FinTOC 2019

I. Title detection

This subtask consists of detecting titles from non titles in a document. To this end, we provide a training dataset containing:

1. The PDF format of the documents

2. The XML representation of the PDFs as given by the Poppler utility libraries; this representation contains the text of the documents as well as layout information about the text (font, bold, italic and coordinates)

3. A CSV file containing the following fields:
    a. *text blocks*: a list of strings computed by a heuristic algorithm; the algorithm segments the documents into homogeneous text regions according to given rules
    b. *begins_with_numbering*: 1 if the text block begins with a numbering such as *1.*, *A/*, *b)*, *III.*, etc….; 0 otherwise
    c. *is_bold*: 1 if the title appear in bold in the PDF document; 0 otherwise
    d. *is_italic*: 1 if the title is in italic in the pdf document; 0 otherwise
    e. *is_all_caps*: 1 if the title is all composed of capital letters; 0 otherwise
    f. *begins_with_cap*: 1 if the title begins with a capital letter; 0 otherwise
    g. *xmlfile*: the xmlfile from which the above features have been derived
    h. *page_nb*: the page number in the PDF where appears the text bock
    i. *label*: 1 if text line is a title, 0 otherwise

    The columns of the CSV file are delimited by tabs.

The test set will have the same format as the training, except that the last column of the CSV file will be empty.

Participants are expected to submit a CSV file similar to that of the testing set with the last column filled with their predictions.

Evaluation metric

The evaluation metric used for this subtask is the weighted F1 score. We provide an evaluation script to all participants (cf shared_task/title_detection/compute_score.py)

II. TOC generation

This subtask consists of predicting a TOC from PDF documents. To this end, we provide a training set containing:

1. The PDF format of the documents

2. The XML representation of the PDFs as given by the Poppler utility libraries

3. The annotated TOCs in the XML format proposed by the ICDAR 2013 competition on book structure extraction [1]. This XML file is composed of *toc-entries* with the following attributes:
    a. *title*: a title of the document
    b. *page*: the page number in the PDF document where appears the title
    c. *matter_attrib*: whether the title appears on the front page, the body, or the back matter of the documents; here is a Wikipedia page explaining the notion of "matter"

    This XML file has the extension "icdar2013.xml".

    Note that the level of a title can be inferred from the hierarchy of the XML file.

The test set will be composed of the PDF and XML format of the documents only.

The train documents are the same as those for the "Title detection" subtask. Participants can either use the segmentation into text blocks suggested in the CSV file provided for the previous subtask, or come up with their own segmentation algorithm ***which we highly encourage***.
In case, a participant only participates to the TOC generation subtask, he/she can use a CSV file we provide in which we give a possible segmentation of the test documents into text blocks as well as their title/non-title label.

We expect participants to submit their predictions in the format described in II.3. They can omit the attribute *matter_attrib* or put a default value for it in the submission. This will not affect their score.

Evaluation metric

The evaluation metric for this subtask is based on the official title-based measure of the ICDAR 2013 competition on book structure extraction [1].
The official title-based measure first searches in the submission the toc-entries whose titles match a title in the groundtruth (the matching being a threshold on a weighted Levenshtein distance with updated costs for edit operations). F1 score is

computed while considering as correct the entries which have the same title and page number.
The official title-based measure from [1] also provides a level accuracy to evaluate the hierarchy of the predicted TOC:

$$level\_acc = nOkLevel / nOk$$

with :

nOkLevel = number of toc-entries with a correct page number and a correct level
nOk = number of toc-entries with a correct page number

Both F1 score and level accuracy are computed at the document level and then averaged. To rank participants, we use a weighted sum of these two average metrics :

$$metric = (F1 + level\_acc) / 2$$

III. Bibliography

[1] A. Doucet, G. Kazai, S. Colutto, and G. Muhlberger, Overview of the ICDAR 2013 Competition on Book Structure Extraction. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference,* pages 1438-1443. IEEE, 2013. https://hal.archives-ouvertes.fr/hal-01073396/document