Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022, pages 89–94
Marseille, 24 June 2022

ISPRAS@FinTOC-2022 Shared Task: Two-stage TOC Generation Model

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

Abstract

This work is connected with participation in FinTOC-2022 Shared Task: "Financial Document Structure Extraction". The competition contains two subtasks: title detection and TOC generation. We describe an approach for solving these tasks and propose the pipeline, consisting of extraction of document lines and existing TOC, feature matrix forming and classification. Classification 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 0.900, 0.778 and 0.558 F1 measure, in the TOC generation task we got 63.1, 41.5 and 40.79 the harmonic mean of Inex F1 score and Inex level accuracy for English, French and Spanish documents respectively. With these results, our approach took first place among English and French submissions and second place among Spanish submissions. As a team, we took first place in the competition in English and French categories and second place in the competition in Spanish.

Keywords: document structure, TOC generation, machine learning

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-2022 offers to solve the problem of extracting structure from financial documents in three languages: English, French and Spanish. 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 fourth time. Similar tasks were solved at FinTOC-2019 (Juge et al., 2019), FinTOC-2020 (Bentabet et al., 2020), FinTOC-2021 (El Maarouf et al., 2021); in 2020, documents in French were added, and the dataset was supplemented with Spanish documents in 2022.

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 FinTOC-2021, for both title detection and TOC generation task, both English and French languages, the best solution (Bourez, 2021) consisted of statistical features extraction on style properties and using XGBoost classifier to predict the needed information.

We also participated in FinTOC-2021 (Kozlov et al., 2021) and took second place in all subtasks. Our decision also relies on XGBoost classifier, that is used separately for solving title detection and TOC item depth prediction subtasks.

In this paper, we describe enhancements of our previous solution to the shared task. As in (Kozlov et al., 2021), 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. We tried to train the classifiers on all data in three languages, as well as on each language separately. In addition, the selection of parameters of the classifiers for each of the subtasks was carried out.

The paper is organized as follows. We describe the given dataset for the competitions and compare it with the previous one in Section 2. We present our approach and its improvement in Section 3. Results and a discussion are given in Section 4 and 5 respectively. Section 6 contains a conclusion about our work.

		English	French	English	French	Spanish
train	Number of documents	49	47	79	81	80
	Mean number of pages	64	30	77	27	119
	Number of TOC	43	4	69	6	74
	Mean number of titles	181	142	225	134	150
	Max title depth	9	6	9	9	7
test	Number of documents	10	10	10	10	10
	Mean number of pages	66	26	102	20	198
	Number of TOC	9	0	8	2	9
		2021		*2022*		

Table 1: Training and test datasets' statistics for 2021 and 2022

Figure 1: Examples of TOCs in Spanish documents

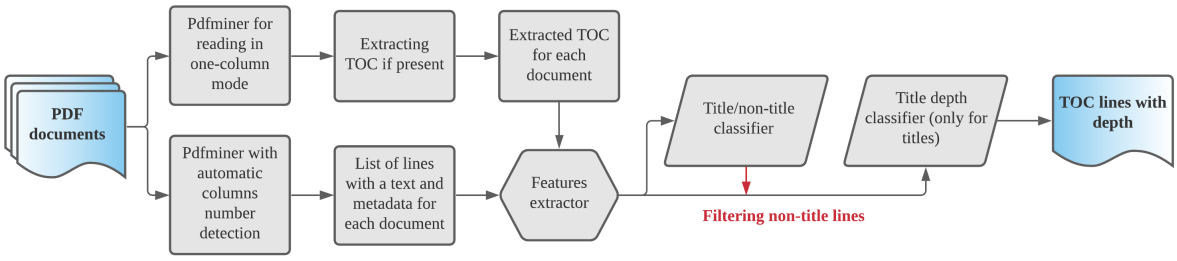

Figure 2: Full pipeline description

2. Datasets

The training data of the FinTOC-2022 shared task consists of 71 English, 81 French and 80 Spanish financial PDF documents with a textual layer. The documents are very heterogeneous, all groups contain documents with and without TOC.

The main information about the datasets of 2021 and 2022 is in the Table 1. Disregarding Spanish documents, the number of documents almost doubled in comparison with the previous year. 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. We should especially mention Spanish documents, which are extremely difficult to parse due to the variety of layouts. Almost all the Spanish documents have a table of contents, still, these TOCs greatly differ from one to another (Figure 1).

There is a set of annotations for each document in the training set. Annotations include only titles with the text and the depth for each title. The number of titles and maximum title depth are different for each document. The number of titles varies from 20 to 1036, from 12 to 527, from 0 to 468 for documents in English, French and Spanish, respectively. Maximum title depth is from 1 to 9 for English and French documents, while it equals from 1 to 7 for Spanish documents. Thus, samples of very different documents are presented at

Features group	Description	Type
Visual	Colour (red, green, blue) and colour dispersion	float
	Font style (bold, italic)	bool
	Indentation, spacing between lines, font size (normalized)	float
Letter, words, and line statistics	The percentage of letters, capital letters, numbers, brackets in a line	float
	The number of words in a line	int
	Normalized page number, line number, and line length	float
TOC	Indicator, if non-empty TOC was extracted for the given document	bool
	Indicator, if the given line is the part of TOC (the page of this line is the page where TOC is located)	bool
	Indicator, if the line is a header included in TOC (the page of this line is mentioned in TOC)	bool
Textual	Indicators, if the line matches regular expressions for different lists like 1), a), I., 1., i), –, etc.	bool
	Indicator, if the line ends with a dot, colon, semicolon, comma	bool
Window bound features	If the line is a list item, the number of predecessors and predecessors with the same indentation in the window of sizes 10, 25, 100 (normalized by the length of the window)	float
	The number of lines with the same indentation in the window of sizes 10, 25, 100 (normalized by the length of the window)	float
Lines depth	The level of numbering for list with dots (like 1.1.1), relative font size and indentation	int
Contextual	The aforesaid features for 3 previous and 3 next lines	float, int, bool

Table 2: Features description

the competition.
The test dataset is similar to the training dataset. It contains 10 documents for each language.

3. Proposed approach

As in the previous year (Kozlov et al., 2021), we propose the 2-stage method for solving the both tasks TD and TOC generation (Figure 2). Each stage includes classification using the XGBoost classifier.

1. The binary classifier classifies each line as title or non-title.

2. For each filtered title from the first stage, its depth is found using the second multiclass classifier.

The main steps of our algorithm are described below.

3.1. Text and metadata extraction.

We extracted text, bold and italic font, colours, etc. of the text with help of PDFMiner (Yusuke Shinyama, 2019), which 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 extraction module. Moreover, we delete empty lines because they aren't useful for our target result.

3.2. 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=0.5, word_margin=1.5, char_margin=100.0, detect_vertical=False)*.

3.3. 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 197 features, some of which are grouped and described in the Table 2.

Option name	Binary classifier			Depth classifier		
	En	Fr	Sp	En	Fr	Sp
learning_rate	0.25	0.1	0.25	0.07	0.4	0.25
max_depth	5	5	4	4	5	3
n_estimators	400	800	600	800	800	600
colsample_bynode	0.8	0.5	0.5	1	1	0.5
colsample_bytree	0.5	0.8	0.5	1	0.5	1
tree_method	*hist*	*approx*	*approx*	*hist*	*exact*	*hist*

Table 3: The resulting classifiers parameters

Model type	TD			TOC		
	En	Fr	Sp	En	Fr	Sp
ISP RAS1	0.79	0.74	0.57	55.8	45.4	42.9
ISP RAS2	0.81	0.73	0.58	57.7	43.4	41.8

Table 4: The mean results from cross-validation on the training dataset

3.4. Classification

For both tasks, we experimented with the XGBoost classifier. During training, we fed the classifiers with different data:

1. **ISP RAS1** – for each language, classifiers were trained only on the documents of that language. Namely, classifiers for English documents were trained only on English documents, etc.

2. **ISP RAS2** – for each language, classifiers were trained on the documents of all available languages.

While training separate classifiers for each language, we selected the best classifiers' options. We tried the grid of possible parameters combinations and found options that gave the highest score. The resulting options are enlisted in the Table 3.
Due to the lack of time, during training classifiers on documents of all languages, we also used the parameters shown in the Table 3.
We use 3-fold cross-validation for evaluate the results of each model. The mean results for both experiments (ISP RAS1 and ISP RAS2) are given in the Table 4. The evaluation script is provided by the organizers.

4. Results

The competition results on test dataset are presented in the table 5 (Title Detection), and tables 6, 7, 8 (TOC generation). In addition, the best three results of the previous year were added. Our approach ranks first among submitted solutions in 2022 for English and French documents, and second for Spanish documents.

5. Discussion

The two-stage model demonstrates high scores for both tasks. But the model has disadvantages. Primarily, the model misclassifies questionable titles, the ground truth

Team run	F1 (EN)	F1 (FR)	F1 (SP)
Christopher B.1	0.822	0.817	–
Christopher B.2	0.830	0.818	–
ISP RAS (2021)	0.813	0.787	–
CILAB	0.738	0.304	0.077
GREYC1	0.790	0.669	0.196
GREYC2	0.793	0.671	0.206
ISP RAS1	**0.900**	**0.778**	0.554
ISP RAS2	0.876	0.758	0.558
swapUNIBA	0.838	0.695	**0.569**

Table 5: Title Detection Competition results

of which are interpreted differently for different documents. For example, one document has a line with some features (color, font size, style, 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. Empty lines, headers and footers were removed from consideration. Extracted lines were transformed to the feature matrix with the vector of predefined features for each line. Then we used a two-stage model for title detection and TOC generation. 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. Optimal parameters for the classifiers were found to improve the

Team run	Inex08-P	Inex08-R	Inex08-F1	Inex08-Title acc	Inex08-Level acc	harm mean
Christopher Bourez1 (2021)	53.3	52	52.5	59	36.5	52.5
Christopher Bourez2 (2021)	55.4	52.6	53.6	60.3	30.6	53.6
ISP RAS (2021)	51.1	45.3	47.6	55.6	31.5	37.9
CILAB	56.2	57.4	56.5	70.7	27.5	36.99
GREYC1	44.0	42.1	42.5	51.3	0.1	0.19
GREYC2	44.6	42.3	42.8	51.7	0.1	0.19
ISP RAS1	**76.3**	**67.2**	**71.3**	**77.5**	55.1	62.16
ISP RAS2	75.2	63.8	68.8	76.8	**58.4**	**63.17**
swapUNIBA	61.4	66.4	63.6	71.4	42.9	51.23

Table 6: TOC Generation Competition on English documents

Team run	Inex08-P	Inex08-R	Inex08-F1	Inex08-Title acc	Inex08-Level acc	harm mean
Christopher Bourez1 (2021)	60.9	54.2	57.3	63.6	39	57.3
Christopher Bourez2 (2021)	60.8	54.3	57.3	63.5	38.7	57.3
ISP RAS (2021)	52.6	38.8	44.5	53.6	39.9	42.1
CILAB	34.9	6.7	11.2	35.5	15.2	12.89
GREYC1	25.8	20.9	22.8	29.1	4.3	7.23
GREYC2	26.0	20.9	22.9	29.3	4.1	6.95
ISP RAS1	52.7	**39.2**	**44.5**	53.7	34.6	38.93
ISP RAS2	**53.2**	38.1	43.9	**54.3**	**39.5**	**41.58**
swapUNIBA	40.0	37.0	38.3	43.8	30.7	34.08

Table 7: TOC Generation Competition on French documents

Team run	Inex08-P	Inex08-R	Inex08-F1	Inex08-Title acc	Inex08-Level acc	harm mean
CILAB	14.8	3.8	4.9	23.8	36.2	8.63
GREYC1	11.4	15.8	6.5	36.0	4.2	5.10
GREYC2	11.7	15.9	6.9	36.1	4.2	5.22
ISP RAS1	**51.6**	35.4	39.4	68.5	42.3	40.79
ISP RAS2	**51.6**	36.9	39.9	**69.1**	40.1	39.99
swapUNIBA	31.8	**59.0**	**40**	65.5	**46.5**	**43.00**

Table 8: TOC Generation Competition on Spanish documents

results, and we used different techniques to train classifiers. The described approach can be used for documents in any language. As a result, our team has taken first place in all categories for English and French documents, and second place for Spanish documents.

7. Bibliographical References

Bentabet, N.-I., Juge, R., El Maarouf, I., Mouilleron, V., Valsamou-Stanislawski, D., and El-Haj, M. (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.

Bourez, C. (2021). Fintoc 2021-document structure understanding. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 89–93.

El Maarouf, I., Kang, J., Aitazzi, A., Bellato, S., Gan, M., and El-Haj, M. (2021). The Financial Document Structure Extraction Shared Task (FinToc 2021). In *The Third Financial Narrative Processing Workshop (FNP 2021)*, Lancaster, UK.

Giguet, E. and Lejeune, G. (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.

Hercig, T. and Kral, P. (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), December. COLING.

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

Kosmajac, D., Taylor, S., and Saeidi, M. (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), December. COLING.

Kozlov, I., Belyaeva, O., Bogatenkova, A., and Perminov, A. (2021). Ispras@ fintoc-2021 shared task: Two-stage toc generation model. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 81–85.

Lin, X. (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.

Premi, D., Badugu, A., and Sharad Bhatt, H. (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), December. COLING.

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

Yusuke Shinyama, P. G. . P. M. (2019). Pdfminer.six documentation.