Building the TD-IDF

2-class Clustering

Preprocess the data

Normalization & Non-linear Transform

Multi-class Clustering

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**