Project 2

**Introduction:**

Clustering algorithms are unsupervised methods for finding groups of data points that

have similar representations in a proper space. K-means clustering is a simple and popular clustering algorithm. In this project, we need to:

1. To find proper representations of the data, s.t. the clustering is efficient and gives

out reasonable results.

2. To perform K-means clustering on the dataset, and evaluate the performance of the

clustering.

3. To try different preprocess methods which may increase the performance of the

clustering.

In order to define the clustering task, we pretend as if the class labels are not available and aim to find groupings of the documents. We then use class labels as the ground truth to evaluate the performance of the clustering task.  
 To get started with a simple clustering task, we take all the documents in the following classes: class 1(com) class 2(rec).

1. **Building the TF-IDF matrix**

We transform the documents into TF-IDF vectors using min\_df=3 and exclude the stopwords.

1. **2-class Clustering**

In this part, we apply K-means clustering to classify TF-IDF data into 2 classes. And then we examine the result with homogeneity score, completeness score, V-measure, adjusted Rand score and adjusted mutual info score.

1. **Preprocess the data**

For the high dimensional sparse TF-IDF vectors, they cannot yield a good result. Also, when the clusters are not round-shaped, K-means may fail to identify the clusters properly. Thus, we use the package in sklearn – Demonstration of k-means assumptions.

To reduce the dimension, we use NMF and LSI method to dimensionality reduction. Through SVD we calculate the variance remained after dimensionality reduction and sweep over parameters for each method, and choose one that yields better results in clustering purity metrics.

1. **Normalization & Non-linear Transform**

First, we visualize the performance of the case with the best clustering result. And then, based on the best r we got, we used 3 methods to see whether they increase the clustering performance. First we use normalization and then non-linear transformation and the combination of both.

1. **Multi- class Clustering**

In this part, we include all the documents and the corresponding terms in the data matrix and find proper representation through dimensionality reduction of the TF-IDF representation. We try different dimensionality reduction techniques and transformations.