## horizontal line



Document Detective

an Intelligent Document Finder

**Created by** Muhammath Azharudeen S, Muhammath Ashik Farhan S

# Problem Statement

Intelligent Document Finder tool that can provide easy and intelligent searches among the document files. The required document type includes presentations, pdf, doc and txt files. The user will enter a few queries in the search page of the tool to look for the most relevant documents.

# Technique Used

# Natural Language Processing [NLP] Techniques.

# Machine Learning [ML] Algorithm - K-means Clustering.

* Programming Language - Python

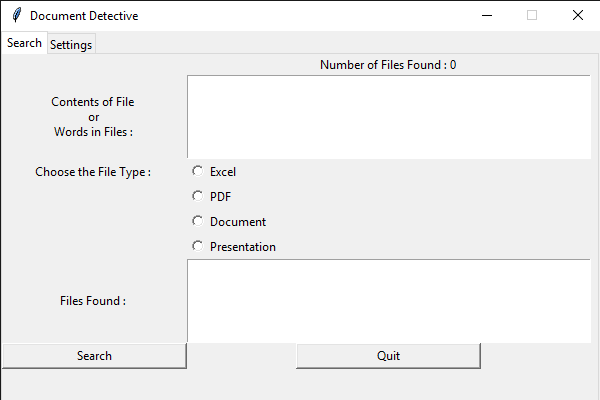
# Idea behind the Tool

**Train** - First detect all the files with required formats. For each file format, read all the data and text elements of files and preprocessing it with **NLP** techniques and train them in **ML Clustering Model** with a certain number of classes. So, now the files belong to specific classes.

**Search** - Some words of content (not consecutive) of the required file. For Example: In Cycle Proposal Presentation, there should be the words of cycle model/name, brand name, contact details like addresses, and more. Like that we need some (core) words which are used in that file. With those words, we can easily find the file.

# Components

In *Search* Tab,

**At Initial**

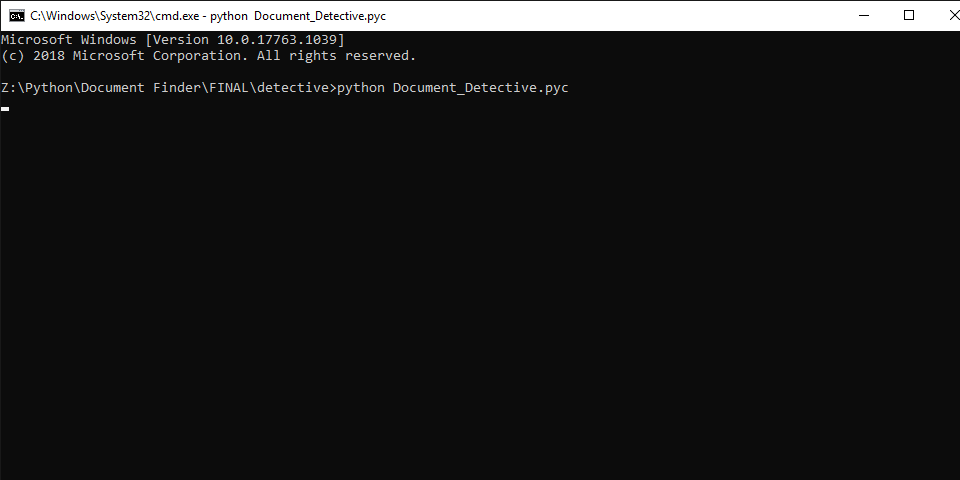
*Contents of File or Words in Files* - Type some words related to the file to be searched

*Choose the File Type* - Choose the file format

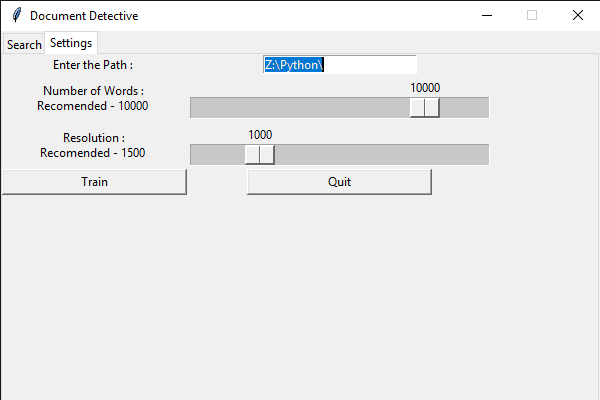
*Search* - Click to start search with contents and format

*Files Found* - Shows the detected file path

*Quit* - Press to Stop the Application

**At Initial**

In *Setting* Tab,

**Before Trained**

*Enter the Path* - Base File Path

*Number of Words* - Configure the number of words to be read from from each file

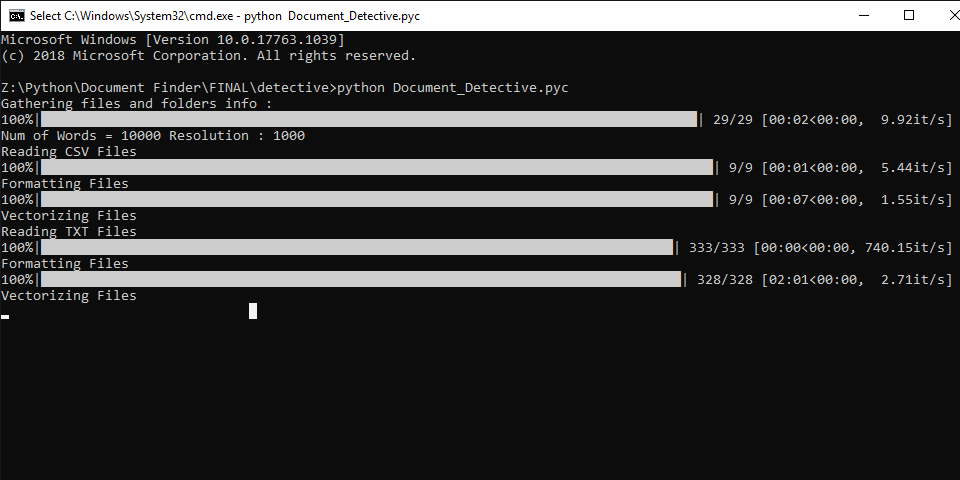
(Note : Large in No. of Words increases accuracy and reduces processing speed and vice versa)

*Resolution Slider* - Configure the optimization value depends on the total number of files

(Note : Number of Words and Resolution Slider configuration is OPTIONAL)

*Update* - To update the changed parameters

Quit - To Stop the Application

**After Trained**

# Procedure

Open the ***Document\_Detector.pyc*** File with Python

(or)

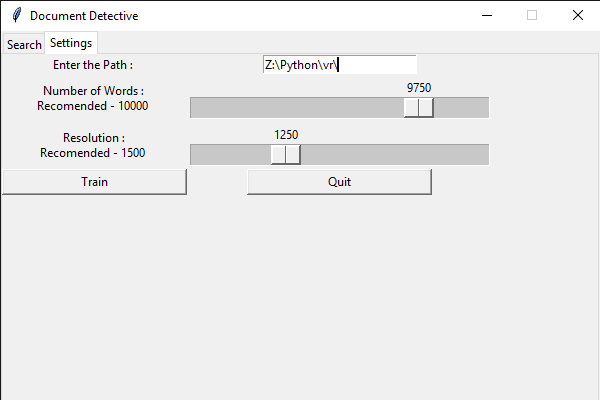
In CMD, type ***python Document\_Detector.pyc***of the file located path

Train the Tool

* Now Document Detective window appears
* Set the Path in Settings tab
* (Optional) - Configure the Sliders
* Click Train

In the background, it will do some process. So please wait until the window appears again.

(Note : Do not disturb the Window until it comes otherwise it will become Unresponsive)



Search with Tool

* Enter some words of file in Contents of File Box
* Choose the file format
* Press Search to start search
* Wait to grab the file
* Once finished we can see the file path in Files Found box

**Note : Search only works if it should atleast train for one time with the correct path.**

**Check the below demo video link:**

https://drive.google.com/open?id=1F2XTplUAa4JMqcmb\_8qTs0UgPqllgXUk

# Achievements

1. Provide faster search of documents.
2. Search with non consecutive words of the document.
3. Flexible with document formats.