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 3: Clustering and Text comparison**

**Tasks to do:**

1. Open the file lab3\_ex011.py and read the code.
2. Open and read 2 files 'Eliot.txt' and 'Tolkien.txt'
3. Divide the each text into several parts (chunks) each one of size **wind,** using the given function **partition\_str()**.
4. Construct the dictionary.
5. Calculate frequency matrix **wordFrequency** in according to the **dictionary**.
6. Calculate the distance matrix **dist**. Reduce the distance matrix into **dist1**.
7. Save the matrix dist1 to the file dist1.npy

np.save('dist1',dist1,allow\_pickle = True)

1. Close the current console.
2. Open the file lab3\_ex012.py and read the code.
3. Load the file dist1.npy.
4. Cluster data **dist1** using ***k*-means algorithm**. Specify ***k* = 2** clusters.
5. Demonstrate the clustering results via bar plot.
6. Analyze the data partition into 2 clusters.
7. Calculate and show the silhouette values for k=2.

**Independent work 1:**

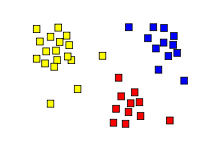
1. Cluster the same data **dist1** into 3 clusters and show the clustering results.
2. Calculate and show the silhouette values for k=3.
3. Use the silhouette method for finding optimal number of clusters. Compare the received values for k=2 and k=3. Make conclusion about optimal number of clusters.

**Independent work2:**

1. Add the text 'DB.txt', repeat the procedure for three given texts.
2. Specify *k* = 2, 3 and 4, use silhouette for finding optimal number of clusters and analyze the results.

**Clustering**

**Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).



**Silhouette Value**

The silhouette value for each point is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters. The silhouette value Si for the ith point is defined as

Si = (bi-ai)/ max(ai,bi)

where ai is the average distance from the ith point to the other points in the same cluster as i, and bi is the minimum average distance from the ith point to points in a different cluster, minimized over clusters.

The silhouette value ranges from –1 to 1. A high silhouette value indicates that i is well matched to its own cluster, and poorly matched to other clusters. If most points have a high silhouette value, then the clustering solution is appropriate. If many points have a low or negative silhouette value, then the clustering solution might have too many or too few clusters. You can use silhouette values as a clustering evaluation criterion with any distance metric.