**README**

**Image Text Extraction and Information Retrieval**

Any good research work first starts with a question. So the question we have is, How can we use python image processing and information retrieval libraries to extract text from images and search through the results with specific queries?

The dataset which we have chosen here to solve the above-mentioned question is typed and hand-written document images and scans. The dataset was created from a combination of multiple datasets. It consists of more or less 1509 data points.

The image text detection libraries were used to convert the document into computer text and consequently allowing us to search through the documents.

**ipynb:**

Machine learning libraries (PIL, Pytesseract, Pandas, Elasticsearch) were first loaded.

Note\*: As we were not being able to import the python libraries on colab, jupyter notebook and pip install should be used for the code to work.

**Data:**

The first dataset in Data/Typed is from <https://guillaumejaume.github.io/FUNSD/>

The second dataset in Data/Written (png-files) is from <https://fki.tic.heia-fr.ch/databases/iam-on-line-handwriting-database>

The third dataset in Data/Written (jpg-files) is from <https://goodnotes.com/gnhk/>

**DataStore:**

The image data is stored in two separate folders and created, converted into a dictionary that stores the images and results into a Panda’s DataFrame. Then the DataFrame was exported to a CSV file: “output.csv”

**Conclusion:**

The inputted searches from user input where the es.search is used to search through the created index for searches inputted by the user. The output is obtained after the user exits. Neither typed nor handwritten texts are extracted from the images with perfect accuracy. Typed texts have higher accuracy in comparison to handwritten ones.