import os

import pandas as pd

def Preprocess\_Tweets(data):

data['Text\_Cleaned'] = data['Text'].str.lower()

## FIX HYPERLINKS

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'https?:\/\/.\*[\r\n]\*', ' ',regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'www.\*[\r\n]\*', ' ',regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('https', '', regex=False)

## FIX INDIVIDUAL SYMBOLS

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(': ', ' ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(', ', ' ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('. ', ' ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('[;\n~]', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace("[]'â€¦\*™|]", '', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('[[()!?"]', '', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('\_', '', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('w/', ' with ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('f/', ' for ', regex=False)

## FIX EMOJIS

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(':)', '', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(':-)', '', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(':(', '', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(':-(', '', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('0\_o', '', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(';)', '', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('=^.^=', '', regex=False)

## FIX % SYMBOL

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('%', ' percent ', regex=False)

## FIX & SYMBOL

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' & ', ' and ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('&amp', ' and ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('&gt', ' greater than ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('cup&handle', 'cup and handle', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('c&h', 'cup and handle', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('head&shoulders', 'head and shoulders', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('h&s', 'head and shoulders', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('point&figure', 'point and figure', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('p&f', 'point and figure', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('s&p', 'SP500', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('q&a', 'question and answer', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('&', ' and ', regex=False)

## FIX USER TAGS AND HASTAGS

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('@[a-z0-9]+', '', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('#[a-z0-9]+', '', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('@', '', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('#', '', regex=False)

## FIX EMBEDDED COMMAS AND PERIODS

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([a-z]),([a-z])', r'\1 \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9]),([0-9])', r'\1\2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])[+]+', r'\1 ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(',', '', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('u.s.', ' us ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('\.{2,}', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([a-z])\.([a-z])', r'\1 \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('pdating', 'updating', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([a-z])\.', r'\1 ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'\.([a-z])', r' \1', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' . ', ' ', regex=False)

## FIX + SYMBOL

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'[+]([0-9])', r'positive \1', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('c+h', 'cup and handle', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('h+s', 'head and shoulders', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('cup+handle', 'cup and handle', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' + ', ' and ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('+ ', ' ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([a-z])[+]([a-z])', r'\1 and \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('+', '', regex=False)

## FIX - SYMBOL

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([a-z])[-]+([a-z])', r'\1 \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([a-z]) - ([a-z])', r'\1 to \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9]) -([0-9\.])', r'\1 to \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r' [-]([0-9])', r' negative \1', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])-([0-9\.])', r'\1 to \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9]) - ([0-9\.])', r'\1 to \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9a-z])-([0-9a-z])', r'\1 \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('[-]+[>]', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' [-]+ ', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('-', ' ', regex=False)

## FIX $ SYMBOL

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('[$][0-9\.]', ' dollars ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('$', '', regex=False)

## FIX = SYMBOL

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('=', ' equals ', regex=False)

## FIX / SYMBOL

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('b/c', ' because ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('b/out', ' break out ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('b/o', ' break out ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('p/e', ' pe ratio ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' [/]+ ', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' 1/2 ', ' .5 ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' 1/4 ', ' .25 ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' 3/4 ', ' .75 ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' 1/3 ', ' .3 ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' 2/3 ', ' .6 ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('[/]{2,}', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([a-z])/([a-z])', r'\1 and \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('[0-9]+/[0-9]+/[0-9]+', '', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9]{3,})/([0-9\.]{2,})', r'\1 to \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9]{2,})/([0-9\.]{3,})', r'\1 to \2', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('[a-z0-9]+/[a-z0-9]+', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('/', '', regex=False)

## FIX < > SYMBOLS

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('[<]+ ', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('<', ' less than ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' [>]+', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('>', ' greater than ', regex=False)

## FIX : SYMBOL

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('[0-9]+:[0-9]+am', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('[0-9]+:[0-9]', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(':', ' ', regex=False)

## FIX UNITS

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('user ', ' ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9]+)dma', r'\1 displaced moving average ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'dma([0-9]+)', r'\1 displaced moving average ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9]+)sma', r'\1 simple moving average ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'sma([0-9]+)', r'\1 simple moving average ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9]+)ema', r'\1 expontential moving average ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'ema([0-9]+)', r'\1 expontential moving average ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9]+)ma', r'\1 moving average ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'ma([0-9]+)', r'\1 moving average ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])mos', r'\1 months ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])minute', r'\1 minute ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])minutes', r'\1 minutes ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])min', r'\1 minute ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])mins', r'\1 minutes ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])day', r'\1 day ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])days', r'\1 days ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])wk', r'\1 week ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' wk ', ' week ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' wknd ', ' weekend ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])wks', r'\1 weeks ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])hours', r'\1 hours ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])hour', r'\1 hour ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])yr', r'\1 year ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])yrs', r'\1 years ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' yr', ' year ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])am', r'\1 am ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])pm', r'\1 pm ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])est', r'\1 ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])ish', r'\1 ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9 ])pts', r'\1 points ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])x', r'\1 times ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])th', r'\1 ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])rd', r'\1 ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])st', r'\1 ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])nd', r'\1 ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('mrkt', 'market', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' vol ', ' volume ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' ptrend', ' positive trend ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' ppl', ' people ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' pts', ' points ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' pt', ' point ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' l(ol){1,}', ' laugh ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('imho', ' in my opinion ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace('prev ', 'previous ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' 1q', ' first quarter ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' 2q', ' second quarter ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' 3q', ' third quarter ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' 4q', ' fourth quarter ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' q1', ' first quarter ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' q2', ' second quarter ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' q3', ' third quarter ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' q4', ' fourth quarter ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' 10q ', ' form 10 ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])million', r'\1 million ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])mil', r'\1 million ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' mil ', ' million ', regex=False)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])billion', r'\1 billion ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])cents', r'\1 cents ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])3d', r'\1 3 dimensional ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])gb', r'\1 3 gigabytes ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])c', r'\1 calls ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])y', r'\1 year ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])p', r'\1 puts ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])d', r'\1 days ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])h', r'\1 hour ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])s', r'\1 ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])k1', r'\1 thousand ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])k', r'\1 thousand ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])m', r'\1 million ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])b', r'\1 billion ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].replace(r'([0-9])([a-z])', r'\1 \2', regex=True)

## FIX EXTRA SPACES AND ENDING PUNCTUATION

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.replace(' +', ' ', regex=True)

data['Text\_Cleaned'] = data['Text\_Cleaned'].str.strip(' .!?,)(:-')

return data