import nltk

import string

import csv

import io

from nltk.corpus import stopwords

from nltk.stem import SnowballStemmer

from nltk.stem import WordNetLemmatizer

from nltk.tokenize import sent\_tokenize, word\_tokenize

snowball\_stemmer = SnowballStemmer('english')

lemmatizer = WordNetLemmatizer()

stop\_words = set(stopwords.words('english'))

contracts = ["8Contracts (Test).txt"]

risky\_sentences = []

safe\_sentences = []

all\_words = []

headers = []

new\_List = [all\_words]

def classifySentences():

for n in range(len(contracts)):

f = open(contracts[n])

raw = f.read()

sent\_tokenize\_list = sent\_tokenize(raw)

for x in range(len(sent\_tokenize\_list)):

is\_risky = False

words = word\_tokenize(sent\_tokenize\_list[x])

for y in range(len(words)):

all\_words.append(words[y])

if "0cf333" in words[y]:

is\_risky = True

if is\_risky == True:

risky\_sentences.append(words)

else:

safe\_sentences.append(words)

def processWords(input\_data):

single\_word\_list = [] #This will store each word in a single cell

unique\_words = [] #This list stores all the unique words, excluding duplicates

filter\_stemming = [] #List for Stemmer (using the porter stemmer algorithm)

filter\_lemmatizer = [] #List for Lemmatizer

final\_stem\_unique = [] #Unique Stem List

print(input\_data)

# Removes punctuation and numbers

for celem in input\_data:

try:

nopunc = celem.translate(str.maketrans('','',string.punctuation))

unique\_words.append(nopunc.strip('0123456789•–“”')) #also removes numbers ranging from 0-9

except:

print("couldn't remove punctuation & numbers")

print(unique\_words)

#Removes stop words

filter\_stop\_list = unique\_words[:]

for elemw in unique\_words:

if elemw in stopwords.words('english'):

filter\_stop\_list.remove(elemw)

filter\_nostopwords = [elemxx for elemxx in filter\_stop\_list if elemxx]

final\_filter = [elemx for elemx in unique\_words if elemx]

for fn in filter\_nostopwords:

filter\_stemming.append(snowball\_stemmer.stem(fn))

for fl in filter\_nostopwords:

filter\_lemmatizer.append(lemmatizer.lemmatize(fl))

#Removing duplicates in stemming

for elem in filter\_stemming:

if elem not in final\_stem\_unique:

final\_stem\_unique.append(elem)

#Removing all single words in the document

excludeWords = [‘cf’,’b’,’c’ , ‘e’,’f’,’g’,’’’,’h’,’ii’,’$’,’wo’,’nt’,’ip’,’’’,’gsm’,’iii’,’a’,’sl’,’dx’,’isl’,’p’,’k’,’l’,’n’,’eg’,’st’,’uptod’,’ait’,’gb’,’iv’,’v’,’vi’,’ wwwgiffgaffsimcardscouk’,’ examplecomgiffgaff’,’ giffgaffexamplecom’,’vii’,’viii’,’x’,’xi’,’xii’, ‘t’,’cs’,’rd’,’I’,’ wwwicoorguk’,’eea’,’ec4a’,’bd’,’pm’,’dpo’,’s’,’o’,’iab’,’ga’,’ck’,’.’,’ag’,’ukon’,’myvm’,’jr’,’sa’,’bb’ ]

for excludeWord in excludeWords:

if excludeWord in final\_stem\_unique:

final\_stem\_unique.remove(excludeWord);

#Places words in a list in csv files ad a excluded generated file.

return final\_stem\_unique

all\_words.sort()

#Puts list in a csv file as an excl generated file

with open('Final Data Collection.csv', 'w') as myfile:

classifySentences();

headers = processWords(all\_words)

wr = csv.writer(myfile, dialect='excel')

wr.writerow(headers)

print("File Successfully saved.")

for sentence in risky\_sentences:

processed\_sentence = processWords(sentence)

attributes\_present = []

for attribute in headers:

if attribute in processed\_sentence:

attributes\_present.append(1)

else:

attributes\_present.append(0)

wr.writerow(attributes\_present)

for sentence in safe\_sentences:

processed\_sentence = processWords(sentence)

attributes\_present = []

for attribute in headers:

if attribute in processed\_sentence:

attributes\_present.append(1)

else:

attributes\_present.append(0)

wr.writerow(attributes\_present)