import PyPDF2

import string

from nltk.corpus import stopwords

import nltk

import pandas as pd

import re

obj=open('JavaBasics-notes.pdf','rb')

PDFreader=PyPDF2.PdfFileReader(obj)

n=PDFreader.numPages

content=''

for i in range(n):

content=content+PDFreader.getPage(i).extractText()

no\_punch=[]

for i in content:

if i not in string.punctuation:

no\_punch.append(i)

final\_clean=''.join(no\_punch)

clean\_content=[]

for word in final\_clean.split():

if word.lower() not in stopwords.words('english'):

clean\_content.append(word.lower())

nlp\_words=nltk.FreqDist(clean\_content)

word=[]

#word=list(word)

for j in clean\_content:

if j not in word:

word.append(j)

freq=[]

for i in word:

freq.append(nlp\_words.get(i))

df=pd.DataFrame({'freq':freq},index=word)

df.to\_excel('with\_numerics.xlsx',sheet\_name='sheet 1',index=True)

"""if the above data is to be represented without numerics """

tok=re.sub('[^a-zA-Z,]',' ',final\_clean)

tok=tok.lower()

clean\_content2=[]

for word2 in tok.split():

if word2 not in stopwords.words('english'):

clean\_content2.append(word2)

nlp\_words2=nltk.FreqDist(clean\_content2)

word2=[]

#word=list(word)

for j2 in clean\_content2:

if j2 not in word2:

word2.append(j2)

freq2=[]

for i2 in word2:

freq2.append(nlp\_words2.get(i2))

df2=pd.DataFrame({'freq':freq2},index=word2)

df2.to\_excel('without\_numerics.xlsx',sheet\_name='sheet 1',index=True)