MES COLLEGE OF ENGINEERING, KUTTIPPURAM
DEPARTMENT OF COMPUTER APPLICATIONS
20MCA245 – MINI PROJECT

PRO FORMA FOR THE APPROVAL OF THE THIRD SEMESTER MINI PROJECT

(Note: All entries of the pro forma for approval should be filled up with appropriate and complete information. Incomplete Pro forma of approval in any respect will be rejected.)

Mini Project Proposal No :
(Filled by the Department)

Academic Year : 2021-2022
Year of Admission : 2020

1. Title of the Project : AUTOMATIC DOCUMENT CLASSIFIER
2. Name of the Guide : Mr. Mohammad Jabir C
3. Number of the Student: MES20MCA-2035
4. Student Details (in BLOCK LETTERS)

	Name	Roll Number	Signature
1.	NABEEL E P	35	

Date: 01/12/2021

Approval Status : Approved / Not Approved

Signature of Committee Members

Comments of The Mini Project Guide — Dated Signature

Initial Submission :

First Review :

Second Review :

Comments of The Project Coordinator — Dated Signature

Initial Submission:

First Review

Second Review

Final Comments :

Dated Signature of HOD

AUTOMATIC DOCUMENT CLASSIFIER

NABEEL E P

Introduction:

The Web use has strengthened the creation of digital information in an accelerated way and about multiple topics. The classification of text is widely used to filter e-mails, classify web pages and organize the results recovered by the web browsers. In the process of recovering information, which includes work-tasks of representation, organization, storage and access to the information, it is desired to have associations of documents by keywords at all times. In this way, classifying documents in an automatic way would allow us to find information in a more efficient way

The classification is a grouping procedure which allows us to group a set of data according to a selected criterion. Generally, the objects or data of the same group share similar characteristics with one another, while the objects of different groups will have less similarity among them. For example, in an organization the documents can be classified by a functional criterion, that is to say, grouping the documents by activities inside of the company, or through a criterion of hierarchical order, where the managers have access to different documents that employees have access to. The goal in the task of classification is to locate the document of an appropriate class.

Having a large number of features makes the classification process to be computationally expensive and that the classes not being well defined. The feature selection focuses on reducing the dimensionality, many approaches concentrate on considering only one subset of features extracted from text. Generally, for the selection of characteristics we have techniques based in the collection of documents and techniques based in typifying each class.

Objectives:

The main objective is to:

• Improve the customer experience and throughput rate of your classification -heavy processes without increasing costs.

• Automate the process of grouping documents and use this information to process an entire volume of documents.

• Take documents and easily organize, extract, and apply key metadata to simplify and organize documents into a content management.

Problem Definition:

Existing System

The use of evolutionary algorithms to solve classification problems has been a recurrent approach. The use of the ACM taxonomy was proposed in order to obtain thy similarity among documents, where each document is formed by a set of keywords. A methodology to obtain the distance between the words in the ACM taxonomy was designed. Such methodology makes use of the FloydWarshall algorithm, which is typically used to obtain the minimum distance between two nodes in a graph. The grouping of scientific documents was proposed as a problem, for which it was designed a genetic algorithm for classification.

• Uses genetic algorithm for text classification.

• ACM taxonomy is used to measure similarities between documents.

• Floyd Warshall Algorithm used to find distance between nodes in a graph.

Proposed System

In the proposed system it is expected to evaluate the efficiency of algorithm with higher number of real articles So, in this way we do the text grouping of document sets of any domain and without being restricted to a few categories. It can be used to decide which document is taken to which cluster, and classification the scientific documents or web documents in a particular cluster.

Additionally, compared the system being developed is capable of classifying documents which belonging to topics which were not provided in the data set into a separate category. It is also capable of identifying duplicate documents present in the input provided if any.

The implementation of this system will be done as a web application and will include User and Admin modules. The admin module will contain some of the main function such as Dataset creation, there are ither functions provided such as verifying user, communication with the user etc. The user module contains functions for uploading documents, also includes the functions of sending feedbacks, complaints etc and also view the communications with the admin. The main function of the user module is to view the final classification under the categories, which is the main output of the application.

Basic functionalities:

Functional Module

Genetic Algorithm

Following is the foundation of GAs based on this analogy –

i. Individual in population compete for resources and mate, here the individuals are keywords of the document and of the dataset.
ii. Those individuals who are successful (fittest) then mate to create more offspring than others.
iii. Genes from "fittest" parent propagate throughout the generation that is sometimes parents create offspring which is better than either parent. Here the parent is the Type of the document or the category to which the document is matched.
iv. Thus each successive generation is expected to have better qualities than previous ones in terms of solutions.

Module Description

- Admin
- Student

Admin

• Verify the user

• Reply to messages

• Add Datasets

Student

Send and receive Messages.

• Upload Documents.

• View Category.

• Identify duplicate documents.

• View history (All documents which user has uploaded along with its categories).

• Display documents under different categories.

Tools / Platform, Hardware and Software Requirements:

Hardware & Software Requirement

Hardware Requirements

The selection of hardware is very important in the existence and proper working of any software. Then selection hardware, the size and capacity requirements are also important.

- Processor : Intel Pentium Core i3 and above, 64 bits
- RAM : Min 4GB RAM
- Hard Disk: 10 GB

Software Requirements

One of the most difficult task is selecting software for the system, once the system requirements is found out then we have to determine whether a particular software package fits for those system requirements. The application requirement:

- OPERATING SYSTEM : WINDOWS 10
- FRONT END : HTML, CSS, JAVASCRIPT
- BACK END : Mysql
- IDE USED : Jetbrains Pycharm, Android studio
- TECHNOLOGY USED : PYTHON JAVA
- FRAME WORK USED : Flask