Python for data analysis project

Meric Théo

Menard Sarah

The Dataset : Blocks Classification

The problem consists in classifying all the blocks of the page layout of a document that has been detected by a segmentation process.

This is an essential step in document analysis in order to separate text from graphic areas.

Indeed, the five classes are:
-text (1)
-horizontal line (2)
-picture (3)
-vertical line (4)
-graphic (5)

The features of the dataset

height: integer. | Height of the block.
lenght: integer. | Length of the block.
area: integer. | Area of the block (height * lenght);
eccen: continuous. | Eccentricity of the block (lenght / height);
p_black: continuous. | Percentage of black pixels within the block (blackpix / area);
p_and: continuous. | Percentage of black pixels after the application of the Run Length Smoothing Algorithm (RLSA) (blackand / area);
mean_tr: continuous. | Mean number of white-black transitions (blackpix / wb_trans);
blackpix: integer. | Total number of black pixels in the original bitmap of the block.
blackand: integer. | Total number of black pixels in the bitmap of the block after the RLSA.
wb_trans: integer. | Number of white-black transitions in the original bitmap of the block.

Data visualisation and Modelisation

-We want to know how we can classify these blocks by working on the features of the dataset we have.

-At first we decided to visualise all the features and how the dataset works. (using pandas, seaborn)

-Then we will make predictions on this dataset and find the best model to classify these blocks. (using scikit-learn)