ISPRAS@FinTOC-2021 Shared Task: Two-stage TOC generation model

Ilya Kozlov **Oksana Belyaeva** **Anastasiya Bogatenkova** **Andrew Perminov**
Ivannikov Institute for System Programming of the RAS
25, Alexander Solzhenitsyn Str., Moscow, 109004, Russia
{kozlov-ilya,belyaeva,nastyboget,perminov}@ispras.ru

Abstract

We propose a two-stage approach for TOC generation from financial documents. This work connected with participation in FinTOC-2021 Shared Task: "Financial Document Structure Extraction". The competition contains two subtasks: title detection and TOC generation. Our model consists of two classifiers: the first binary classifier separates title lines from non-title, the second one determines the title level. In the title detection task, we got 0.813 and 0.787 F1 measure, in the TOC generation task we got 37.9 and 42.1 the harmonic mean between Inex F1 score and Inex level accuracy for English and French documents respectively. With these results, our approach took third place among all submissions. As a team, we took second place in the competition in all categories.

1 Introduction

Currently, electronic documents have become widespread. A large number of documents are presented in a PDF format, but only a few of them contain an automatic table of contents (TOC). However, there may be the need for a quick search of information and it may be a problem for large documents. One example is financial documents, which can be over 100 pages long. Financial documents contain a lot of important information and can have different appearances and structures. The task of automatically extracting the table of contents from financial documents seems to be relevant and its solution is not obvious.

FinTOC-2021 (El Maarouf et al., 2021) offers to solve the problem of extracting structure from financial documents in two languages: English and French. The results of solving two subtasks are evaluated:

- **Title detection (TD)** - selection from all lines of the document only those that should be included in the table of contents;
- **Table of contents (TOC) generation** - identification nesting depths of selected titles.

The competition is held for the third time. Similar tasks were solved at FinTOC-2019 (Juge et al., 2019) and FinTOC-2020 (Bentabet et al., 2020); in 2020, documents in French were added.

In FinTOC-2019, the best solution (Tian and Peng, 2019) for title detection is based on the LSTM with augmentation and attention. The best solution (Giguet and Lejeune, 2019) for the TOC generation task relies on the decision tree classifier DT 10 and TOC page detection.

In FinTOC-2020, the best solution (Hercig and Kral, 2020) for title detection (French) was obtained with the maximum entropy classifier. For title detection in English documents (Premi et al., 2020) LSTM, CharCNN, and a fully connected network with some handcrafted features were used. The best approach for TOC generation (Kosmajac et al., 2020) consisted in extracting linguistic and structural information and using the Random Forest classifier.

In this paper, we describe our solution to the shared task. This work is a continuation of (Bogatenkova et al., 2020). We make a list of features for each document line and use two classifiers for the consequent solution of both title detection and TOC generation tasks.

The paper is organized as follows. We describe in detail the given dataset for the competitions in Section 2. We present our approach to solving the task in Section 3. Results and a discussion are given in Section 4 and 5 respectively. Section 6 contains a conclusion about our work.

2 Dataset

2.1 Train dataset

The training data of the FinTOC-2021 shared task consists of 49 English and 47 French financial PDF documents with a textual layer. The documents

	English	**French**
Number of documents	49	47
Mean number of pages	64	30
Number of extracted lines	191373	79071
Number of TOC	43	4
Mean number of titles	181	142
Max title depth	9	6

Table 1: Training dataset statistics

	English	**French**
Number of documents	10	10
Mean number of pages	66	26
Number of extracted lines	42100	13027
Number of TOC	9	0

Table 2: Test dataset statistics

are very heterogeneous, both groups contain documents with and without TOC. Moreover, not only existing TOC should be included in the result, but also smaller titles of each document.

The main information about the training dataset is in the Table 1. The mean of pages' numbers is 64 and 30 for English and French documents respectfully. But the size of documents varies greatly from 3 to 285 pages. The dataset contains one-column, two-column, and even three-column documents. At the same time, a different number of columns may occur within one document. Moreover, documents are different in their appearance (e. g. the appearance of titles or existing TOC) and logical structure. So, there is no way to extract a complete TOC using regular expressions and we need to use machine learning techniques.

There is a set of annotations for each document. Annotations include only titles with the text and the depth for each title. The number of titles and maximum title depth are also different for each document. The number of titles varies from 20 to 1004 and from 33 to 527 for documents in English and French, respectively. Maximum title depth is from 3 to 9 for English documents while it equals from 4 to 6 for French documents. Thus, a sample of very different documents is presented at the competition.

2.2 Test dataset

The test dataset is similar to the training dataset. It contains 10 documents in each English and French set. The documents are also very diverse, none of the French documents contain a table of contents. More statistics are shown in the Table 2.

3 Proposed approach

We proposed the 2-stage method (see figure 1) for solving the both tasks TD and TOC generation. Each stage includes classification using the XGBoost classifier. In the first stage, the binary classifier classifies each line as title or non-title. Thus, the first stage is a process of filtering all lines of the document. In the second stage of our method hierarchy levels are found for each filtered title from the first stage.

Text and metadata extraction. Since the input PDF documents have a textual layer, we extracted text, bold and italic font, and colors of the text with help of PDFMiner (Yusuke Shinyama, 2019). PDFMiner has different layout reading modes. To read the entire document we have chosen the universal layout mode for multi-column documents with parameters *LAParams(line_margin=1.5, line_overlap=0.5, boxes_flow=0.5, word_margin=0.1, detect_vertical=False)* . Thus the list of lines with text and metadata is extracted from the input documents. To obtain lines with labels we matched the provided labelled titles and the extracted lines using a Levenshtein distance with 0.8 threshold.

As preprocessing, we remove footers and headers from a document using the method (Lin, 2003). It helps to improve the quality of the binary classifier and the TOC extration module. The problem with headers and footers is that they are similar to titles and can predicted as the element of TOC.

Existing TOC extraction. As additional information, we separately extract a table of content (TOC) for each document. We look for the keywords of the TOC heading in the document (for example, "Table of contents", "CONTENT") as the beginning of TOC. Then, we detect the TOC's body using regular expressions.

Most tables of contents in the given documents are one-column regardless of the number of columns in the whole document. The TOC extraction module requires PDFMiner to be run in the single column mode because the TOC text may be read automatically as a multi-column. In this case, PDFMiner should be run with the parameters *LAParams(line_margin=3.0, line_overlap=0.1, boxes_flow=-1, word_margin=1.5, char_margin=100.0, detect_vertical=False)*.

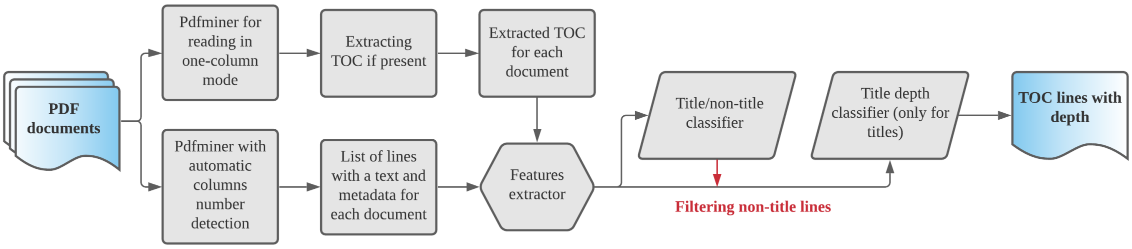
PDF documents
Pdfminer for reading in one-column mode
Pdfminer with automatic columns number detection
Extracting TOC if present
List of lines with a text and metadata for each document
Extracted TOC for each document
Features extractor
Title/non-title classifier
Filtering non-title lines
Title depth classifier (only for titles)
TOC lines with depth

Figure 1: Full pipeline description

Features extraction. The list of extracted lines and extracted TOCs (if present) are processed to obtain a vector of features for each extracted line. We formed a vector from 184 features some of which are enlisted below:

- **visual features**: bold or italic text, indentation, spacing between lines, normalized font size, text color;
- **features from letter statistics**: the percentage of letters, capital letters, numbers, the number of words in line, normalized line's length;
- **features from regular expressions for lines beginning**: indicators of matching some regular expressions for list items;
- **features from regular expressions for lines end**: indicators of ending with a dot, colon, semicolon, comma;
- **features connected with lines depth**: the level of numbering for list with dots (like 1.1.1), relative font size and indentation;
- **TOC features**: indicator of being in the existing TOC (we extracted it automatically), indicator of being the TOC header;
- **other features**: normalized page number and line number;
- **context features**: the same features for 3 previous and 3 next lines.

Training process and experiments. For both tasks we experimented with two models: one-stage and two-stage classifiers. Under the one-stage model we call the model without the first stage (without the binary classifier). In this case the input lines are not pre-filtered. We use the XGBoost classifier in both models. The training process ran with parameters 0.1 learning rate and 100 estimators. We use 3-fold cross-validation for evaluate the results of each model. The mean results for English documents are given in the Table 3. The evaluation script is provided by the organizers (El Maarouf et al., 2021).

Model type	F1 (TD)	Inex08-F1 (TOC)
XGBoost 1-stage	0.77	50.5
XGBoost 2-stage	0.81	55.7

Table 3: The results from cross-validation on the training dataset (English)

Team run	F1 (English)	F1 (French)
Christopher Bourez2	0.830	0.818
Christopher Bourez1	0.822	0.817
ISP RAS (our)	**0.813**	**0.787**
Yseop Lab	0.728	0.639
Cilab_fintoc2	0.514	–
NovaFin	0.507	0.562
Daniel	0.465	0.606
Cilab_fintoc1	0.456	–

Table 4: Title Detection Competition results

The two-stage model performed better than the single-stage one. Thus, we've chosen two-stage model for solving the task on the test dataset.

4 Results

The competition results on test dataset (see table 2) are presented in the table 4 (Title Detection), and tables 5, 6 (TOC generation). Our approach ranks third among 8 and 6 submitted solutions for English and French documents respectfully. As a team we took the second place in all categories.

5 Discussion

The two-stage model demonstrates high scores for both tasks. But the model has disadvantages. Pri-

Team run	**Inex08-P**	**Inex08-R**	**Inex08-F1**	**Inex08-Title acc**	**Inex08-Level acc**	**harm mean**
Christopher Bourez2	55.4	52.6	53.6	60.3	30.6	53.6
Christopher Bourez1	53.3	52	52.5	59	36.5	52.5
ISP RAS (our)	**51.1**	**45.3**	**47.6**	**55.6**	**31.5**	**37.9**
Yseop Lab	61.1	50.3	53.4	68.2	12.4	20.1
Cilab_fintoc2	30.5	18.6	22.6	38.9	31.4	26.3
NovaFin	22.2	21.5	21.2	35.7	31.4	25.3
Daniel	52.8	18.6	25.1	54.3	0	–
Cilab_fintoc1	26.6	14.4	17.6	34.2	34.8	23.4

Table 5: TOC Generation Competition on English documents

Team run	**Inex08-P**	**Inex08-R**	**Inex08-F1**	**Inex08-Title acc**	**Inex08-Level acc**	**harm mean**
Christopher Bourez2	60.8	54.3	57.3	63.5	38.7	57.3
Christopher Bourez1	60.9	54.2	57.3	63.6	39	57.3
ISP RAS (our)	**52.6**	**38.8**	**44.5**	**53.6**	**39.9**	**42.1**
NovaFin	29.7	24.7	26.7	34.6	32	29.1
Yseop Lab	46.8	28.1	34.4	47.3	16.6	22.4
Daniel	49.7	28.6	35.8	53.7	7.1	11.8

Table 6: TOC Generation Competition on French documents

marily, the model misclassifies questionable titles, the ground truth of which are interpreted differently for different documents. For example, one document has a line with some features (color, font size, style and etc.) as a title, but the equivalent line in another document is not a title. Also, we don't combine adjacent titles together as in the ground truth of the data sets.

As well, a two-stage model accuracy in the title detection task is limited by the binary classifier. If the model filters out the title lines in the first step, it will not be able to determine their depths in the second one. Therefore, the accuracy of the two-stage model will not exceed the accuracy of the binary classifier.

As a development of the work, we propose to consider more advanced and complicated models, e. g. the LSTM model. This model can give greater accuracy through the use of long-term memory. Thus, we will be able to remember the previous predictions made up to this point in the document.

6 Conclusion

We proposed the approach for automatic title detection and TOC generation for PDF financial documents with a textual layer. We extracted lines with metadata using Pdfminer and found existing TOCs using the regular expressions. This information we transform to the feature matrix with the vector of predefined features for each document line. Then we use a two-stage model for solving both title detection and TOC generation problems. First, we filter titles from all document lines using the XGBoost binary classifier. Then, we find the depths of the filtered lines using the second XGBoost classifier. The described approach does not depend on the presence of the table of contents in the document and can be used for both English and French financial documents. As a result, our team has taken second place in all categories in the FinTOC-2021 competition.

References

Najah-Imane Bentabet, Remi Juge, Ismail El Maarouf, Virginie Mouilleron, Dialekti Valsamou-Stanislawski, and Mahmoud El-Haj. 2020. The Financial Document Structure Extraction Shared Task (FinToc 2020). In *The 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (FNP-FNS 2020)*, Barcelona, Spain.

Anastasiya Olegovna Bogatenkova, Ilya Sergeyevich Kozlov, Oksana Vladimirovna Belyaeva, and Andrey Igorevich Perminov. 2020. Logical structure extraction from scanned documents. *Proceedings* *of the Institute for System Programming of the RAS*, 32(4):175–188.

Ismail El Maarouf, Juyeon Kang, Abderrahim Aitazzi, Sandra Bellato, Mei Gan, and Mahmoud El-Haj. 2021. The Financial Document Structure Extraction Shared Task (FinToc 2021). In *The Third Financial Narrative Processing Workshop (FNP 2021)*, Lancaster, UK.

Emmanuel Giguet and Gaël Lejeune. 2019. Daniel@ fintoc-2019 shared task: toc extraction and title detection. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 63–68.

Tomáš Hercig and Pavel Kral. 2020. UWB@FinTOC-2020 shared task: Financial document title detection. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 158–162, Barcelona, Spain (Online). COLING.

Rémi Juge, Imane Bentabet, and Sira Ferradans. 2019. The fintoc-2019 shared task: Financial document structure extraction. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 51–57.

Dijana Kosmajac, Stacey Taylor, and Mozhgan Saeidi. 2020. DNLP@FinTOC'20: Table of contents detection in financial documents. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 169–173, Barcelona, Spain (Online). COLING.

Xiaofan Lin. 2003. Header and footer extraction by page association. In *Document Recognition and Retrieval X*, volume 5010, pages 164–171. International Society for Optics and Photonics.

Dhruv Premi, Amogh Badugu, and Himanshu Sharad Bhatt. 2020. AMEX-AI-LABS: Investigating transfer learning for title detection in table of contents generation. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 153–157, Barcelona, Spain (Online). COLING.

Ke Tian and Zi Jun Peng. 2019. Finance document extraction using data augmentation and attention. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 1–4.

Philippe Guglielmetti Pieter Marsman Yusuke Shinyama. 2019. Pdfminer.six documentation.