**Documents Splitting**

1. **Target**

* Scanned document (PDF) contains a several types of documents. Many applications need to split the scanned document image to get the information for key index fields to operates expected task.
* The major task is to index different types of documents, which helps in separation and extraction information from a variety of complex scanned document.
* The target documents spitting was represented following the table below, where 0 represents to start index, 1 represents to index element.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Page Number | | | | | | | | | |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | … |
| Type | A | | | B | | | | C | | … |
| Target | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |  |

1. **Background**

* Input: scanned document in PDF format, includes a variety of complex scanned document (image with text).
* Output: Classify each page of document into two-class, class 1: start index document, class 0: index element
* Programing language: Python
* Method: Machine Learning / Deep Learning
* Evaluation method: confusion matrix, f1-score, accuracy

1. **Solution Engine**

Step 1: PDF to Image conversion

* Module: pdf2image, PyMuPDF

Step 2: Feature extraction (text extraction)

* OCR (Optical Character Recognition) engine: Tesseract, EasyOCR

Step 3: Tokenization

* Module: Gensim, TfidfVectorizer, HashingVectorizer

Step 4: Classification

* Module: Gensim, classification deep learning model

1. **Pipeline**