

**School of Computer Science and Engineering**

**Register Number: 18BCE0745**

**Name: Gourishetty Sreemanth**

**Code**

# -\*- coding: utf-8 -\*-

"""

Created on Thu Jul 30 16:07:35 2020

@author: Sreemanth

"""

import nltk

#nltk.download('wordnet')

#nltk.download('averaged\_perceptron\_tagger')

from nltk.corpus import stopwords

from nltk.stem import PorterStemmer

from nltk.stem import WordNetLemmatizer

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

import pandas as pd

import re

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

print("#####COMMON WORDS -1")

words1 = []

#Group in a list the words common for two text files and show their total count

f1 = open("Artificaial intelligence.txt").readlines()

f2 = open("machine learning.txt").readlines()

if len(f1) != 0 | len(f2) != 0:

uniq1 = set(words for line in f1 for words in line.strip().split())

uniq2 = set(wordss for lines in f2 for wordss in lines.strip().split())

for words in uniq1:

for wordds in uniq2:

if words == wordds:

words1.append(words);

words1 = [w for w in words1 if not w in stop\_words]

print(len(words1))

with open('index.txt', 'w') as f:

for item in words1:

f.write("%s\n" % item)

readwords = []

# opening the text file

with open('index.txt','r') as file:

# reading each line

for line in file:

# reading each word

for word in line.split():

# displaying the words

readwords.append(word)

ps = PorterStemmer()

lemmatizer = WordNetLemmatizer()

stems = []

lemma = []

for w in readwords:

print(ps.stem(w), " - ", lemmatizer.lemmatize(w))

stems.append(ps.stem(w))

lemma.append(lemmatizer.lemmatize(w))

frequency1 = {}

for word in stems:

count = frequency1.get(word,0)

frequency1[word] = count + 1

frequency\_list1 = frequency1.keys()

print(len(frequency\_list1))

frequency2 = {}

for word in lemma:

count = frequency2.get(word,0)

frequency2[word] = count + 1

frequency\_list2 = frequency2.keys()

print(len(frequency\_list2))

if(len(frequency\_list1) <= len(frequency\_list2)):

with open('index.txt', 'w') as f:

for item in stems:

f.write("%s\n" % item)

import os

if(len(frequency\_list1) > len(frequency\_list2)):

print("hello")

with open('index.txt', 'w') as f:

for item in lemma:

f.write("%s\n" % item)

os.rename('index.txt', 'final-index.txt')

finalwords = []

# opening the text file

with open('index.txt','r') as file:

# reading each line

for line in file:

# reading each word

for word in line.split():

# displaying the words

finalwords.append(word)

tagged = nltk.pos\_tag(finalwords)

print(tagged)

df = pd.DataFrame(tagged)

print(df)

**Output**







