Document Segmentation

BigTobacco: Contains 3 folders named train, validation and test. Each folder contains several H5DF files where each file contains the following information from BigTobacco dataset:

1. Image: An Image from BigTobacco scaled to 256x256
2. Ocr: Text present in the image extracted with pytest react
3. Label: The value is 1 when the Image is the first of a Document

BigTobacco_filtered:Contains 3 folders named train, validation and test but in this case the files present in train have been cleaned in order to balance the number of occurrences of pages that are the start of a new document and those that are the continuation of a document . Each folder contains several H5DF files where each file contains the following information from BigTobacco dataset:

1. Image: A path to an Image from BigTobacco
2. Ocr: Text present in the image extracted with pytest react
3. Label: The value is 1 when the Image is the first of a Document

BigTobacco_Layout:Contains 3 folders named train, validation and test generated from BigTobacco. Each folder contains several H5DF files prepared for LayoutLM_V2 model. The contents for each file are the following ones:

1. Image: A path to an Image from BigTobacco
2. Ocr: Text present in the image extracted with pytest react
3. Ocr_type:Indicate if text is part from document A or B
4. Attention_mask: Matrix that indicates which tokens are significant
5. box: Coordinates of the bounding box that contains the text
6. Label: The value is 1 when the Image is the first of a Document

Tobacco800:Contains 1 folder named test. Each folder contains several H5DF files where each file contains the following information from Tobacco800 dataset:

1. Image: An Image from BigTobacco scaled to 256x256
2. Ocr: Text present in the image extracted with pytest react
3. Label: The value is 1 when the Image is the first of a Document

Tobacco800_split:Contains 3 folders named train, validation and test. Each folder contains several H5DF files where each file contains the following information:

1. Image: An Image from BigTobacco scaled to 256x256
2. Ocr: Text present in the image extracted with pytest react
3. Label: The value is 1 when the Image is the first of a Document