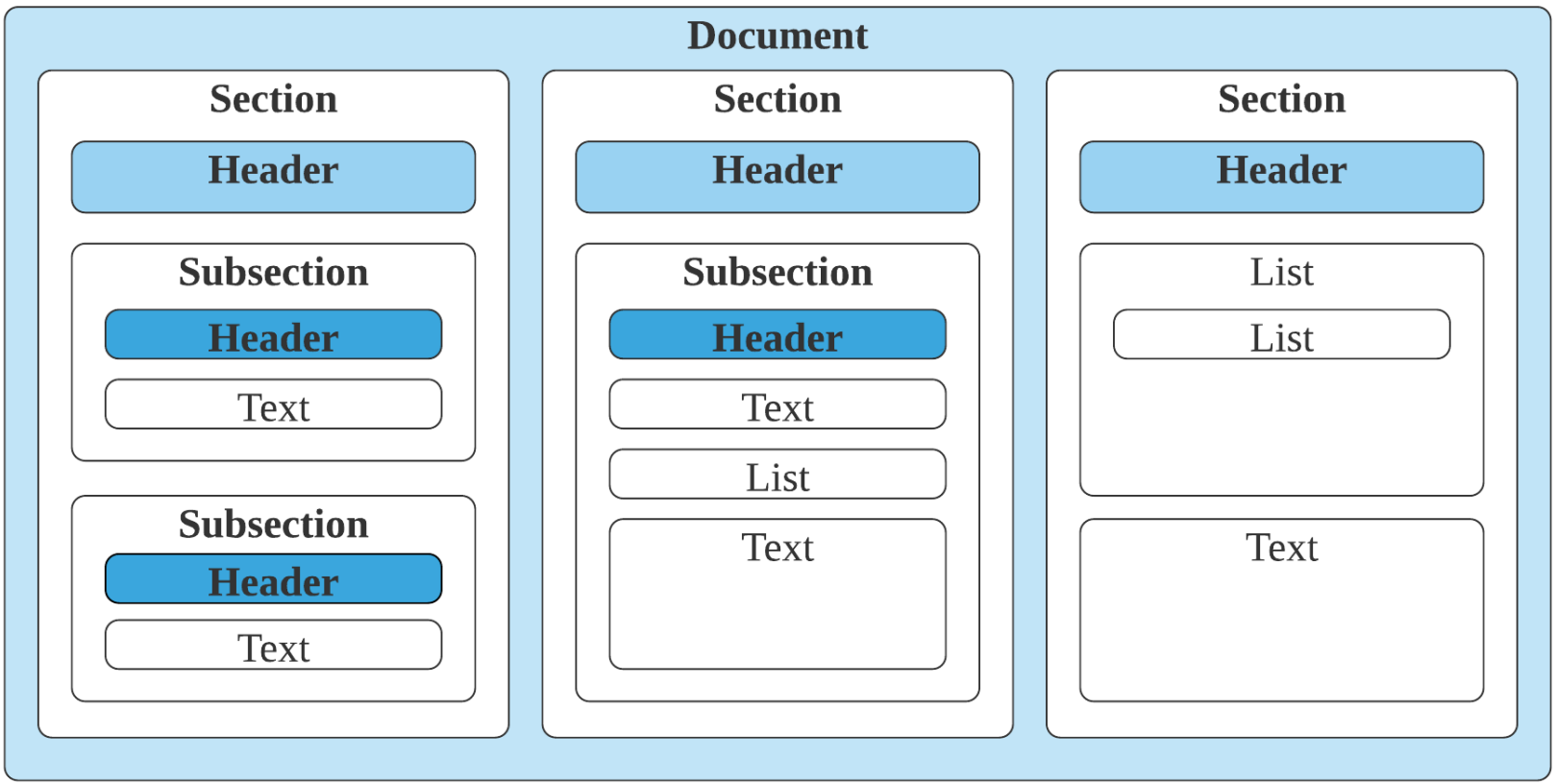
Logical structure extraction from scanned documents

Bogatenkova A.O., Kozlov I.S., Belyaeva O.V., Perminov A.I.

September 25, 2020

Introduction

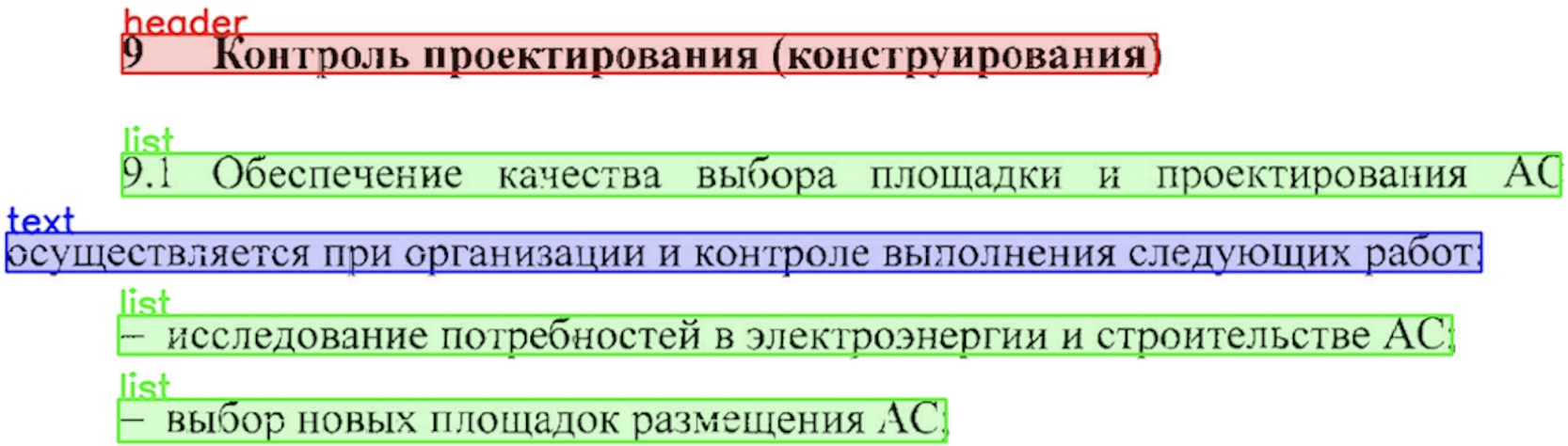
A huge number of scanned documents exist that one needs to work with. Extraction the structure from such documents may be useful for their analysis.



The purpose of the work

The task of extracting a logical structure from documents is being solved. We will describe the pipeline for scanned documents processing. The method is based on the multiclass classification of document lines. The set of classes includes:

- headers;
- list elements;
- textual lines.



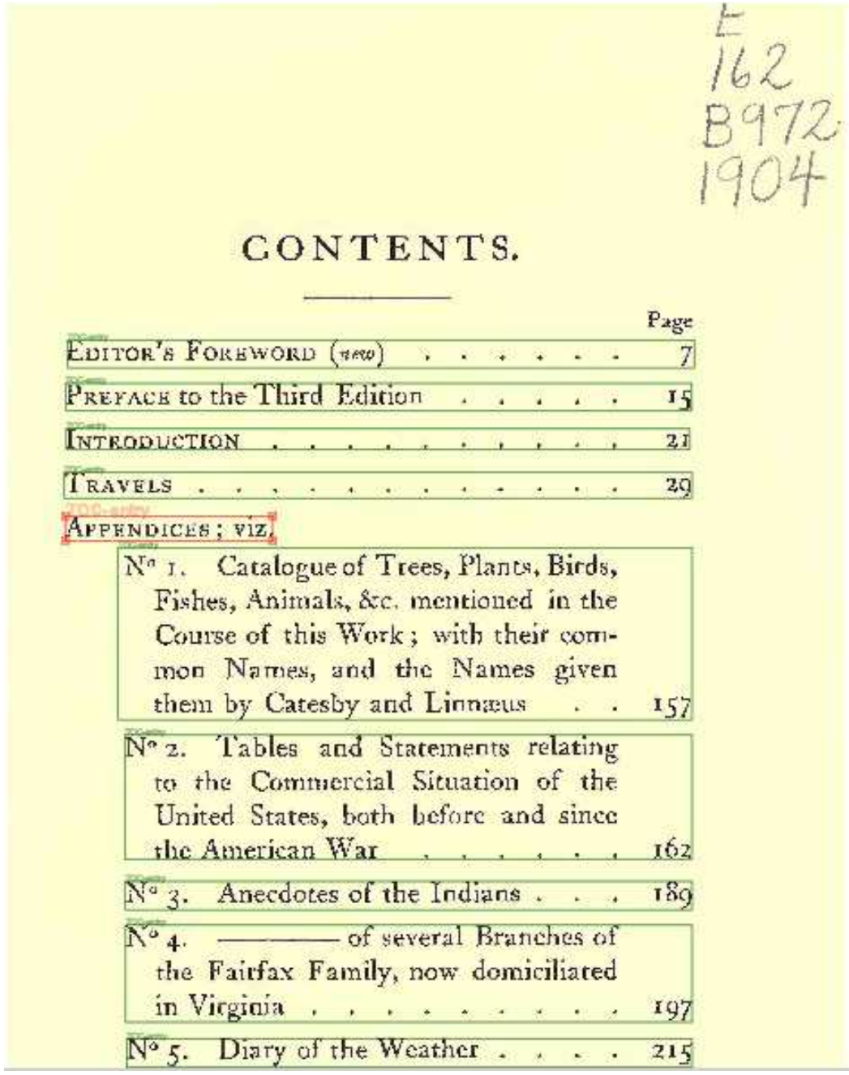
Related work

Several approaches for structure extraction from documents exist:

- methods based on table of contents (TOC);
- rule based methods;
- methods based on machine learning.

1. Structure extraction using table of contents

A lot of ICDAR competitions connected with the analysis of documents are held. In one of these competitions, the structure was extracted from books. It was proposed that books consist of pages, paragraphs, chapters, such structure was extracted using the table of contents, which was present in most books.

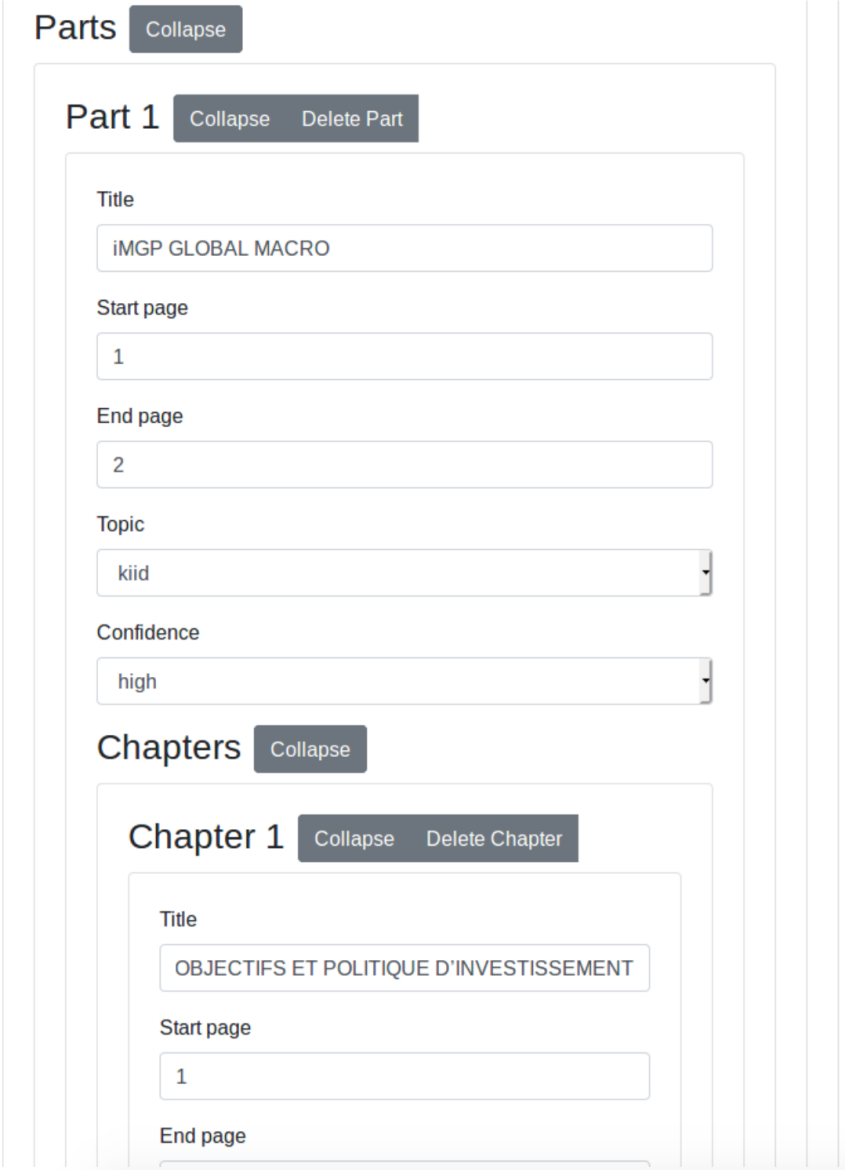


E
162
B972
1904

CONTENTS.

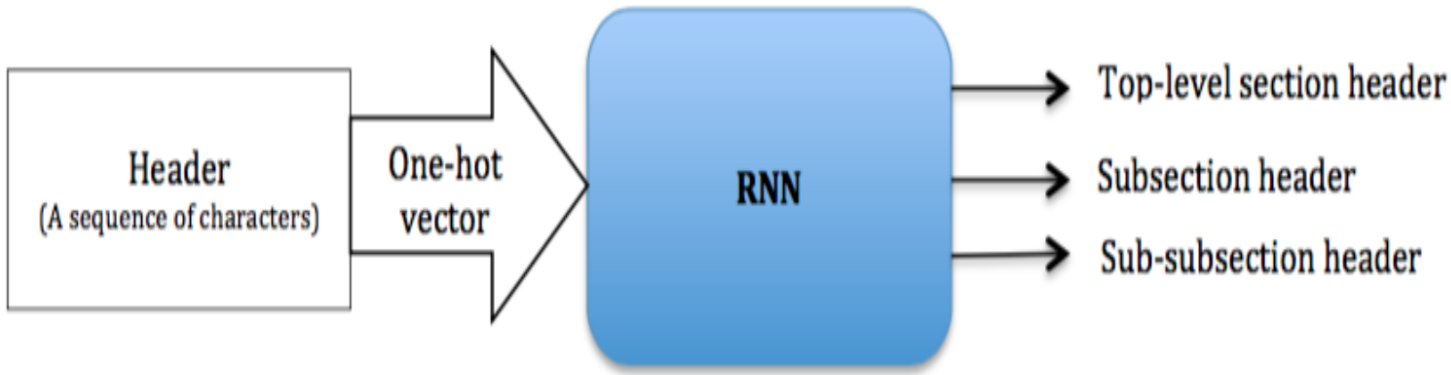
2. Rule based methods

In 2019 on the FinTOC competitions a structure was extracted from financial documents in the form of a hierarchy of document headings with different levels.
One of the participating teams extracted the necessary structure using the table of contents and rules (line spacing, indentation, font, numbering).



3. Methods based on machine learning.

In a 2017 article, document structure was extracted using machine learning techniques including deep learning. Firstly, the classifier for lines determined if the line was the title or not, then the title lines were classified more precisely by section classifiers.

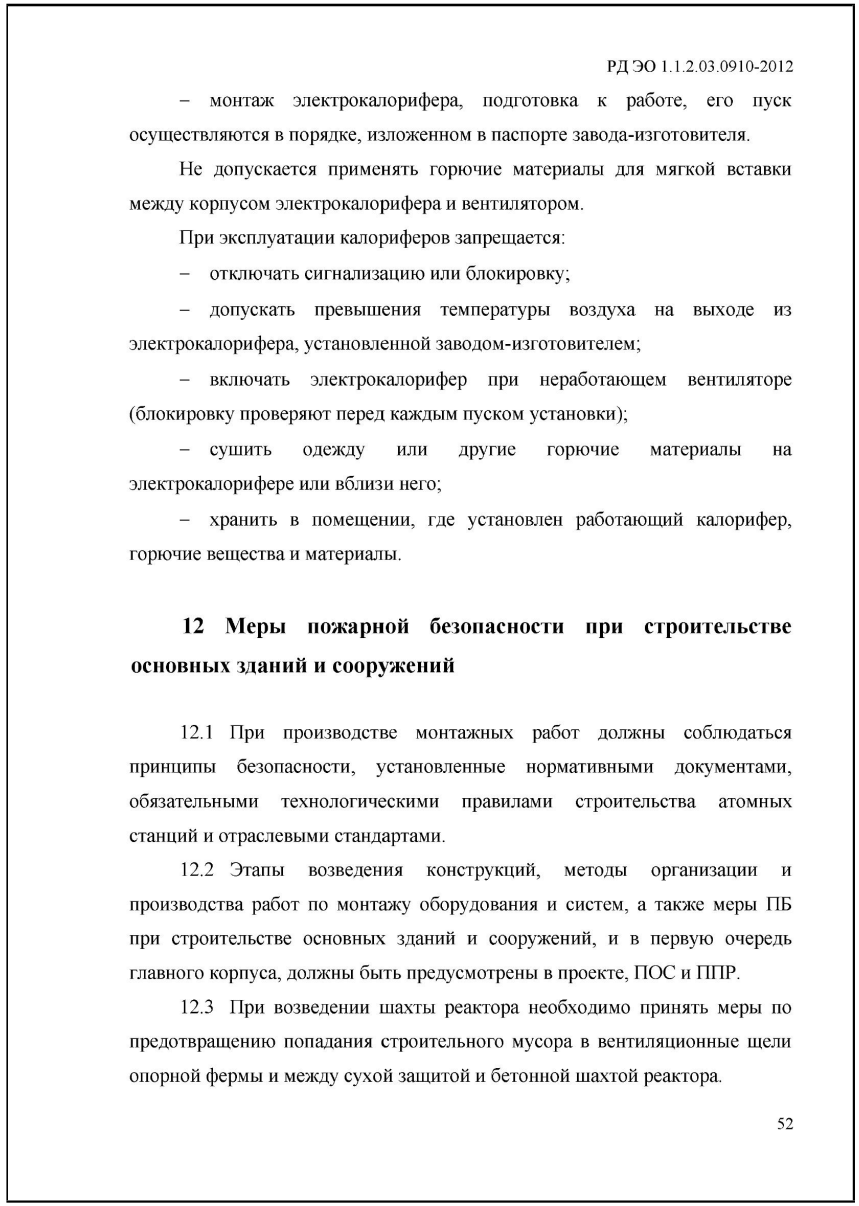


Difficulties in using existing solutions

- Documents can contain deeply nested headings or numbered list items.
- As a rule, existing solutions are focused on the extraction of the hierarchy of headers, for our task it is also necessary to extract the elements of the lists.
- In most of the examples given above, the classification of text blocks is carried out, in our task, we should classify each line of the document.

Dataset description

The dataset is a collection of document JPEG images downloaded from zakupki.gov.ru. The images are scanned copies of government procurements documents. Each image considered as a separate document. The dataset didn't include documents containing tables, figures, frames and other non-text elements.



– монтаж электрокалорифера, подготовка к работе, его пуск осуществляются в порядке, изложенном в паспорте завода-изготовителя.

Не допускается применять горючие материалы для мягкой вставки между корпусом электрокалорифера и вентилятором.

При эксплуатации калориферов запрещается:

– отключать сигнализацию или блокировку;

– допускать превышения температуры воздуха на выходе из электрокалорифера, установленной заводом-изготовителем;

– включать электрокалорифер при неработающем вентиляторе (блокировку проверяют перед каждым пуском установки);

– сушить одежду или другие горючие материалы на электрокалорифере или вблизи него;

– хранить в помещении, где установлен работающий калорифер, горючие вещества и материалы.

12 Меры пожарной безопасности при строительстве основных зданий и сооружений

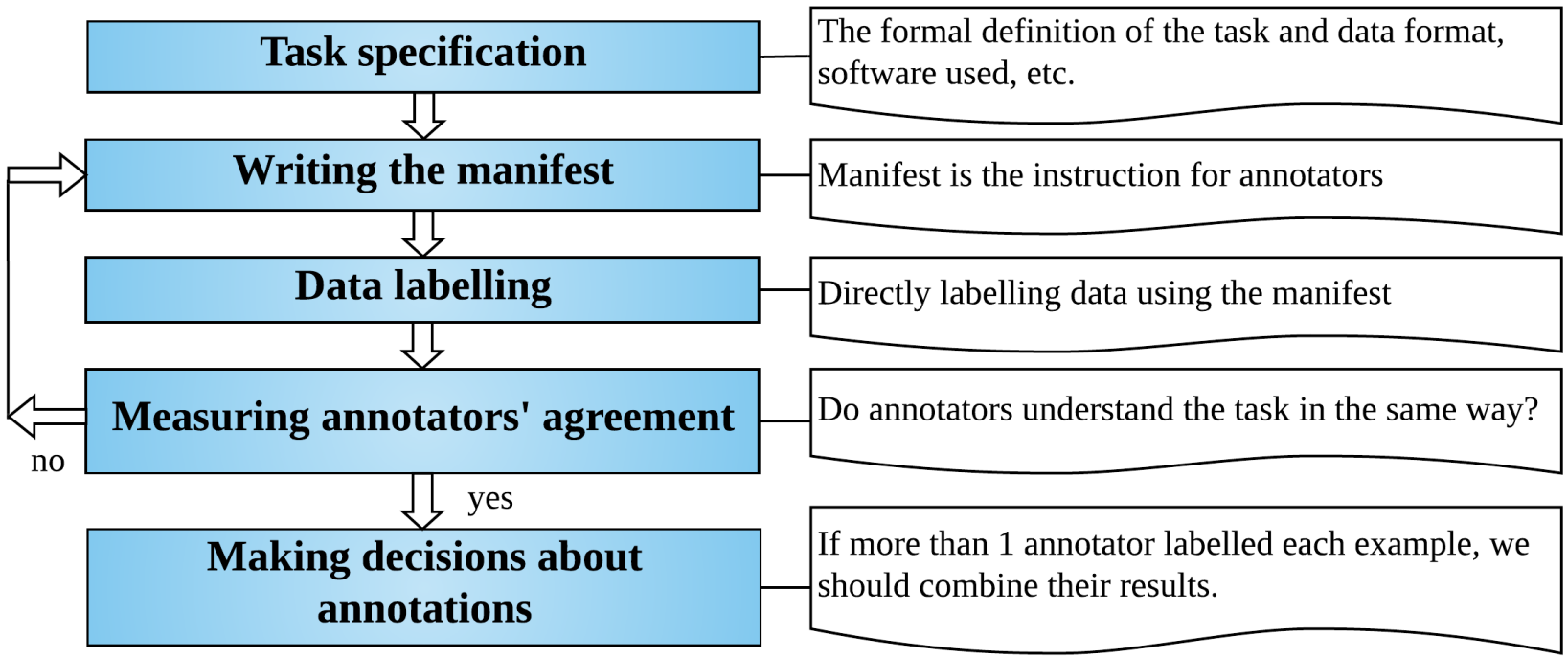
12.1 При производстве монтажных работ должны соблюдаться принципы безопасности, установленные нормативными документами, обязательными технологическими правилами строительства атомных станций и отраслевыми стандартами.

12.2 Этапы возведения конструкций, методы организации и производства работ по монтажу оборудования и систем, а также меры ПБ при строительстве основных зданий и сооружений, и в первую очередь главного корпуса, должны быть предусмотрены в проекте, ПОС и ППР.

12.3 При возведении шахты реактора необходимо принять меры по предотвращению попадания строительного мусора в вентиляционные щели опорной фермы и между сухой защитой и бетонной шахтой реактора.

Data labelling

The process of creating labeled data includes the following steps:



Task specification

Our task is a multiclass classification of document lines. The images of the scanned document with the next line for labelling outlined in a blue frame were sequentially shown to annotators. Annotators should assign the text in the frame to one of the predefined classes:

- header;
- list item;
- text;
- other.

The manifest

To define the labelling rules for annotators, the manifest has been developed. It contains formal rules for classifying document lines.

Манифест

Как правило, документы имеют логическую структуру: название, разбиение на главы, подглавы и т. д., нумерованные и маркированные списки. Мы занимаемся извлечением структурных элементов из сканированных документов. Выделение такой логической структуры документа может пригодиться для автоматизированного анализа документов. Мы хотим решать эту задачу как задачу классификации, нам нужно для каждой строки текста определить, к какому типу она относится.

Мы выделяем следующие типы строк: заголовок, элемент списка, текст.

На вход вам будут подаваться документы, в которых выделена прямоугольником одна строка. Вам необходимо для каждой выделенной строки документа определить её тип. Необходимо «Заголовок» пометить цифрой 1, «Список» - 2, «Текст» - 3, «Другое» - 4.

1. Заголовок

   Название главы, секции, подглавы, параграфа. Строка помечается заголовком, если:

   - текст визуально (полностью) выделяется жирностью;

     15.3 Проведение временных огневых работ

   - текст полностью выделяется шрифтом (курсив, подчеркнутый, другой шрифт, другой размер шрифта);

     - *Привести в текстовой части*

     при этом если текст строки выделен шрифтом частично, то заголовком это не считается;

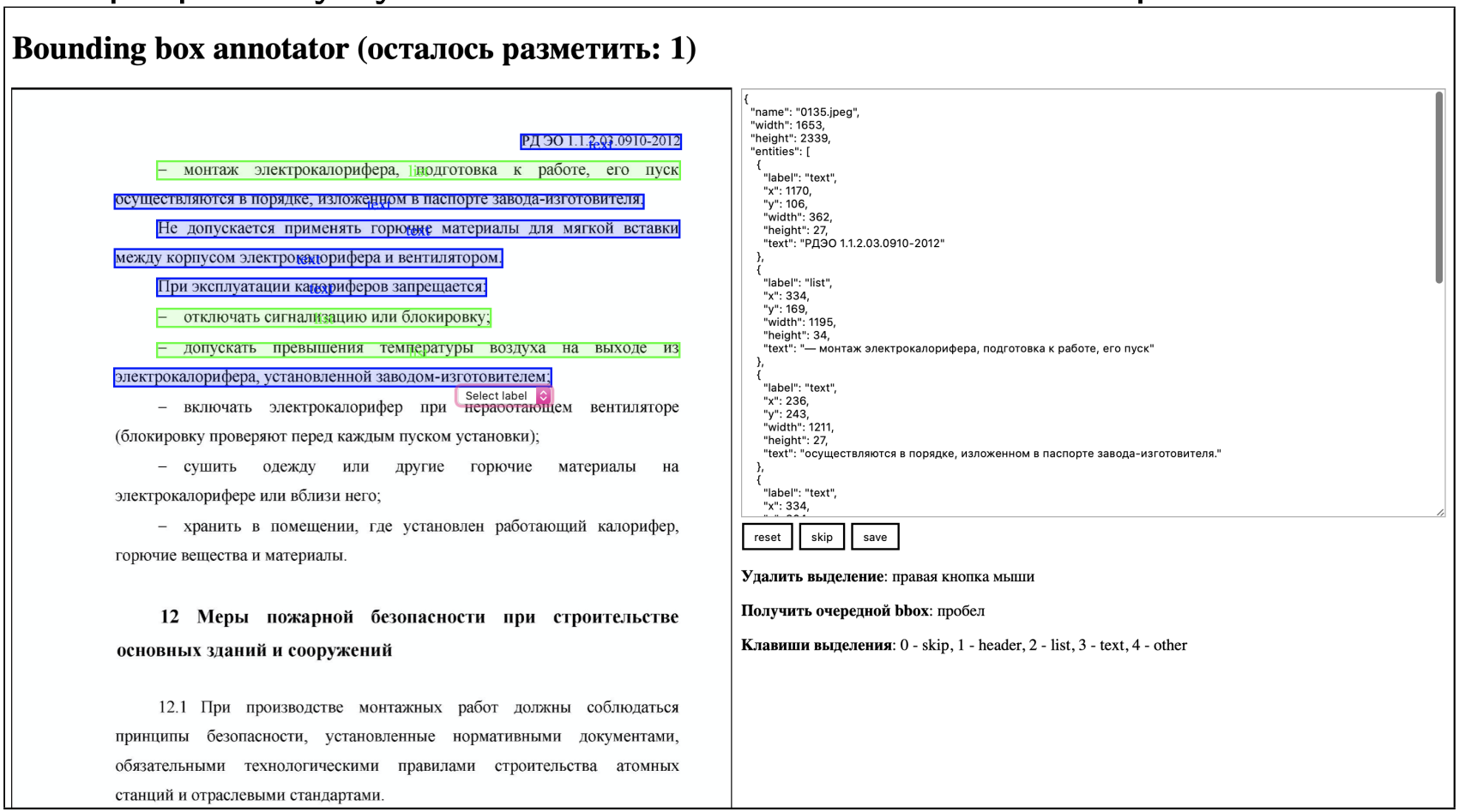
     3.44 пожарное депо: Объект пожарной охраны, в котором

   - текст выделяется отступом (расположен по центру);

     Технические условия
     на проектирование системы АПС и оповещения людей о пожаре на объектах
     ООО «КАМАЗ-Энерго»

Data labelling

The proprietary system was used for the annotation process.



Annotators' agreement

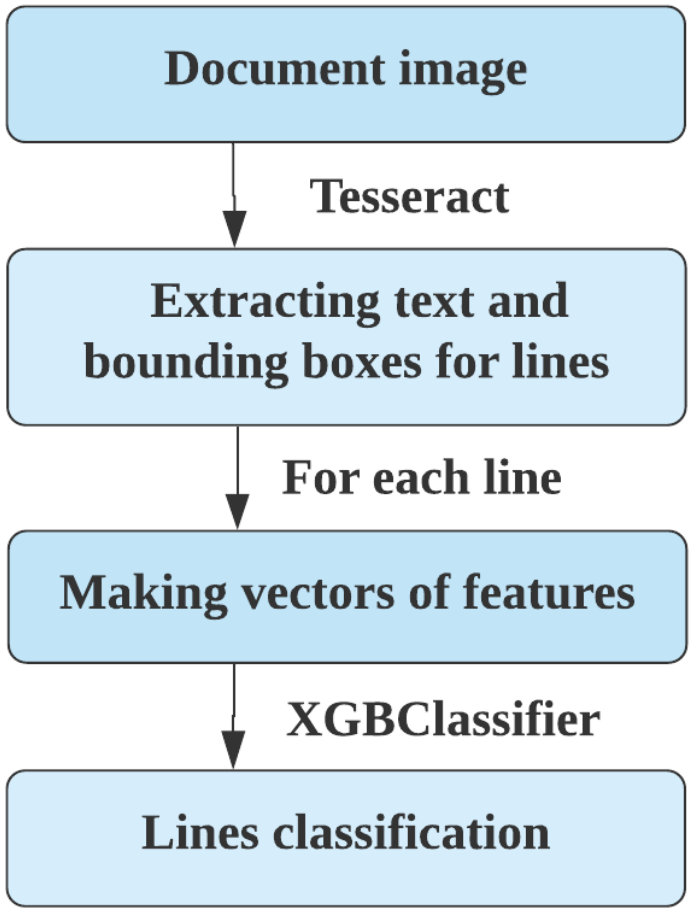
Cohen's kappa κ statistic was calculated to check the correctness of the labelling.

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \kappa \leq 1$$

p_o - the relative observed agreement among raters (accuracy);
p_e - the hypothetical probability of chance agreement.
The closer κ is to 1, the higher the agreement between annotators.
After labelling 10 documents (407 lines) by two annotators, the value of the κ statistic was 0.975, which is considered as a high level of agreement.
Then 600 documents (21350 lines) were labelled.

Method description

Pipeline for documents processing:



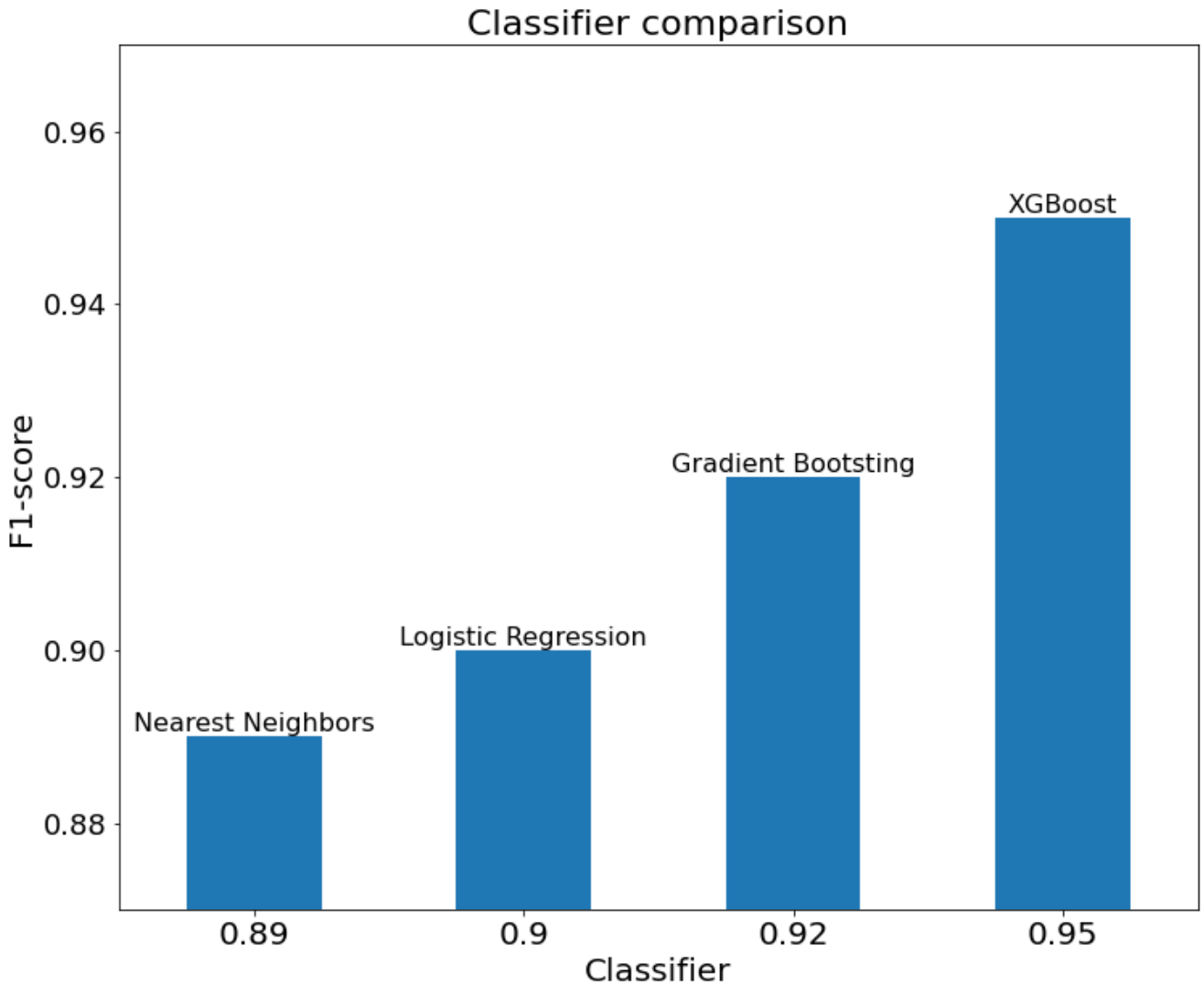
Feature extraction

3 types of features were extracted:

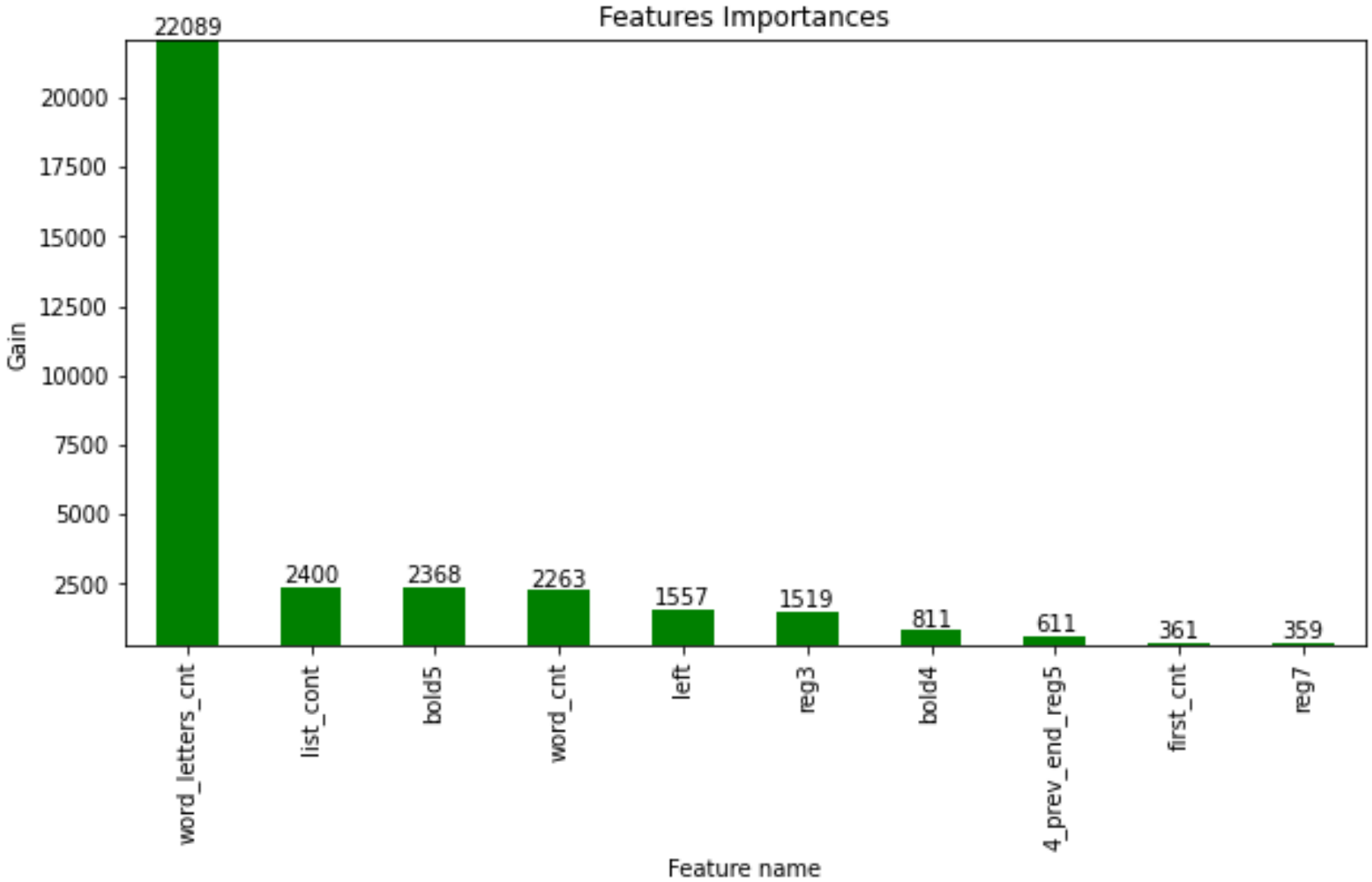
- **Regular expression-based features** (the line starts with an uppercase or lowercase letter, number, dash, etc., the line ends with a letter, period or semicolon).
- **Textual features** (number of letters in the first word, line length, number of words in a line).
- **Visual features** (left indentation, font weight, text height).

In addition, the features of the four previous (subsequent) lines and the average values of some features were used.

Classifier selection



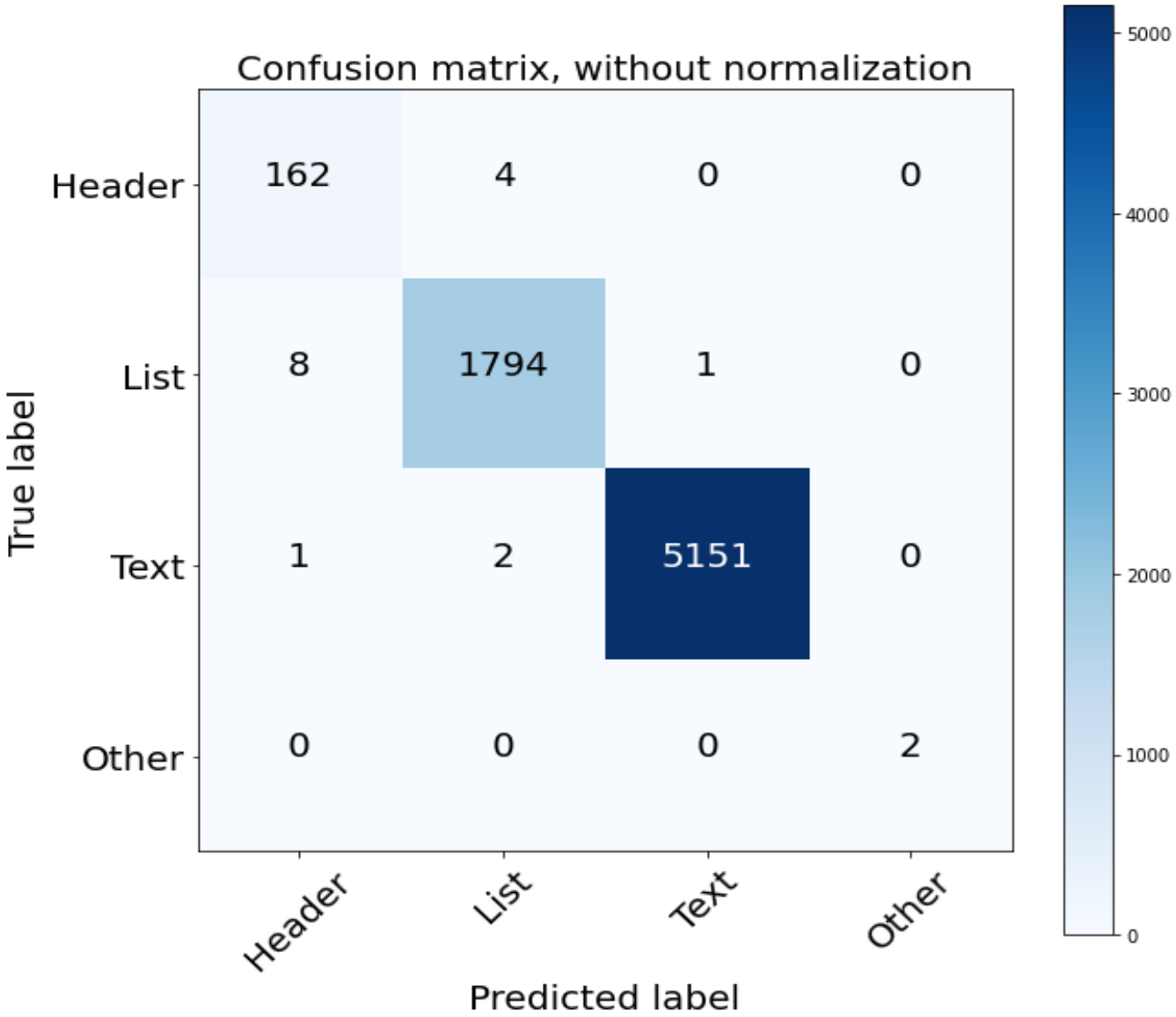
Features' importances analysis



Results

- XGBClassifier parameters after model tuning:
  $learning\_rate = 0.1$
  $n\_estimators = 1000$
  $max\_depth = 7$
  $min\_child\_weight = 2$
  $gamma = 0$
  $subsample = 1$
  $colsample\_bytree = 1$
  $alpha = 0.01$
- F1-score on the cross-validation: 0.98995.

Errors analysis



Conclusion

1. The method for extracting the logical structure based on the classification of document lines is developed.
2. The pipeline is implemented, consisting of processing documents using the Tesseract program for extracting lines and bounding boxes, making feature vectors and training the classifier.
3. The dataset obtained using manual labelling is available.