# Major Technical Project
# Report 1

## Title of MTP

Building a Search Engine

## Abstract of MTP Proposal

A web search engine is basically a tool which is widely used to search information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a specialist in web pages, images, information and other types of files. The idea here is to build a search engine based on our own text analysis which will be dependent on several factors. Broadly speaking, when a user enters a word or a string to search, the most relevant pages containing related and best content should be retrieved by the application to the user in a ranked order. The page towards the top of the final result will be ranked the highest among all the possible options.

Keywords : Search Engine, PageRank, Spider, Crawling, Indexing, Big Data, Map Reduce, Hubs and Authorities

## Course of Action for MTP

The Entire project will be divided into three subparts which will eventually lead to the formation of search engine. The subproblems are :

1. Web Crawling
2. Indexing
3. Searching

For the Implementation, decomposing the above mentioned into the important parts of search engine gives us:

- Crawler

  Crawler browses the web, downloads documents(pages) and saves them.

- Parser

  Parser transforms downloaded documents into some internal format suitable for indexing.

- Indexer

  Indexer processes the documents parsed by crawler, and creates the necessary structures.

- Ranker

  Ranker is the part that arranges documents in some particular order. The goal is to put the most relevant documents on top of a results page. Usually machine learning is used here, but there is the difficulty of obtaining data-sets(for test and training). One option is to use BM25 or some other text ranking function.

- Snippets Generator

  It generates short extracts from the documents. They are usually put under the link.

- Searcher

  Searcher takes queries entered by users, transforms them somehow, parses them and finds documents containing words of query, gets all the necessary stuff like annotations and finally generates and presents results pages.

## Work Carried out till now

One of the important aspects of search engine that determines the relevance of various significant pages i.e. Page Rank and various other algorithms namely Topic Sensitive Page Rank and Hubs & Authorities have been studied and implemented on big data set. Some of the other things that are required for the successful working of search engine like Web Crawler and Text Processing have been browsed hence giving a basic insight of these and their implementation towards a functional Search Engine.

## Team

Name : Sahil Mutneja
Roll No : B11031
Email : sahil_mutneja@students.iitmandi.ac.in
Major : Computer Science and Engineering

Signature

## Faculty supervisor name and signature

Dr. Dileep A.D.
Assistant Professor
School of Computing and Electrical Engineering

Signature