_date	news_final_date
0	900378212	news10006	https://www.bluradio.com/economia/precio-dolar...	2022-07-30	2022-08-14
1	900378212	news10011	https://www.semana.com/economia/macroeconomia/...	2022-07-30	2022-08-14
2	860034313	news10011	https://www.semana.com/economia/macroeconomia/...	2022-07-30	2022-08-14
3	900378212	news10015	https://elcomercio.pe/respuestas/que/gustavo-p...	2022-07-30	2022-08-14
4	900166896	news10015	https://elcomercio.pe/respuestas/que/gustavo-p...	2022-07-30	2022-08-14
...	...	...	...	...	...
74704	800230209	news99997	https://www.laopinion.com.co/economia/en-cucut...	2022-07-15	2022-07-30
74705	890209174	news99997	https://www.laopinion.com.co/economia/en-cucut...	2022-07-15	2022-07-30
74706	830069311	news99997	https://www.laopinion.com.co/economia/en-cucut...	2022-07-15	2022-07-30
74707	830130106	news99997	https://www.laopinion.com.co/economia/en-cucut...	2022-07-15	2022-07-30
74708	890110188	news99997	https://www.laopinion.com.co/economia/en-cucut...	2022-07-15	2022-07-30

74709 rows × 5 columns

In [ ]:
```
```