Submission Format:

Please Submit one **ZIP FILE** that contains:

* **\*.py** files with necessary code and
* **\*.docx** file with IDs and students names, results and explanations.

The name of the ZIP FILE should be: lab<#>\_<IDnumber1>\_<IDnumber2>, where # is the lab number.

**Lab 4: Clustering and Text comparison**

**Tasks to do:**

1. Reminder: Use the files lab3\_ex011.py and lab3\_ex012.py as example.

**Independent work:**

1. Open and read the file 'text1.txt' containing ***k*** books.
2. Preprocess the data from the given file.
3. Cluster the data using ***k*-means algorithm**. Specify ***k* = 2**, **3** and **4** clusters.
4. Show the clustering results in subplots.
5. Use silhouette for finding the optimal number of clusters. Show in subplots the silhouette values for ***k***=2***, k***=3 and ***k***=4.
6. Make conclusion about the optimal number of clusters.

Write your conclusion with explanations in the Word file **result.docx**.

1. Divide the given text in ***k*** different books in according to the received results up to resolution of 5000 symbols. Add to **result.docx** your conclusion about the place where each book starts (show symbol number interval of size 5000, e.g. “book 2 begin in the interval from 20000 to 25000”). Explain your answer.
2. Write a function for automatic detection of the boundaries between books in the original file.