Natural Language Processing
Homework nr. 1

Task 1. Reading the .txt file.

Download the folder with data from here: https://drive.google.com/drive/folders/1QiTxxzmNGqqxsEmyVwdXa-1lTbVmtHZj?usp=sharing

Open with the python open function the file "*class_11_biology_chapter_1_0.txt*" from the "*TEST/biology*" folder. Extract the full content of the file in one unique string at the lower case.

NOTE: The following tasks should be done on the string in the lower case.

Task 2. Obtain the file metadata.

Get the following metadata about the content of the file.

1. The number of sentences extracted by **sent_tokenize** from nltk.
2. The number of unique tokens obtained with **word_tokenize** from nltk.
3. The number of unique tokens obtained with **casual_tokenize** from nltk.
4. The number of unique tokens obtained with **MWETokenizer** from nltk.
5. The mean number of words (tokens) per sentence in the text by every tokenizer.

HINT: Save the unique tokens obtained by every tokenizer in a set you will need it for the next task.

Task 3. Comparing the Stemmers.

Get the union of all unique tokens found by every tokenizer. Use this new created set to find the stemmed form of this token generated by every of the following stemmers - **PorterStemmer**, **LancasterStemmer** and **SnowballStemmer**. Save the result in a pandas DataFrame of the following form.

original_token	porter	Lancaster	Snowball
token	token	token	token

Compare the result and differences of every stemmer.

Task 4. Frequencies of words.

Using the lower case string obtained from the file, get the frequencies of the word. What are the 10 most frequent tokens?

Extract the list of the hapaxes.

Plot the frequency of the words in the descending order as a scatter plot, How similar it is to the zipf law? Why?

Task 5. Part of speech tagging.

Extract from the text the part of speech of every token, and group them in dictionary with the following form:

{“POS” : [list of tokens with this POS] }

What is the most common part of speech and why? What is the part of speech that has the smallest frequency?

Task 1. Citirea fișierului .txt.

Descarcă mapa cu date de aici: https://drive.google.com/drive/folders/1QiTxxzmNGqqxsEmyVwdXa-1lTbVmtHZj?usp=sharing

Deschide cu funcția open din python fișierul "*class_11_biology_chapter_1_0.txt*" din folderul "*TEST/biology*". Extrage tot conținutul fișierului într-un string unic la lower case.
NOTE: Următoarele task-uri trebuie realizate cu string-ul dat în lower case.

Task 2. Obține metadatele fișierului.

Obține următoarele metadate din conținutul fișierului:

1. Numărul de propoziții extrase de **sent_tokenize** din nltk.
2. Numărul de tokenuri unice obținute cu **word_tokenize** din nltk.
3. Numărul de tokenuri unice obținute cu **casual_tokenize** din nltk.
4. Numărul de tokenuri unice obținute cu **MWETokenizer** din nltk.
5. Numărul mediu de cuvinte (tokenuri) pe propoziție în text pentru fiecare propoziție.

HINT: Salvează setul de tokenuri unice obținute de către fiecare tokenizator într-un set, vei avea nevoie de ele în task-ul următor.

Task 3. Compararea stemmerilor.

Obține uninea tuturor seturilor cu tokenuri unice obținute de fiecare tokenizator. Cu acest nou creat set găsește stemmed for a acestui token obținută de fiecare din următorii stemmeri - **PorterStemmer**, **LancasterStemmer** și **SnowballStemmer**. Salvează rezultatul într-un pandas DataFrame în următoarea formă.

original_token	porter	Lancaster	Snowball
token	token	token	token

Compară rezultatul și diferențele dintre fiecare stemmer.

Task 4. Frecvența cuvintelor.

Utilizând stringul obținut din fișier, obține frecvențele fiecărui cuvânt. Care sunt cele mai întâlnite 10 cuvinte.

Extrage liste the hapaxe.

Creează un un grafic scatter plot în ordinea descendentă a frecvenței cuvintelor în text. Cât de tare urmează aceste legea lui zipf? De ce?

Task 5. Extragerea părților de vorbire.

Extrage din text partea de vorbire a fiecărui token, și grupeaza-le într-un dicționar cu următoarea formă.

{“POS” : [lista tokenurilor cu aceasta parte de vorbire]

Care este cea mai întâlnită parte de vorbire și de ce? Care este partea/părțile de vorbire care are cea mai mică frecvență?