Machine Learning Synopsys

In the Partial fulfilment of B. Tech. - III year

Course requirement of

Subject: Machine Learning and Data Mining

BML MUNJAL UNIVERSITY™

FROM HERE TO THE WORLD

Project Title

Document Clustering

Submitted To: -

Dr. Atul Mishra

Assistant Professor

School of Engineering & Technology

BML Munjal University

Submitted by: -

1. Ratandeep(1700265C203)

2. Pujith Sai Kumar (1700225C203)

Introduction

Document clustering is an automatic grouping of text documents into clusters (groups) so that documents within a cluster have higher degree of similarity but are dissimilar to documents in another cluster. The goal of a document clustering scheme is to minimize intra-cluster distances (using an appropriate distance measure between documents). A distance measure (or similarity measure) thus lies at the core of document clustering. The large variety of documents makes it almost impossible to create a general algorithm which can work best in case of all kinds of data sets.

Libraries Being Used: -

NLTK

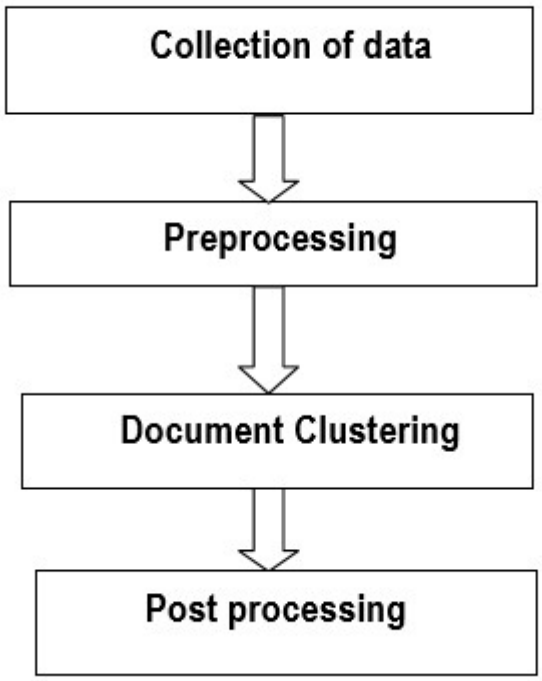
Scikit Learn

NUMPY

Methodology

Since we are new in this field, we are studying various resources available for text processing and NLP taking help of various researches that has been done in this field which has been mentioned in our status report. We try our best to show you the best output possible till the course end after which we can extend the work if possible.

General steps for document clustering are: -



Currently we are learning about various ways of text processing such as TF-IDF matrix and other concepts such as Euclidean distance etc and trying to process the text with these.

Note: Due to unforeseen circumstances, we forget to send the synopsis on time. Hope you would understand. And you can be sure that we are putting our best efforts on this. Also please bear with me that I can't submit the synopsis of my **(Pujith)** another project right now as we are busy in working on a new project to handle this Covid Pandemic to help **GoI** regarding which I am going to contact you soon.

Thank you!