***\*\*PLEASE ARRANGE THE REPORT IN THIS STRUCTURE***

**Chapter 1 :** Introduction and Objectives

*(give the problem description, with diagrams)*

**Chapter 2 :** Our approach

*( describe the modules in short and the architecture, mention that all the)*

Modules Description

**Chapter 3 :** Algorithms

*(mention about the common algorithms used and give a brief introduction. However, do not describe the algorithms here. Just mention the purpose of the algorithms. This should be of 1-2 pages only.)*

TFIDF

Clustering

**Chapter 4 :** Preprocessing

Preprocessing, why its necessary.

Stopword removal *(why and how? Give examples of initial text with stopword and final text without stopword)*

Punctuation removal (why?)

Numbers removal (why?)

---- removal

*(mention and describe the four preprocessing tasks that we have done. Also, if possible find some other common text preprocessing tasks from the net and insert it)*

**Chapter 5 :** Term Weighting

TFIDF *(explain everything about TFIDF here.)*

**Chapter 6 :** Document Similarity Computation

5. Comparison of TFIDF files for clustering

5.1. Method1

5.2. Method2

5.3. Method3

**Chapter 7 :** Clustering

Introduction

Comparison of clustering algorithms

Selection of an algo for our purpose

DBSCAN

Introduction

Advantages

Disadvantages

**Chapter 8** : Clustering Performance Evaluation

Approaches to measure clustering performance

Adjusted Rand Index

Mutual Information based scores

Problems in clustering

Single Cluster Explosion

Cluster Explosion

Unified Cluster formation

Estimating initial threshold from average scores

Estimation of best threshold for clustering

**Chapter 9** : Project Implementation

*(this is in short the documentation, with module and functions description)*

All the modules

**Chapter 10** : Dataset

*(describe the dataset in detail)*

List of documents

List of genres or topics.

A brief description on the documents of each topic and a personal description on what the similarities might be between documents.

**Chapter 11** : Results

*(show all the results here in this order:)*

Screenshot of final result.

Graphical visualization of the results. (the circle diagrams)

Graph of variance of cluster performance with varying threshold

**Chapter 12** : Conclusion

*(will add this at last)*

**Chapter 13** : References

*(later)*