**Lab 3: K-means Clustering and Text comparison**

**Tasks to do:**

1. Open the file lab3\_ex1.py and read the code.
2. Open and read 2 files 'Eliot.txt' and 'Tolkien.txt'
3. Divide each text into parts (chunks) of size **wind,** using the given function **partitionStr()**.
4. Construct the dictionary.
5. Calculate a frequency matrix **wordFrequency** according to the **dictionary**.
6. Calculate a reduced frequency matrix **wordFrequency**1.
7. Cluster data **word\_frequency1** using ***k*-means algorithm**. Specify ***k* = 2** clusters.
8. Demonstrate the clustering results via bar plot.
9. Analyze the data partition into 2 clusters.
10. Calculate and show the silhouette values for k=2.

**Independent work 1:**

1. Open and read 3 the files 'Eliot.txt','Tolkien.txt', and HP\_Small.txt
2. Generate the reduced **word\_frequency**, cluster the data into 3 clusters and show the ***k*-means** clustering results using the silhouette method.
3. Use k-means with *k* = 2, 3 and 4.
4. Use the silhouette method for finding the optimal number of clusters. analyze and explain your results.

**Independent work2:**

1. Concatenate the text files into a single string.
2. Partition the string of texts into chunks with window size of 20000 characters.
3. Calculate the reduced word frequency and use k-means with *k* = 2, 3 and 4.
4. Use the silhouette method for finding the optimal number of clusters. analyze and explain your results.

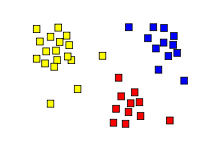
Format of submission

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.

**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.

