Text Clustering Project

# Introduction

The dataset contains text data from Siena courses, we will use clustering to decide which courses belong in which group. We will specify a different number of clusters, 3 clusters, 33 clusters and 57 clusters. After which we will analyze the optimal number of clusters for this dataset. After extracting the appropriate data, we will use k-means, agglomerative clustering and LDA clustering algorithms. After obtaining these results we will use metrics such as a silhouette score and adjusted rand score to see which algorithm gives us the best results and discuss findings related to each algorithm.

# Data Preprocessing and Clustering

The data we have is in a txt file, after storing it in a variable and removing the columns of index we do not need, we use the CountVectorizer() method to transform the data appropriately and fitting and transforming the “decriptions.txt” file. This process sets the stage for the data which will be used for clustering.

### K-means

We use the KMeans method and fit our data to the model and specify the 3 data clusters we need.

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

*Figure 1.1 Figure 1.2*

Figure 1.1 showcases the silhouette score for clusters ranging from 2 to 57. We can see that the silhouette score is decreasing as the number of clusters increases, most notably the sharp change in the start which we see the silhouette scores drop from zero and going towards -1. This trend clearly shows that we go from a value near 0 which means the cluster boundaries are close but different but just before 10 clusters we cross the zero mark. As the number of clusters increase, we are drifting closer to being some outliers and clusters are mixing with each other. The adjusted rand score calculates how correct the clustering was and the different results from the clusters is visualized in Figure 1.2 and we can conclude that this algorithm was not too efficient with assigning clusters due to the low values. Below we attempt to visualize the clusters at the cluster which yielded the best silhouette score.