AP Lab 5 documentation

# Introduction

This lab is mainly about creating an elaborate search engine for our own local system. We have to create a crawler to visit all the directory and files. Next we have to index all the searchable content, including the content in the text files.

# Problem

There are 3 main tasks to be done in this lab. First we need to get all the directory names and file names. We also need their paths. Second we need to parse the content in the text files, so that we can ultimately index it. Lastly we need to index all the information based on the names of files, directories and words found in text files. Once done, if user enters a keyword we can return all the files & directories which contains the instances of that specific keyword.

* The first is to develop a crawler to get the paths of all the files and the directories
* Next we need to separate out the .txt files from rest of the files
* Open files and parse their content. Make content key and value should be the path of that file.

# Solution

* Python has a library called os and in that library there is a function called walk().Generate the file names in a directory tree by walking the tree either top-down or bottom-up. For each directory in the tree rooted at directory top including top itself), it yields a 3-tuple (dirpath, dirnames, filenames).
* Checking for .txt extension for all the files in fileNames.
* Done