**Lab 2: Text comparison II**

**Tasks to do:**

1. Open the file lab2\_ex1.py and read the code.
2. Open and read 3 files 'DB.txt', 'HP\_small.txt','Tolkien.txt'
3. Divide the text 'Tolkien.txt' into 3 parts.
4. For building a dictionary concatenate the 3 files to **allFilesStr**.
5. For text comparison, build the dictionary from array **allFilesStr.**
6. For the all 5 text parts built in sections 2 and 3 generate a frequency matrix **wordFrequency** according to the dictionary.
7. Decrease the frequency matrix size by building a new frequency matrix **wordFrequency1** with the condition: **total word frequency**>20.

**Independent work:**

1. Analyze the texts similarity using distance matrices **dist** (built from **wordFrequency**) and **dist1** (to be built from **wordFrequency1**)**.** Write results of analysis with explanations in **lab2\_result.docx**
   1. What is the meaning of the distance matrix results?
   2. Which distance matrix gives better results?
2. Analyze the texts similarity of the texts 'Eliot.txt' divided into 3 parts and 'Tolkien.txt' divided into 3 parts, and 'HP\_small.txt' divided into 2 parts.
3. Decrease the frequency matrix size by building a new frequency matrix **wordFrequency2** with the condition: **total word frequency**>40. Analyze and explain the new results.
4. Write the results of the analysis and add explanations in **lab2\_result.docx.**

Format of submission

Submission contains .py files with necessary code and .docx file with the students names and IDs, the results and explanations which has to be archived in .zip file with the following name:

lab<#>\_<IDnumber1>\_<IDnumber2>, where # is a lab number.