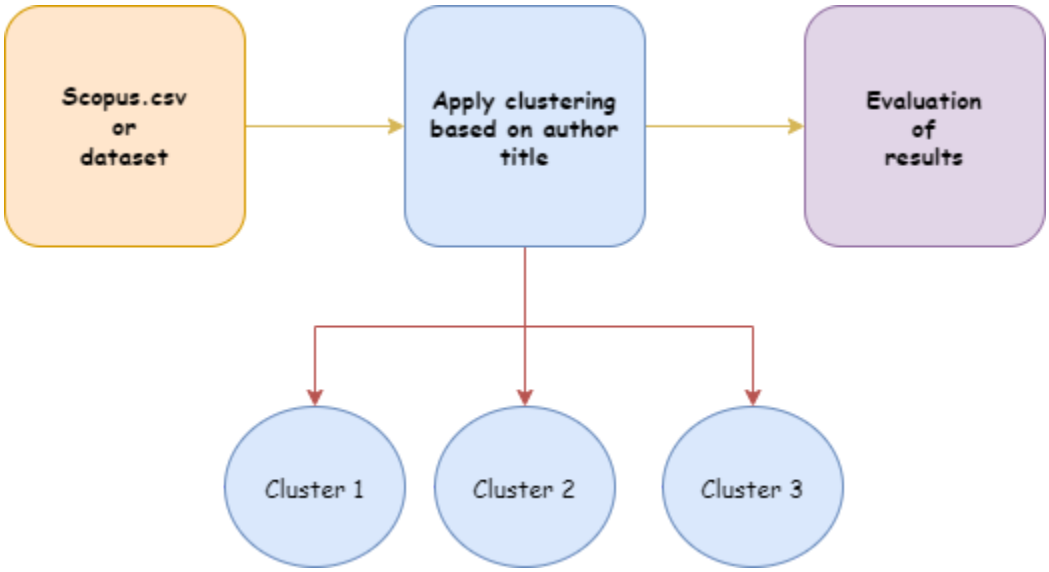
Introduction

In this document we take you through the workflow of our clustering process and there by explain in detail the process and the evaluation methods which were applied.

Dataset:

The dataset contains a result of manual SLR process with their primary studies. The dataset also contains a primary column which indicates the row in the dataset as a primary study, which if true tells us that this study is relevant to the data.

Workflow:



A rough workflow of the process can be seen in the above figure. The workflow of the above process are as follows,

1. We collect data and perform preprocessing on it
2. The processed data is later applied to a clustering algorithm
3. We have used KMeans algorithm inorder to perform clustering
4. The number of clusters are chosen by the elbow method used
5. The results obtained are later evaluated against a labelled dataset
6. A rand score is generated based on the obtained results

Data Preprocessing:

The obtained data is preprocessed, in the sense the data is cleaned. This process is done manually as the required data is kept and the unwanted data is removed. In the end we end up with clean structured data

Clustering and choosing number of clusters:

Clustering is performed on the authors title column, grouping the title of the paper according to their relevance. We use KMeans algorithm as a clustering algorithm which clusters relevant titles into one cluster.

KMeans usually is a centroid selection based algorithm which requires an input of a number of clusters which has to be manually entered. This selection of clusters manually will impact the results. It will be hard to manually select the number of clusters which there by introducing bias in the algorithm and the results fetched cannot be considered as consistent.

Inorder to solve the above problem, the manual selection of clusters needed is done with the help of Elbow method. This method gives us an indication as to how many clusters can be relevant by scanning through the data. Based on the result obtained by the Elbow method we select the number of clusters.

Once the number of clusters are selected, we apply KMeans clustering algorithm on the data and fetch the results.

Evaluation of clustering:

A cluster can be evaluated by measuring minimal intra cluster distance and maximal inter cluster distance. This is performed by using rand metrix.

Rand Metrix is function that measures the similarity of the two assignments, ignoring permutations. Once the clusters are computed we apple rand metrix inorder to perform cluster evaluation.

Results:

Here are the results after the clustering and evaluation process has been applied. We can see that we have a rand score of 0.54 after the clustering process.

```
Primary: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Cluster_predicted [1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0]
Rand Score: 0.5466666666666666
Mallows Score: 0.7393691004272944
```

Summary:

1. Clustering can help in selecting primary studies:
   This has been proved from the above results that clustering can be used for the selection of primary studies which can help reduce the time.

2. Clustering with manual intervention can provide better results.
   We can see that clustering alone will not fetch the results which we wanted inorder to get the results. Clustering with manual intervention will make sure we have proper results. This would be proper inorder to determine the results.

3. Different clustering algorithms can also be incorporated.