Exercise 9

Programs should be written in the programming language Python according to the respective exercise description. The program must be correct in terms of syntax and semantic. If there exists a minimal solution calling only a single pre-defined function, this function is not allowed.

The weekly exercise should be uploaded on ILIAS as a single Jupyter Notebook (.ipynb). Processing time of each exercise runs from Wednesday, 8:00 until following Tuesday, 8:00, if not communicated differently. Make sure that answers to questions are contained within the Jupyter Notebook that is uploaded on ILIAS.

1. The file `3000words.txt`[1] contains a list of 3000 most frequent english words. Write a function `load_wordset(filename)` that reads the file and stores all words in a python set and returns the set.

Example:

```
wordset = load_wordset("3000words.txt")
print(wordset)
```

Example output:

```
{'EVERYBODY', 'GLANCE', 'DISTINGUISH', 'THICK', 'SCOPE', ...}
```

2. Write a program that loads the wordset of file `3000words.txt` and creates a histogram of word lengths. Add proper diagram title, axis labels and legends and use your preferred colors.

3. Write a function `get_words_of_len(wordset, wordlen)`, that takes a word set and a word length as arguments and returns a list of words in `wordset` with length `wordlen`.

Example:

```
wordlist = get_words_of_size(wordset, 2)
print(wordlist)
```

Example output:

```
['TO', 'NO', 'ME', 'BY', 'WE', 'MR', 'BE', 'HI', 'IT', 'ON', 'DO', 'MY', ...]
```

4. We want to know, if the relative frequency of shared words (k-tuples) of length $k = 2$ is a reasonable approximation of the relative pairwise identity between two protein sequences. Therefore we calculated for several pairs of homologous protein sequences from different species the relative number of shared k-tuples and the relative global pairwise identity. The results can be found in file `ktup_approximation.txt`. Create a scatter plot visualizing the data. Add proper title, axis lables, legends and render the markers in your preferred colors and style.

5. In exercise `ExerciseW7.pdf` you worked with the file `bacteria_abundances_W7.txt`. Create a bar diagram showing the average „abundance“ of each of the bacterial species/genera grouped by patients with and without Chron's disease. Add proper diagram title, axis labels and legends and render the bars in your preferred colors.

6. Based on task 5, try to add standard deviation as error bar for each of the bars in the diagram.

[1] https://www.ef.de/englisch-hilfen/englische-vokabellisten/3000-worter