**RESEARCH PAPERS SEARCH ENGINE**

**PHASE 3 – Additional Enhancements**

**ABSTRACT – 16PW24, 16PW33**

**Project Abstract:**

The aim is to develop a search engine which would search for papers related to a paper of interest to the user. It displays the user a list of relevant papers to the query. All the documents shown contain at least one of the query words and in the results shown, all query words are bold marked. The entire project is implemented in python.

The additional enhancements are

1. Long Query Algorithm
2. Similar Papers
3. Categorisation of the Research Papers

**Long Query Algorithm:**

Allowing users to specifya long query, defined as having more than 10 words after preprocessing the query. Theidea is that the user, instead of searching for a paper by its title, should specify ageneral idea and describe it in more detail to allow us to find a relevant paper that presents the same ideas.

The phrases are selected using the NLTK RAKE library. Furthermore, we have chosen to implement a limit of four words for each of the extracted phrases, this is done to limit the amount of keywords extracted in total. After extracting all the phrases we combine them into one query and remove all duplicate words.

**Similar Papers:**

Suppose a user finds a paper useful to his search, and wishes to find more papers similar to this one to formulate the search query using the selected document’s abstract. The abstract of a document is usually a summary of a paper consisting of the main keywords and features and hence using this as the search query made sense.

**Categorisation Of the Research Papers:**

Each research paper belongs to a category. The training set is prepared with preprocessed text. CountVectorizer is used considering 2500 features. Multinomial Naïve Bayes Algorithm is used to train. With the help of the model the categories are predicted for the research papers. So when a user types a query along with the relevant paper’s abstract and url to access it , the category under which the paper belongs to is also shown.