Homework 1

You can submit the homework as *.py or *.ipynb file. Answers to theoretical questions can be written in a separate WORD document.

1.) Setup a python environment on your computer (Pycharm / google colab).
2.) Install the following libraries: pytorch, NLTK, spaCy
3.) Write a program that loads the file spam.csv (source: KAGGLE)
4.) Compute simple statistics on the data: number of SMS message, number of spams, total word count, averaged number of words per message, 5 most frequent words in the file, number of rare words (appear only once in the file).
5.) Perform tokenization twice: by using NLTK library and by Spacy library. Write what is the difference between the results.
6.) Perform lemmatization by NLTK and by Spacy libraries. Write what is the difference between the results.
7.) Perform stemming by NLTK twice on both results of (6). Write what is the difference between the results.
8.) Identify a spam message where its removal from the spam.csv dataset would:
    a) Reduce the total number of stemmed tokens
    b) Maintain the exact same number of lemmatized tokens

    If no such message exists, provide a detailed explanation of why this scenario is impossible or improbable.

9.) Identify a spam message where its removal from the spam.csv dataset would:
    a) Reduce the total number of lemmatized tokens
    b) Maintain the exact same number of stemmed tokens

    If no such message exists, provide a detailed explanation of why this scenario is impossible or improbable.

10.) Compare the results of NLTK before and after tokenization, lemmatization and stemming
11.) Compare the results of Spacy before and after tokenization and lemmatization.
12.) Install the library BeautifulSoup
13.) Choose any url page. You can choose a page of yourself in a social media (e.g., facebook, Instagram, Linkdin, Twitter), or anything else
14.) Perform data scrapping of this url. You can learn how to do it from the following links:
    https://www.edureka.co/blog/web-scraping-with-python/
    https://zindi.africa/learn/a-beginners-guide-to-scraping-data-from-social-media
    or, ask CHATGPT to program it for you.
15.) Perform steps 5-11 on the text achieved in (14).
16.) Download one of your whatsapp chats into *.txt file (at least 50 messages in Hebrew).
17.) Repeat steps 5-11 on the text file achieved in (16)