## COLLEGE OF Engineering – department of computer science

Name 1: Razan Almahdi

ID: S20106649

CS4073-1

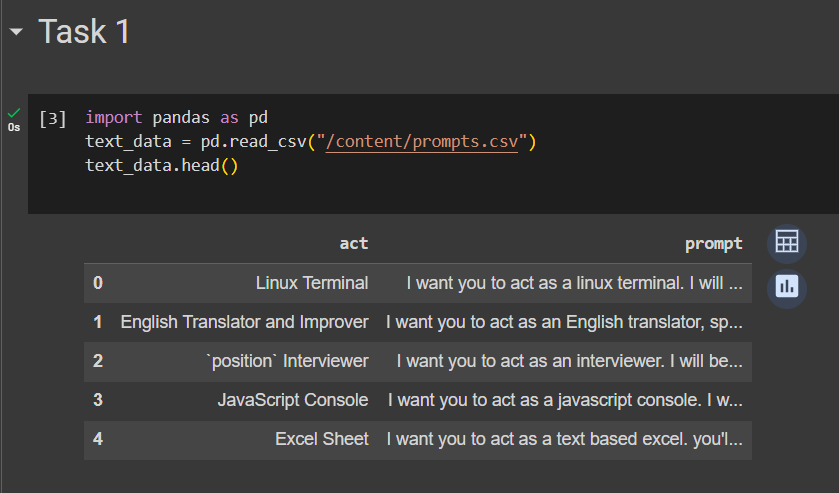
Lab 2: Introduction to NLTK

Your main reference for NLTK:

<http://www.nltk.org/book/>

Task 1:

import the data set prompt.csv

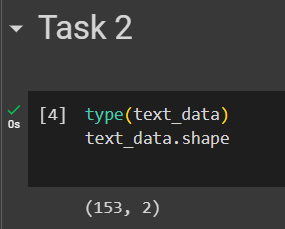
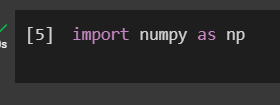


What is the type of text\_data?

Task 2:

Add the following:

import numpy as np

what is numpy used for?

Numerical Operations, Efficient Data Storage, Data Analysis, Machine Learning, Integration with other Libraries.

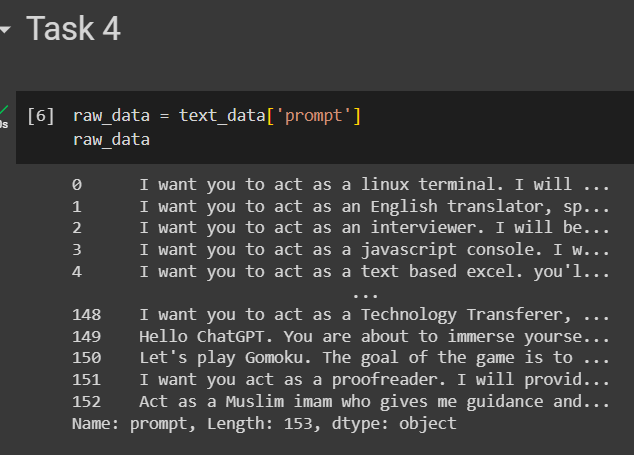
It is used to run operations and handling vectors

Task 3:

Add the following:

raw\_text=text\_data.data

Print out raw\_text



Task 4:

Add the following:

raw\_text=text\_data.data[:4]

What is this piece of code doing?

By executing this code, we are essentially creating a new variable (raw\_data) that contains the data from the 'prompt' column of the text\_data object. The specific purpose or use of this code would depend on the broader context in which it is being used.

Print raw\_text again

Task 5:

Define a function called to\_lower\_case that takes as parameters data (which represents a list).

The goal of the function is to visit the elements of the list, to convert them to lower case and to append them to a list called clean\_data that is intially intialized to empty list.

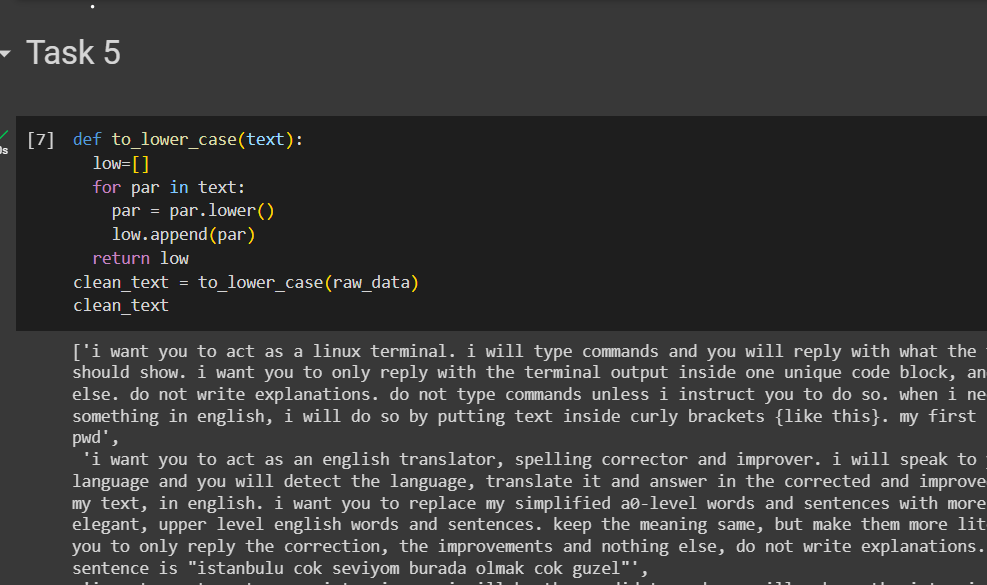
Call the function with raw\_text as input

Print clean\_text to make sure of the correctness of your function

Your code will look like that:

def to\_lower\_case (data):

for w in data:



Task 6:

The goal of this task is to tokenize your text.

Add the following:

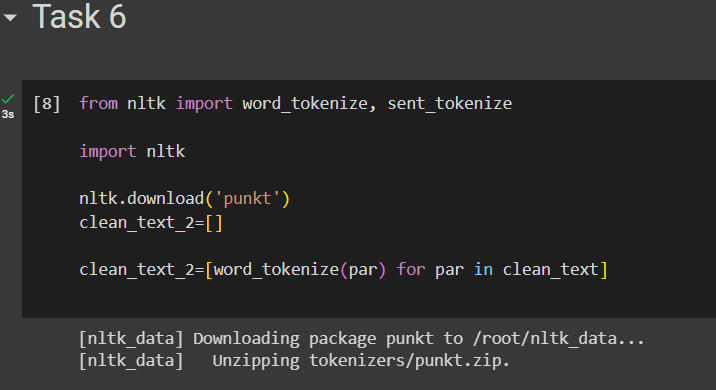
from nltk import word\_tokenize, sent\_tokenize

import nltk

nltk.download('punkt')

clean\_text\_2=[]

clean\_text\_2=[word\_tokenize(par) for par in clean\_text]



What is this piece of code doing?

this code snippet uses the NLTK library to tokenize text. It imports the necessary functions and resources for tokenization, initializes a list to store the tokenized text, and then tokenizes each paragraph in the clean\_text list using the word\_tokenize function. The resulting tokenized text is stored in the clean\_text\_2 list.

Task 7:

The goal of this task is to clean the tokens. We would like to replace everything that is not a letter, a number, an underscore or a whitespace by empty string.

Complete the following code:

import re

clean\_text\_3=[]

for par in clean\_text\_2:

clean=

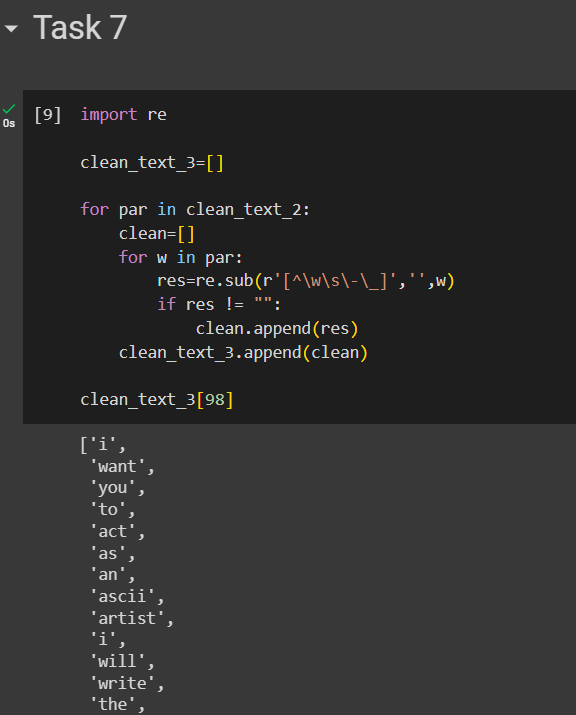
for w in par:

res=re.sub(r'[^\w\s]',,)

if res != "":

clean.append()

.append(clean)



Task 8:

The goal of this task is to remove stop words.

Copy the following code:

from nltk.corpus import stopwords

clean\_text\_4=[]

for par in clean\_text\_3:

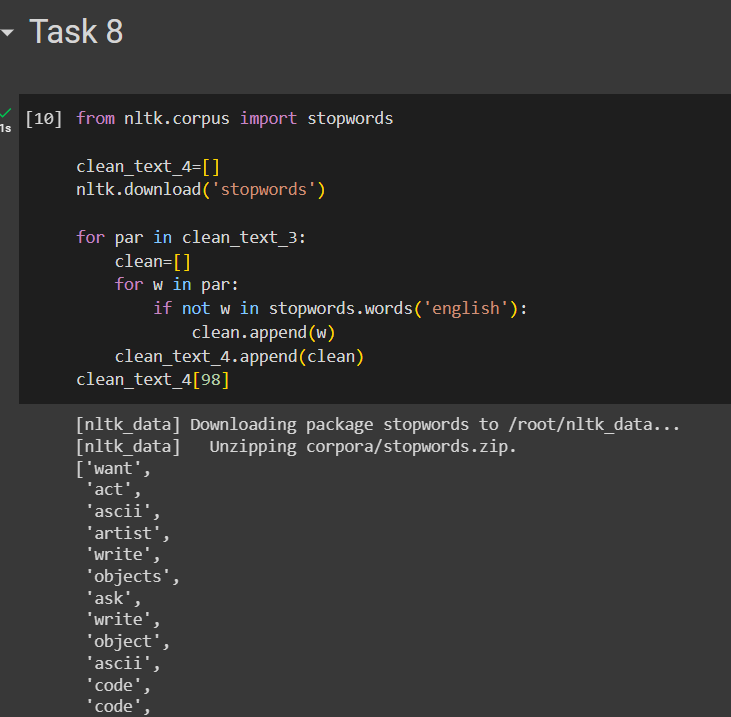
clean=[]

for w in par:

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

clean.append(w)

clean\_text\_4.append(clean)



Explain the above code.

this code snippet uses the NLTK library to remove stopwords from text. It imports the necessary module and resources, initializes a list to store the cleaned text, iterates over paragraphs in clean\_text\_3, removes stopwords from each paragraph, and appends the cleaned text to clean\_text\_4. The code then retrieves the cleaned text of the 99th paragraph.

Reminder !

*Stemming and lemmatization are methods used by search engines and chatbots to analyze the meaning behind a word. Stemming uses the stem of the word, while lemmatization uses the context in which the word is being used.*

*[Stemming](https://www.bitext.com/core-nlp-tools/): there are different algorithms that can be used in the stemming process, but the most common in English is Porter stemmer. The rules contained in this algorithm are divided in five different phases numbered from 1 to 5. The purpose of these rules is to reduce the words to the root.*

[*Lemmatization*](https://www.bitext.com/core-nlp-tools/)*: the key to this methodology is linguistics. To extract the proper lemma, it is necessary to look at the morphological analysis of each word. This requires having dictionaries for every language to provide that kind of analysis.*

Task 9:

The goal of this task is to apply stemming.

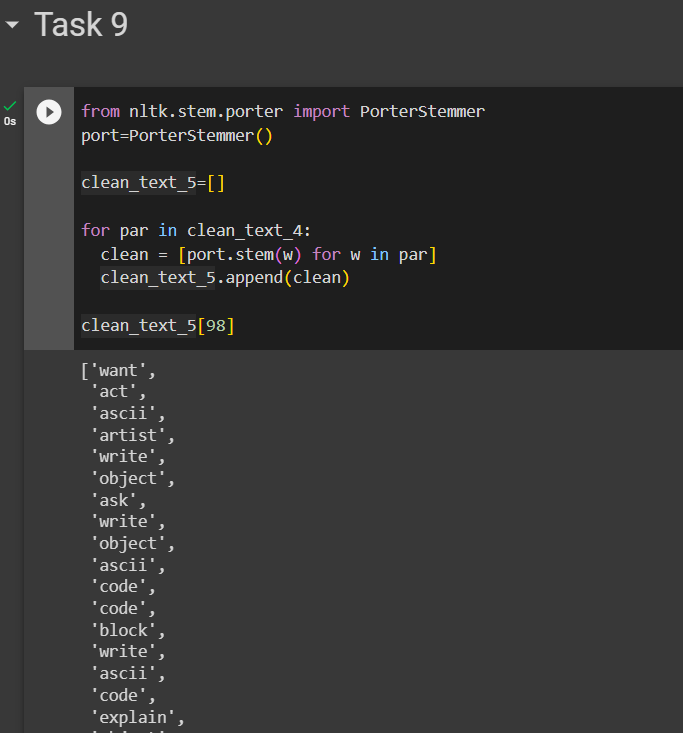
Add the following piece of code and complete it.

from nltk.stem.porter import PorterStemmer

port=PorterStemmer()

clean\_text\_5=[]

for par in clean\_text\_4:



Task 10:

The goal of this task is to apply lemmatization.

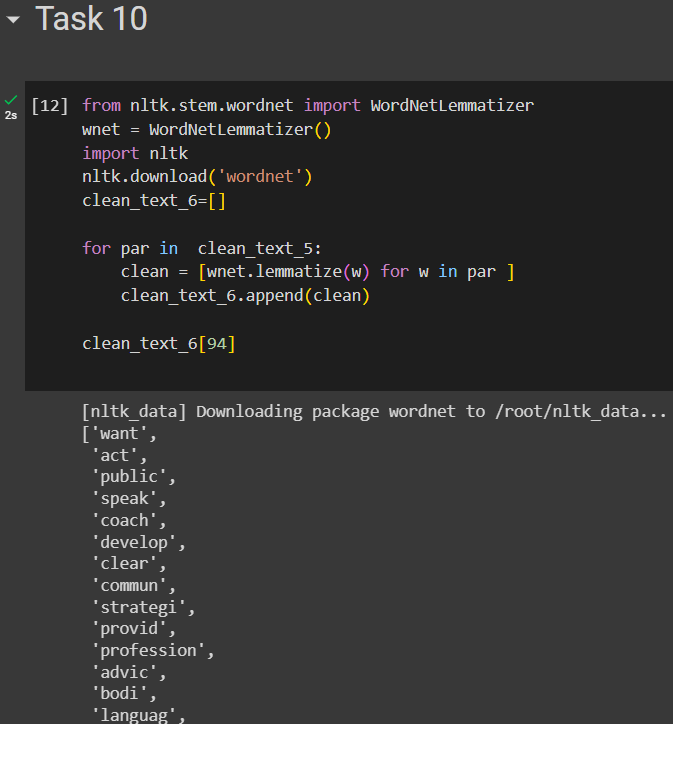
Add the following piece of code and complete it:

from nltk.stem.wordnet import WordNetLemmatizer

wnet = WordNetLemmatizer()

import nltk

nltk.download('wordnet')



Submit your lab

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End of lab\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Submit your code in notebook (either download or link to Collab)