**Indexing PDF documents using FSCrawler**

This crawler helps to index binary documents such as PDF, Open Office, MS Office.

***References: https://fscrawler.readthedocs.io/en/fscrawler-2.6/***

1: Download **FSCrawler** from **Github**

2: Create a directory in FScrawler along with a job name

After downloading, go to fscrawler bin directory using cmd console :

***C:\tmp\fscrwaler\bin>fscrawler --config\_dir ./Directory index\_name***

**Example: *C:\tmp\fscrwaler\bin>fscrawler --config\_dir ./DS data\_science***

On running the above command, it will ask whether you want to create the directory if it is not created, say yes and create the directory along with the job name.

Once done, check whether the directory is created inside FSCrawler folder or not,

FSCrawler 🡪 DS 🡪 data\_science 🡪 \_settings.json/ settings.yaml

Now, make changes in ***data\_science (\_setting.json)***file directing to the original path of the pdf files which are to be indexed.

There are some changes required in the ***'\_setting.json'*** file before running the fscrawler for indexing the pdf's.

**Step 1 :** Set url path to location of pdf files (Here, double forward slash is required while defining the path for fscrawler to run)

*Set* ***"url" : "D:\\ELK\_SETUP\\PDF\\DS\_books",***

**Step 2 :** If you don’t want to delete the pdf document fron elasticsearch in case it is deleted from the original path/ folder containing the pdfs after it is indexed (set 'remove deleted' to false)

*Set* ***"remove\_deleted" : false,***

**Step 3 :** If you don’t want to index the raw metadata of the pdfs.

Set ***"raw\_metadata" : false,***

Now after making changes in the '\_settings.json' we can run the fscrawler to index the pdfs.

On the cmd prompt run the following command inside fscrwaler bin-

***bin/fscrawler --config\_dir ./DS data\_science --loop 1***

**Note:** --loop 1: In case we only want to index the document once.

**Query**:

GET data\_science/\_search

{ "\_source": "path",

"query":{

"match\_phrase": {

"content": "Multiple Logistic Regression"

}}}