Format of submission

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

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

**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 dictionary concatenate the all files to **allFilesStr**.
5. For text comparison, build the dictionary from array **allFilesStr.**
6. For 5 text parts build the frequency matrix **wordFrequency** in according to the dictionary.

Independent work:

1. To decrease the frequency matrix dimension build the new frequency matrix **wordFrequency2** in according to the condition (**sumArray**>20). Present results in in the Word file **result.docx**.
2. Analyze the texts similarity using distance matrices **dist** (built from **wordFrequency**) and **dist2**(built from **wordFrequency2**)**.** Write results of analysis in **result.docx** (What is the meaning of distance matrix values? Which distance matrix gives better results?) .
3. Analyze the texts similarity of the texts 'Eliot.txt'(divide into 2 parts) and 'Tolkien.txt' (4 parts). Write results of analysis in **result.docx.**