**Coursework 2: CLUSTERING OF LARGE UNLABELED DATASETS**

**Task description:**

As we understood, the purpose of this project is to be familiar with cluster analysis using k-mean algorithm, which is the main task for an exploratory data mining and finding underlying patterns. We will use Spark for doing this coursework and Python, mllib implementation for evaluation of the results.

**Use cases:**

1. To demonstrate the use of a K-Means cluster analysis model to get an overall understanding of the various naturally occurring users clusters. This use case will show how cluster analysis can be applied to market research for partitioning a population of users into groupings. This can be useful, for example, for user segmentation and targeting marketing campaigns. (/data/stackoverflow/Users dataset)
2. We are going to implement analysis of the text data to cluster the documents by topics using k-means approach. (/data/stackoverflow/Posts dataset)
3. Analysis of the posts-answers lengths to understand which length represents the short, medium and long answer. (/data/stackoverflow/Posts dataset)
4. Using the post data and votes for the answers we could also cluster all our users as ninja, senior, middle and junior developers. (/data/stackoverflow/Votes and /data/stackoverflow/Users dataset)

**Additional project goals:**will try all of the proposed tasks.

**Extra-mile:**another approaches for clusterisation, C-means, Latent Semantic Analysis (LSA).