Anexo 2 - Creación del conjunto de datos (Twitter y Google News),

Las fuentes de información utilizadas en este trabajo son Twitter y Google News. En este documento se encuentra el programa utilizado para:

- Descargar la información de Twitter
- Descargar la información de Google News
- Importar información a base de datos DynamoDB - AWsS
- Unificar el conjunto de datos

Descarga de informacion de Twitter

Es utilizada el API de Twitter para obtener la informacion

In [47]:

```
from __future__ import print_function # Python 2/3 compatibility
import boto3
import json
import decimal
import pandas as pd
```

In [ ]:

```
```

In [ ]:

```
# Clase para convertir DynamoDB item a JSON.
class DecimalEncoder(json.JSONEncoder):
    def default(self, o):
        if isinstance(o, decimal.Decimal):
            if abs(o) % 1 > 0:
                return float(o)
            else:
                return int(o)
        return super(DecimalEncoder, self).default(o)
```

In [3]:

```
# Funcion para eliminar elementos vacios dentro de un diccionario (
def clean_empty(d):
    if not isinstance(d, (dict, list)):
        return d
    if isinstance(d, list):
        return [v for v in (clean_empty(v) for v in d) if v]
    return {k: v for k, v in ((k, clean_empty(v)) for k, v in d.ite
```

In [7]:

```
# Funcion para importar elementos a dynamobd
def import_aws_dynamobd(path,table):
    #path = '/Users/christianorrego/Downloads/udem-reputacion/'
    #file_name =  'Top 10 Merco 4 Nov -7 dias Colombia/alpinaAT.txt
    with open(path, 'r') as f:
        lines = f.readlines()

    lines  = [json.loads(x,parse_float = decimal.Decimal) for x in

    #agregar llave e importar linea
    for i in range(len(lines)):
        lines[i]["key"] = path.split("/")[-2]+"/"+path.split("/")[-

        element = lines[i]

        element = clean_empty(element) #eliminar elementos vacios -

        response = table.put_item(
           Item=element
        )
        print("Exitoso "+str(i)+"--------------------")
        #print("PutItem succeeded:")
        #print(json.dumps(response, indent=4, cls=DecimalEncoder))
```

In [22]:

```
# http://dynamodb.us-east-1.amazonaws.com  - production url
dynamodb = boto3.resource('dynamodb', region_name='us-east-1', endp
table = dynamodb.Table('twitter')
```

Importar datset descargado de Twitter a DynamoBD - local

Previamente fue creada la tabla Twitter donde se almacenaran todos los json resultados de las peticiones al API de Twitter. Script de creacion de tabla en terminal (linea de comandos linux)
aws dynamodb --endpoint-url http://localhost:8000 (http://localhost:8000) create-table
--table-name twitter
--attribute-definitions
AttributeName=key,AttributeType=S
--key-schema AttributeName=key,KeyType=HASH
--provisioned-throughput ReadCapacityUnits=1,WriteCapacityUnits=1

In [12]:

```
origin = '/Users/christianorrego/Downloads/udem-reputacion'
folders = ['Top 10 Merco 3 DIC -7 dias Colombia', 'Top 10 Merco 4 N
file_names = ['alpinaAT.txt', 'alpinaHASH.txt', 'bancolombiaAT.txt'

paths = []

#creacion del array de los nombres de archivos a subir
for folder in folders:
    for file in file_names:
        paths.append(origin+"/"+folder+"/"+file)

# importacion de archivos
for path in paths:
    import_aws_dynamobd(path,table)
    print("----------------------- EXITO archivo "+path)
```

```
Exitoso 0------------------
Exitoso 1------------------
Exitoso 2------------------
Exitoso 3------------------
Exitoso 4------------------
Exitoso 5------------------
Exitoso 6------------------
Exitoso 7------------------
Exitoso 8------------------
Exitoso 9------------------
Exitoso 10-----------------
Exitoso 11-----------------
Exitoso 12-----------------
Exitoso 13-----------------
Exitoso 14-----------------
Exitoso 15-----------------
Exitoso 16-----------------
Exitoso 17-----------------
Exitoso 18-----------------
```

In [23]:

```
#Cantidad de elementos almacenados en la BD
print(table.item_count)
```

```
28069
```

Transformar Dynambo BD y crear Pandas Dataframe

In [24]:

```
response = table.scan()
data = response['Items']

while 'LastEvaluatedKey' in response:
    response = table.scan(ExclusiveStartKey=response['LastEvaluated
    data.extend(response['Items'])
```

In [121]:

```
dataset = []

# Obtener los campos requeridos de la respuesta del api de twitter
for ele in data:
    tweet = {}
    tweet['key'] = ele['key']
    tweet['key_row'] = ele['key'].split('-')[-1].replace("row=","")
    tweet['key_folder'] =  ele['key'].split('/')[-2]
    tweet['key_file'] =  ele['key'].split('/')[-1].split("-")[-2]
    tweet['company'] = ele['key'].split('/')[-1].split("-")[-2].rep
    tweet['id_str'] = ele['id_str']
    tweet['created_at'] = ele['created_at']
    tweet['lang'] = ele['lang']
    tweet['in_reply_to_status_id_str'] = ele.get('in_reply_to_statu
    tweet['re_tweet_created_at'] = ele.get('retweeted_status',{}).g
    tweet['re_tweet_id_str'] = ele.get('retweeted_status',{}).get('
    tweet['re_tweet_in_reply_to_status_id_str'] = ele.get('retweete
    tweet['retweeted_status'] = 'retweet' if ele.get('retweeted_sta
    tweet['truncated'] = ele.get('truncated', None)
    tweet['user_id_str'] = ele['user']['id_str']
    tweet['user_location'] = ele.get('user',{}).get('location',None
    tweet['user_name'] = ele['user']['name']
    tweet['user_screen_name'] = ele['user']['screen_name']
    tweet['user_description'] = ele.get('user',{}).get('description
    tweet['user_followers_count'] = ele.get('user',{}).get('followe
    tweet['text'] = ele['text']
    dataset.append(tweet)

pdDataset =  pd.DataFrame(dataset)
# Eliminar registros repetidos
pdDataset.drop_duplicates(subset='id_str', inplace = True)

#pdDataset.to_excel('Dataset-reputacion-12-12-2019.xls')
#pdDataset.to_csv('Dataset-reputacion-12-12-2019.csv',sep="|")
```

Peticiones al API de Amazon web service comprehend

Permite tener una respuesta del sentimiento detectado por AWS. Respuestas de sentimientos: POSITIVO, NEGATIVO, NEUTRO y MEZCLADO (multimples sentimientos detectados)

Documentacion del api:
https://boto3.amazonaws.com/v1/documentation/api/latest/reference/services/comprehen
(https://boto3.amazonaws.com/v1/documentation/api/latest/reference/services/comprehen

In [123]:

```
#Funcion para realizar la peticion al servicio de Amazon web servic
def aws_sentiment(text,client):
    response = client.detect_sentiment(
        Text=text,
        LanguageCode='es'
    )
    return response
```

In [126]:

```
client = boto3.client('comprehend')

respuestas = []
i = 0
for index, row in clean_pdDataset.iterrows():
    text = row['text']
    id_str = row['id_str']
    response =  aws_sentiment(text,client)
    #response = {'Sentiment': 'NEUTRAL', 'SentimentScore': {'Positi
    response['id_str'] =  id_str
    #
    respuestas.append(response)
    i =  i+1
    print("-----OK: "+ str(i))
```

```
-----OK: 1
-----OK: 2
-----OK: 3
-----OK: 4
-----OK: 5
-----OK: 6
-----OK: 7
-----OK: 8
-----OK: 9
-----OK: 10
-----OK: 11
-----OK: 12
-----OK: 13
-----OK: 14
-----OK: 15
-----OK: 16
-----OK: 17
-----OK: 18
-----OK: 19
```

In [127]:

```
respuestas
```

Out[127]:

```
[{'Sentiment': 'NEUTRAL',
  'SentimentScore': {'Positive': 0.00011496161459945142,
   'Negative': 6.106298678787425e-05,
   'Neutral': 0.9998083710670471,
   'Mixed': 1.5553940102108754e-05},
  'ResponseMetadata': {'RequestId': 'a10652dd-bdca-4fad-a03e-a0763bcfda5c',
   'HTTPStatusCode': 200,
   'HTTPHeaders': {'x-amzn-requestid': 'a10652dd-bdca-4fad-a03e-a0763bcfda5c',
    'content-type': 'application/x-amz-json-1.1',
    'content-length': '166',
    'date': 'Thu, 12 Dec 2019 22:54:34 GMT'},
   'RetryAttempts': 0},
  'id_str': '1193353981252317185'},
 {'Sentiment': 'NEUTRAL',
  'SentimentScore': {'Positive': 0.2310795933008194.
```

In [128]:

```
with open('aws_comprehend_response.txt', 'w') as f:
    for item in respuestas:
        f.write("%s\n" % item)
    print("Respuestas almacenadas")
```

```
Respuestas almacenadas
```

In [138]:

```
responses_array = []
for response in respuestas:
    row = {}
    row['id_str'] = str(response['id_str'])
    row['aws_sentimiento'] = response['Sentiment']
    row['aws_score_positivo'] = response['SentimentScore']['Positiv
    row['aws_score_negativo'] = response['SentimentScore']['Negativ
    row['aws_score_neutro'] =   response['SentimentScore']['Neutral
    row['aws_score_mixed'] =     response['SentimentScore']['Mixed'
    responses_array.append(row)
```

In [149]:

```
result_pdDataset =  pd.merge(pdDataset,pd.DataFrame(responses_array
result_pdDataset['manual_response'] = None

result_pdDataset.to_excel('resultado_dataset-12-12-2019.xls')
```

In [154]:

```
result_pdDataset.groupby(by=['company','key_folder'])['key'].count(
```

Out[154]:

```
company           key_folder
BAVARIA_OFICIAL   Top 10 Merco 10 Dic - 7 dias Colombia
98
                  Top 10 Merco 11 Nov -7 dias Colombia
223
                  Top 10 Merco 3 DIC -7 dias Colombia
83
                  Top 10 Merco 4 Nov -7 dias Colombia
462
Cementos_Argos    Top 10 Merco 10 Dic - 7 dias Colombia
65
                  Top 10 Merco 11 Nov -7 dias Colombia
74
                  Top 10 Merco 3 DIC -7 dias Colombia
129
                  Top 10 Merco 4 Nov -7 dias Colombia
39
```

ECOPETROL_SA Top 10 Merco 10 Dic – 7 dias Colombia
999
Top 10 Merco 11 Nov –7 dias Colombia
989
Top 10 Merco 3 DIC –7 dias Colombia
1041
Top 10 Merco 4 Nov –7 dias Colombia
1526
EPMestamosahi Top 10 Merco 10 Dic – 7 dias Colombia
820
Top 10 Merco 11 Nov –7 dias Colombia
1522
Top 10 Merco 3 DIC –7 dias Colombia
1021
Top 10 Merco 4 Nov –7 dias Colombia
891
Nestle Top 10 Merco 10 Dic – 7 dias Colombia
43
Top 10 Merco 11 Nov –7 dias Colombia
65
Top 10 Merco 3 DIC –7 dias Colombia
75
Top 10 Merco 4 Nov –7 dias Colombia
25
UNColombia Top 10 Merco 10 Dic – 7 dias Colombia
2905
Top 10 Merco 11 Nov –7 dias Colombia
1323
Top 10 Merco 3 DIC –7 dias Colombia
2611
Top 10 Merco 4 Nov –7 dias Colombia
1540
alpina Top 10 Merco 10 Dic – 7 dias Colombia
155
Top 10 Merco 11 Nov –7 dias Colombia
164
Top 10 Merco 3 DIC –7 dias Colombia
47
Top 10 Merco 4 Nov –7 dias Colombia
182
bancolombia Top 10 Merco 10 Dic – 7 dias Colombia
944
Top 10 Merco 11 Nov –7 dias Colombia
1762
Top 10 Merco 3 DIC –7 dias Colombia

```
1216
                Top 10 Merco 4 Nov -7 dias Colombia
1700
grupo_nutresa    Top 10 Merco 10 Dic - 7 dias Colombia
36
                Top 10 Merco 11 Nov -7 dias Colombia
36
                Top 10 Merco 3 DIC -7 dias Colombia
94
                Top 10 Merco 4 Nov -7 dias Colombia
37
gruposura        Top 10 Merco 10 Dic - 7 dias Colombia
145
                Top 10 Merco 11 Nov -7 dias Colombia
113
                Top 10 Merco 3 DIC -7 dias Colombia
223
                Top 10 Merco 4 Nov -7 dias Colombia
91
Name: key, dtype: int64
```

In [151]:

```
result_pdDataset.columns
```

Out[151]:

```
Index(['key', 'key_row', 'key_folder', 'key_file', 'co
mpany', 'id_str',
       'created_at', 'lang', 'in_reply_to_status_id_st
r',
       're_tweet_created_at', 're_tweet_id_str',
       're_tweet_in_reply_to_status_id_str', 'retweete
d_status', 'truncated',
       'user_id_str', 'user_location', 'user_name', 'u
ser_screen_name',
       'user_description', 'user_followers_count', 'te
xt', 'aws_sentimiento',
       'aws_score_positivo', 'aws_score_negativo', 'aw
s_score_neutro',
       'aws_score_mixed', 'manual_response'],
      dtype='object')
```

Google

In [66]:

```
import pandas as pd
import json
import os
import ast
import re
import glob
```

In [67]:

```
def clean_text(text):
    text = text.strip('\n')
    text = text.replace('\\xa0',' ')
    return text
```

In [70]:

```
os.chdir('/Users/christianorrego/Downloads/udem-reputacion/GoogleNe

files = glob.glob('*')
google_news = []
for file in files:
    lines = []
    print(file)
    with open('./'+file, 'r',encoding='utf8') as f:
        lines = f.readlines()


    lines = [ast.literal_eval(x) for x in lines]
    for x in lines:
        x['company'] = file.replace('1-1000.txt','')

    google_news.extend(lines)

data = pd.DataFrame(google_news)


data['full_text'] = data['title'].map(str) + '. ' + data['desc'].ma
data['full_text_cln'] = data['full_text'].map(clean_text)
data.index.name = 'id'
#apply(lambda x : clean_text(x))
```

```
grupo_nutresa1-1000.txt
alpina1-1000.txt
unal1-1000.txt
epm1-1000.txt
bavaria1-1000.txt
bancolombia1-1000.txt
nestle1-1000.txt
ecopetrol1-1000.txt
argos1-1000.txt
grupo_sura1-1000.txt
```

In [109]:

```
import requests

def meaningCloud_sentiment(text):
    key ='d78557076586c0f9b1c9c35cecfbc3e3'
    url = "https://api.meaningcloud.com/sentiment-2.1"
    payload = "key="+key+"&lang=es&txt="+text+"&txtf=plain"
    headers = {'content-type': 'application/x-www-form-urlencoded'}
    response = requests.request("POST", url, data=payload.encode('u
    return response.text
```

In [99]:

```
data.head()
```

Out[99]:

id	title	media	date	desc	
0	Mayores ventas impulsan el beneficio del Grupo...	Portafolio.co (Comunicado de prensa) (blog)	25 oct. 2019	La utilidad neta del mayor productor de alimen...	https://www.portafolio.co/nego
1	Grupo Nutresa suma a su línea de café un negoc...	La República	18 sep. 2019	El Grupo Nutresa, líder en el mercado de alime...	https://www.larepublica.co/er
2	Grupo Nutresa: alimento para el mundo	Semana.com	25 may. 2019	Uno de ellos, el Grupo Nutresa, muestra el alt...	https://ww e
3	Así quedó el dividendo de Nutresa para este año	Dinero.com	26 mar. 2019	En la asamblea de accionistas de Nutresa se di...	https://www.dinero.com/emp
4	Grupo Nutresa cierra planta procesadora de Cun...	La República	1 sep. 2019	Alimentos Cárnicos S.A.S. aseguró que las oper...	https://www.larepublica.co/er

In [112]:

```
respuestas = []
i = 0
for index, row in data.iterrows():
    text = row['full_text_cln']
    id_str = index
    response =  ast.literal_eval(meaningCloud_sentiment(text))
    #response = {"status":{"code":"0","msg":"OK","credits":"1","rem
    response['id_google'] =  id_str
    #
    respuestas.append(response)
    i =  i+1
    print("-----OK: "+ str(i))
```

```
-----OK: 1
-----OK: 2
-----OK: 3
-----OK: 4
-----OK: 5
-----OK: 6
-----OK: 7
-----OK: 8
-----OK: 9
-----OK: 10
-----OK: 11
-----OK: 12
-----OK: 13
-----OK: 14
-----OK: 15
-----OK: 16
-----OK: 17
-----OK: 18
-----OK: 19
```

In [113]:

```
with open('goolge_response.txt', 'w') as f:
    for item in respuestas:
        f.write("%s\n" % item)
    print("Respuestas almacenadas")
```

```
Respuestas almacenadas
```

In [124]:

```
result_pdDataset_google =  pd.merge(data,pd.DataFrame(respuestas).s
result_pdDataset_google['manual_response'] = None

result_pdDataset_google.to_excel('resultado_dataset_google-12-12-20
```

In [123]:

```

```

Out[123]:

id	title	media	date	desc	
0	Mayores ventas impulsan el beneficio del Grupo...	Portafolio.co (Comunicado de prensa) (blog)	25 oct. 2019	La utilidad neta del mayor productor de alimen...	https://www.pc
1	Grupo Nutresa suma a su línea de café un negoc...	La República	18 sep. 2019	El Grupo Nutresa, líder en el mercado de alime...	https://www.la
2	Grupo Nutresa: alimento para el mundo	Semana.com	25 may. 2019	Uno de ellos, el Grupo Nutresa, muestra el alt...	
3	Así quedó el dividendo de Nutresa para este año	Dinero.com	26 mar. 2019	En la asamblea de accionistas de Nutresa se di...	https://www.c
4	Grupo Nutresa cierra planta procesadora de Cun...	La República	1 sep. 2019	Alimentos Cárnicos S.A.S. aseguró que las oper...	https://www.la

...	...	...	...	...	
1945	Grupo Sura avanza en el Índice de Sostenibilid...	Estrategia y Negocios	2 oct. 2018	Grupo SURA es parte de las 317 empresas selecc...	https://www.est
1946	Grupo Sura actualizó su direccionamiento estra...	Valora Inversiones (blog)	23 mar. 2018	El Grupo de Inversiones Suramericana (Grupo Su...	https://www.v
1947	En rebalanceo del Hcolsel de septiembre no se ...	valoraanalitik.com	19 sep. 2019	“En una menor proporción, resaltan las compras...	https://www.'
1948	El aprendizaje fue el reto en la implementació...	Dinero.com	10 jun. 2016	Desaprender para volver a aprender fue el reto...	https://www.c
1949	Suramericana SA cerró 2018 con un patrimonio n...	La República	29 ene. 2019	La filial de Grupo Sura, Suramericana S.A., di...	https://www.

1950 rows × 19 columns

In [118]:

```
data.columns
```

Out[118]:

```
Index(['title', 'media', 'date', 'desc', 'link', 'img'
, 'company', 'full_text',
       'full_text_cln'],
      dtype='object')
```

In [111]:

```
respuestas
```

Out[111]:

```
[{'status': {'code': '0',
   'msg': 'OK',
   'credits': '1',
   'remaining_credits': '19980'},
  'model': 'general_es',
  'score_tag': 'P',
  'agreement': 'AGREEMENT',
  'subjectivity': 'OBJECTIVE',
  'confidence': '100',
  'irony': 'NONIRONIC',
  'sentence_list': [{'text': 'Mayores ventas impulsan
el beneficio del Grupo Nutresa a ....',
    'inip': '0',
    'endp': '60',
    'bop': 'y',
    'confidence': '100',
    'score_tag': 'P',
    'agreement': 'AGREEMENT',
```

Twitter

In [1]:

```
import os
import tweepy as tw
import pandas as pd
import sys
import jsonpickle
```

In [2]:

```
consumer_key= 'fcNy8aVGV4fwAjjVzfY66ZRU5'
consumer_secret= 'OdeX5i61NFw04ljI7uoCe1NoHSQKardNjlnAvRgFIGCWnnS10
access_token= '2912019929-iHaP2u4rPI1bjKkpEEkUxh7McOG02sdQGhZCDPI'
access_token_secret= 'HyrP8T2REOhfeA7AiQGJzcNJcBPanXIPrwFSXuYQwntFx
```

In [5]:

```
auth = tw.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tw.API(auth, wait_on_rate_limit=True)
```

In [ ]:

```
search_words = "bancolombia -filter:retweets"
date_since = "2019-06-01"
```

In [ ]:

```
# Collect tweets
tweets = tw.Cursor(api.search,
              q=search_words,
              lang="es",
              since=date_since).items(1)
# Iterate and print tweets
for tweet in tweets:
    print(tweet)


#users_locs = [[tweet.user.screen_name, tweet.user.location] for tw
#users_locs
```

In [ ]:

```
auth = tw.AppAuthHandler(consumer_key, consumer_secret)

api = tw.API(auth, wait_on_rate_limit=True, wait_on_rate_limit_noti

if (not api):
    print ("Can't Authenticate")
    sys.exit(-1)
```

In [7]:

```
def busquedaTweeter(empresa):
    colombia = "4.570868,-74.29733299999998,680mi"
    original_searchQuery = empresa # this is what we're searching f

    searchQuery = "@"+original_searchQuery
```

```
# at
maxTweets = 10000000 # Some arbitrary large number
tweetsPerQry = 100  # this is the max the API permits
fName = original_searchQuery+'AT.txt' # We'll store the tweets


# If results from a specific ID onwards are reqd, set since_id
# else default to no lower limit, go as far back as API allows
sinceId = None

# If results only below a specific ID are, set max_id to that I
# else default to no upper limit, start from the most recent tw
max_id = -1000000

tweetCount = 0
print("Downloading max {0} tweets".format(maxTweets))
with open(fName, 'w') as f:
    while tweetCount < maxTweets:
        try:
            if (max_id <= 0):
                if (not sinceId):
                    new_tweets = api.search(q=searchQuery,geoco
                else:
                    new_tweets = api.search(q=searchQuery,geoco
                                            since_id=sinceId)
            else:
                if (not sinceId):
                    new_tweets = api.search(q=searchQuery,geoco
                                            max_id=str(max_id -
                else:
                    new_tweets = api.search(q=searchQuery,geoco
                                            max_id=str(max_id -
                                            since_id=sinceId)
            if not new_tweets:
                print("No more tweets found")
                break
            for tweet in new_tweets:
                f.write(jsonpickle.encode(tweet._json, unpickla
                        '\n')
            tweetCount += len(new_tweets)
            print("Downloaded {0} tweets".format(tweetCount))
            max_id = new_tweets[-1].id
        except tw.TweepError as e:
            # Just exit if any error
            print("some error : " + str(e))
```

```
            break

print ("Downloaded {0} tweets, Saved to {1}".format(tweetCount,



# HASH
searchQuery = "#"+original_searchQuery

maxTweets = 10000000 # Some arbitrary large number
tweetsPerQry = 100  # this is the max the API permits
fName = original_searchQuery+'HASH.txt' # We'll store the tweet


# If results from a specific ID onwards are reqd, set since_id
# else default to no lower limit, go as far back as API allows
sinceId = None

# If results only below a specific ID are, set max_id to that I
# else default to no upper limit, start from the most recent tw
max_id = -1000000

tweetCount = 0
print("Downloading max {0} tweets".format(maxTweets))
with open(fName, 'w') as f:
    while tweetCount < maxTweets:
        try:
            if (max_id <= 0):
                if (not sinceId):
                    new_tweets = api.search(q=searchQuery,geoco
                else:
                    new_tweets = api.search(q=searchQuery,geoco
                                            since_id=sinceId)
            else:
                if (not sinceId):
                    new_tweets = api.search(q=searchQuery,geoco
                                            max_id=str(max_id -
                else:
                    new_tweets = api.search(q=searchQuery,geoco
                                            max_id=str(max_id -
                                            since_id=sinceId)
            if not new_tweets:
                print("No more tweets found")
                break
            for tweet in new_tweets:
```

```
                    f.write(jsonpickle.encode(tweet._json, unpickla
                            '\n')
                tweetCount += len(new_tweets)
                print("Downloaded {0} tweets".format(tweetCount))
                max_id = new_tweets[-1].id
            except tw.TweepError as e:
                # Just exit if any error
                print("some error : " + str(e))
                break

    print ("Downloaded {0} tweets, Saved to {1}".format(tweetCount,
```

In [8]:

```
empresas =["bancolombia","grupo_nutresa","alpina","gruposura","ECOP
for x in empresas:
    busquedaTweeter(x)
```

```
Downloading max 10000000 tweets
Downloaded 100 tweets
Downloaded 200 tweets
Downloaded 300 tweets
Downloaded 400 tweets
Downloaded 500 tweets
Downloaded 600 tweets
Downloaded 700 tweets
Downloaded 800 tweets
Downloaded 900 tweets
Downloaded 1000 tweets
Downloaded 1051 tweets
No more tweets found
Downloaded 1051 tweets, Saved to bancolombiaAT.txt
Downloading max 10000000 tweets
Downloaded 53 tweets
No more tweets found
Downloaded 53 tweets, Saved to bancolombiaHASH.txt
Downloading max 10000000 tweets
Downloaded 46 tweets
```

In [ ]:

```
max_id
```

In [ ]:

```
sinceId
```

In [ ]:

```
original_searchQuery = 'alpina'  # this is what we're searching for

searchQuery = "#"+original_searchQuery
```

In [ ]:

```
searchQuery
```

In [ ]:

```

```

In [6]:

```
from GoogleNews_mod import GoogleNews_mod #modulo modificado para o
```

In [11]:

```
googlenews = GoogleNews_mod()
```

In [12]:

```
googlenews.search('Universidad Nacional')
```

In [13]:

```

#googlenews.getpage()
googlenews.searchThousand(starpage=1)
#lista = googlenews.gettext()

#googlenews.getlinks()
#for x in lista:
```

```
#     print(x)
```

```
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=0
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=0)
'NoneType' object has no attribute 'get'
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=10
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=10)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=20
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=20)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=30
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=30)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=40
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=40)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=50
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=50)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=60
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=60)
'NoneType' object has no attribute 'get'
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=70
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=70)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=80
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=80)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=90
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=90)
https://www.google.com/search?q=Universidad+Nacional&t
```

bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=100
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=100)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=110
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=110)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=120
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=120)
'NoneType' object has no attribute 'get'
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=130
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=130)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=140
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=140)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=150
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=150)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=160
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=160)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=170
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=170)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=180
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=180)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=190
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=190)
'NoneType' object has no attribute 'get'
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=200
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=200)

```
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=210
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=210)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=220
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=220)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=230

(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=230)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=240
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=240)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=250
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=250)
exit pagina 26  229  229
```

In [14]:

```
googlenews.result()
googlenews.export("unal1-1000.txt")
#googlenews.clear()
```

Codigo no usado para consultas a traves del api de google

In [ ]:

```
from googleapiclient.discovery import build
my_api_key = "AIzaSyD9bKSVP4YXkm1pKblGA__66dtgqX8uYTc"
my_cse_id = "009524471252104230700:ak9azxvojac"


def google_search(search_term, api_key, cse_id, **kwargs):
    service = build("customsearch", "v1", developerKey=api_key)
    res = service.cse().list(q=search_term, cx=cse_id, **kwargs).ex
    return res

google_search("bancolombia",my_api_key,my_cse_id)
```

In [ ]:

```

```

Anexo 2 - Creación del conjunto de datos (Twitter y Google News),

Las fuentes de información utilizadas en este trabajo son Twitter y Google News. En este documento se encuentra el programa utilizado para:

- Descargar la información de Twitter
- Descargar la información de Google News
- Importar información a base de datos DynamoDB - AWsS
- Unificar el conjunto de datos

Descarga de informacion de Twitter

Es utilizada el API de Twitter para obtener la informacion

In [47]:

In [ ]:

In [ ]:

In [3]:

In [7]:

In [22]:

Importar datset descargado de Twitter a DynamoBD - local

Previamente fue creada la tabla Twitter donde se almacenaran todos los json resultados de las peticiones al API de Twitter. Script de creacion de tabla en terminal (linea de comandos linux)
aws dynamodb --endpoint-url http://localhost:8000 (http://localhost:8000) create-table
--table-name twitter
--attribute-definitions
AttributeName=key,AttributeType=S
--key-schema AttributeName=key,KeyType=HASH
--provisioned-throughput ReadCapacityUnits=1,WriteCapacityUnits=1

In [12]:

```
Exitoso 0------------------
Exitoso 1------------------
Exitoso 2------------------
Exitoso 3------------------
Exitoso 4------------------
Exitoso 5------------------
Exitoso 6------------------
Exitoso 7------------------
Exitoso 8------------------
Exitoso 9------------------
Exitoso 10-----------------
Exitoso 11-----------------
Exitoso 12-----------------
Exitoso 13-----------------
Exitoso 14-----------------
Exitoso 15-----------------
Exitoso 16-----------------
Exitoso 17-----------------
Exitoso 18-----------------
```

In [23]:

```
28069
```

Transformar Dynambo BD y crear Pandas Dataframe

In [24]:

In [121]:

Peticiones al API de Amazon web service comprehend

Permite tener una respuesta del sentimiento detectado por AWS. Respuestas de sentimientos: POSITIVO, NEGATIVO, NEUTRO y MEZCLADO (multimples sentimientos detectados)

Documentacion del api:
https://boto3.amazonaws.com/v1/documentation/api/latest/reference/services/comprehen
(https://boto3.amazonaws.com/v1/documentation/api/latest/reference/services/comprehen

In [123]:

In [126]:

```
-----OK: 1
-----OK: 2
-----OK: 3
-----OK: 4
-----OK: 5
-----OK: 6
-----OK: 7
-----OK: 8
-----OK: 9
-----OK: 10
-----OK: 11
-----OK: 12
-----OK: 13
-----OK: 14
-----OK: 15
-----OK: 16
-----OK: 17
-----OK: 18
-----OK: 19
```

In [127]:

Out[127]:

```
[{'Sentiment': 'NEUTRAL',
  'SentimentScore': {'Positive': 0.00011496161459945142,
   'Negative': 6.106298678787425e-05,
   'Neutral': 0.9998083710670471,
   'Mixed': 1.5553940102108754e-05},
  'ResponseMetadata': {'RequestId': 'a10652dd-bdca-4fad-a03e-a0763bcfda5c',
   'HTTPStatusCode': 200,
   'HTTPHeaders': {'x-amzn-requestid': 'a10652dd-bdca-4fad-a03e-a0763bcfda5c',
    'content-type': 'application/x-amz-json-1.1',
    'content-length': '166',
    'date': 'Thu, 12 Dec 2019 22:54:34 GMT'},
   'RetryAttempts': 0},
  'id_str': '1193353981252317185'},
 {'Sentiment': 'NEUTRAL',
  'SentimentScore': {'Positive': 0.2310795933008194.
```

In [128]:

Respuestas almacenadas

In [138]:

In [149]:

In [154]:

Out[154]:

company	key_folder	
BAVARIA_OFICIAL	Top 10 Merco 10 Dic – 7 dias Colombia	98
	Top 10 Merco 11 Nov –7 dias Colombia	223
	Top 10 Merco 3 DIC –7 dias Colombia	83
	Top 10 Merco 4 Nov –7 dias Colombia	462
Cementos_Argos	Top 10 Merco 10 Dic – 7 dias Colombia	65
	Top 10 Merco 11 Nov –7 dias Colombia	74
	Top 10 Merco 3 DIC –7 dias Colombia	129
	Top 10 Merco 4 Nov –7 dias Colombia	39
ECOPETROL SA	Top 10 Merco 10 Dic – 7 dias Colombia	

In [151]:

Out[151]:

```
Index(['key', 'key_row', 'key_folder', 'key_file', 'co
mpany', 'id_str',
       'created_at', 'lang', 'in_reply_to_status_id_st
r',
       're_tweet_created_at', 're_tweet_id_str',
       're_tweet_in_reply_to_status_id_str', 'retweete
d_status', 'truncated',
       'user_id_str', 'user_location', 'user_name', 'u
ser_screen_name',
       'user_description', 'user_followers_count', 'te
xt', 'aws_sentimiento',
       'aws_score_positivo', 'aws_score_negativo', 'aw
s_score_neutro',
       'aws_score_mixed', 'manual_response'],
      dtype='object')
```

Google

In [66]:

In [67]:

In [70]:

```
grupo_nutresa1-1000.txt
alpina1-1000.txt
unal1-1000.txt
epm1-1000.txt
bavaria1-1000.txt
bancolombia1-1000.txt
nestle1-1000.txt
ecopetrol1-1000.txt
argos1-1000.txt
grupo_sura1-1000.txt
```

In [109]:

In [99]:

Out[99]:

id	title	media	date	desc	
0	Mayores ventas impulsan el beneficio del Grupo...	Portafolio.co (Comunicado de prensa) (blog)	25 oct. 2019	La utilidad neta del mayor productor de alimen...	https://www.portafolio.co/nego
1	Grupo Nutresa suma a su línea de café un negoc...	La República	18 sep. 2019	El Grupo Nutresa, líder en el mercado de alime...	https://www.larepublica.co/er
2	Grupo Nutresa: alimento para el mundo	Semana.com	25 may. 2019	Uno de ellos, el Grupo Nutresa, muestra el alt...	https://ww e
3	Así quedó el dividendo de Nutresa para este año	Dinero.com	26 mar. 2019	En la asamblea de accionistas de Nutresa se di...	https://www.dinero.com/emp
4	Grupo Nutresa cierra planta procesadora de Cun...	La República	1 sep. 2019	Alimentos Cárnicos S.A.S. aseguró que las oper...	https://www.larepublica.co/er

In [112]:

```
-----OK: 1
-----OK: 2
-----OK: 3
-----OK: 4
-----OK: 5
-----OK: 6
-----OK: 7
-----OK: 8
-----OK: 9
-----OK: 10
-----OK: 11
-----OK: 12
-----OK: 13
-----OK: 14
-----OK: 15
-----OK: 16
-----OK: 17
-----OK: 18
-----OK: 19
```

In [113]:

```
Respuestas almacenadas
```

In [124]:

In [123]:

Out[123]:

id	title	media	date	desc	
0	Mayores ventas impulsan el beneficio del Grupo...	Portafolio.co (Comunicado de prensa) (blog)	25 oct. 2019	La utilidad neta del mayor productor de alimen...	https://www.pc

1	Grupo Nutresa suma a su línea de café un negoc...	La República	18 sep. 2019	El Grupo Nutresa, líder en el mercado de alime...	https://www.la
2	Grupo Nutresa: alimento para el mundo	Semana.com	25 may. 2019	Uno de ellos, el Grupo Nutresa, muestra el alt...	
3	Así quedó el dividendo de Nutresa para este año	Dinero.com	26 mar. 2019	En la asamblea de accionistas de Nutresa se di...	https://www.c
4	Grupo Nutresa cierra planta procesadora de Cun...	La República	1 sep. 2019	Alimentos Cárnicos S.A.S. aseguró que las oper...	https://www.la
...	...	...	...	...	
1945	Grupo Sura avanza en el Índice de Sostenibilid...	Estrategia y Negocios	2 oct. 2018	Grupo SURA es parte de las 317 empresas selecc...	https://www.est
1946	Grupo Sura actualizó su direccionamiento estra...	Valora Inversiones (blog)	23 mar. 2018	El Grupo de Inversiones Suramericana (Grupo Su...	https://www.v
1947	En rebalanceo del Hcolsel de septiembre no se ...	valoraanalitik.com	19 sep. 2019	“En una menor proporción, resaltan las compras...	https://www.'
1948	El aprendizaje fue el reto en la implementació...	Dinero.com	10 jun. 2016	Desaprender para volver a aprender fue	https://www.c

					el reto...	
1949	Suramericana SA cerró 2018 con un patrimonio n...	La República	29 ene. 2019	La filial de Grupo Sura, Suramericana S.A., di...		https://www.

1950 rows × 19 columns

In [118]:

Out[118]:

```
Index(['title', 'media', 'date', 'desc', 'link', 'img'
, 'company', 'full_text',
       'full_text_cln'],
      dtype='object')
```

In [111]:

Out[111]:


```
[{'status': {'code': '0',
   'msg': 'OK',
   'credits': '1',
   'remaining_credits': '19980'},
  'model': 'general_es',
  'score_tag': 'P',
  'agreement': 'AGREEMENT',
  'subjectivity': 'OBJECTIVE',
  'confidence': '100',
  'irony': 'NONIRONIC',
  'sentence_list': [{'text': 'Mayores ventas impulsan
el beneficio del Grupo Nutresa a ....',
    'inip': '0',
    'endp': '60',
    'bop': 'y',
    'confidence': '100',
    'score_tag': 'P',
    'agreement': 'AGREEMENT',
```


Twitter

In [1]:

In [2]:

In [5]:

In [ ]:

In [ ]:

In [ ]:

In [7]:

In [8]:

```
Downloading max 10000000 tweets
Downloaded 100 tweets
Downloaded 200 tweets
Downloaded 300 tweets
Downloaded 400 tweets
Downloaded 500 tweets
Downloaded 600 tweets
Downloaded 700 tweets
Downloaded 800 tweets
Downloaded 900 tweets
Downloaded 1000 tweets
Downloaded 1051 tweets
No more tweets found
Downloaded 1051 tweets, Saved to bancolombiaAT.txt
Downloading max 10000000 tweets
Downloaded 53 tweets
No more tweets found
Downloaded 53 tweets, Saved to bancolombiaHASH.txt
Downloading max 10000000 tweets
Downloaded 46 tweets
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [6]:

In [11]:

In [12]:

In [13]:

```
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=0
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=0)
'NoneType' object has no attribute 'get'
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=10
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=10)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=20
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=20)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=30
(https://www.google.com/search?q=Universidad+Nacional&
tbm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=30)
https://www.google.com/search?q=Universidad+Nacional&t
bm=nws&tbs=lr:lang_1es&lr=lang_es&gl=CO&start=40
(https://www.google.com/search?q=Universidad+Nacional&
```

In [14]:

Codigo no usado para consultas a traves del api de google

In [ ]:

```

```

In [ ]:

```

```