Извлечение иерархической логической структуры из текстовых документов в формате docx

---

Автор: Богатенкова Анастасия Олеговна

Научный руководитель: Козлов Илья Сергеевич

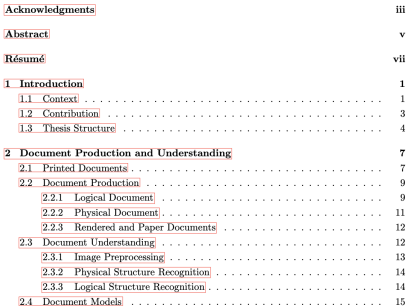
22 декабря 2020 г.

ИСП РАН

Логическая структура документа

ИСП РАН

Логическая структура документа несет в себе некоторый смысл. Например, научные статьи состоят из аннотации, введения, обзора существующих работ и других секций.



Acknowledgments iii

Abstract v

Résumé vii

1 Introduction 1
- 1.1 Context . . . 1
- 1.2 Contribution . . . 3
- 1.3 Thesis Structure . . . 4

2 Document Production and Understanding 7
- 2.1 Printed Documents . . . 7
- 2.2 Document Production . . . 9
  - 2.2.1 Logical Document . . . 9
  - 2.2.2 Physical Document . . . 11
  - 2.2.3 Rendered and Paper Documents . . . 12
- 2.3 Document Understanding . . . 12
  - 2.3.1 Image Preprocessing . . . 13
  - 2.3.2 Physical Structure Recognition . . . 14
  - 2.3.3 Logical Structure Recognition . . . 14
- 2.4 Document Models . . . 15

Актуальность темы

ИСП РАН

Для автоматизированного анализа документов полезно уметь извлекать логическую структуру из документов.

id = 0 ; type = root

Пример документа id = 0.0 ; type = raw_text

Глава 1 id = 0.1 ; type = named_header

Какие то определения id = 0.1.0 ; type = raw_text

Статья 1 id = 0.1.1 ; type = named_header

Определим опрделения id = 0.1.1.0 ; type = raw_text

Статья 2 id = 0.1.2 ; type = named_header

Дадим пояснения id = 0.1.2.0 ; type = raw_text

id = 0.1.2.1 ; type = list

1.2.1. Поясним за непонятное id = 0.1.2.1.0 ; type = list_item

1.2.2. Поясним за понятное id = 0.1.2.1.1 ; type = list_item

id = 0.1.2.1.1.0 ; type = list

а) это даже ежу понятно id = 0.1.2.1.1.0.0 ; type = list_item

б) это ежу не понятно id = 0.1.2.1.1.0.1 ; type = list_item

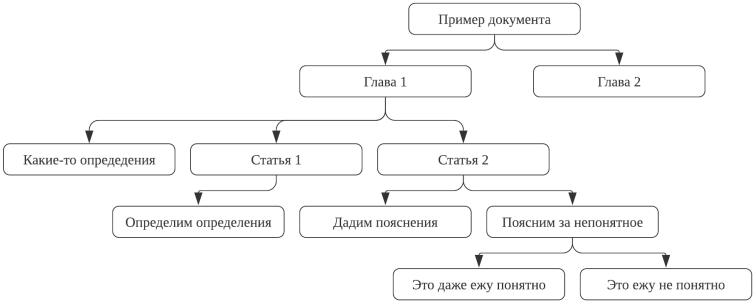
1.2.3. id = 0.1.2.1.2 ; type = list_item

id = 0.1.2.1.2.0 ; type = raw_text

Иерархическая логическая структура

ИСП РАН

Документ можно представить в виде последовательности вложенных друг в друга элементов.



Особенности docx

- В настоящее время большое количество документов создаётся и хранится в формате docx.
- Формат позволяет описывать только физическую структуру документа.
- Библиотеки по работе с форматом существуют, однако не позволяют выделить в документах все стили и текст нумерации элементов списков.

Постановка задачи

Что уже сделано:

- Провести обзор некоторых форматов документов;
- Провести обзор способов представления логической структуры документа;
- Описать особенности формата docx;
- Реализовать извлечение текста и необходимых метаданных из документов в формате docx;
- Создать обучающий набор документов и осуществить его разметку (в процессе).

Извлечение метаданных

ИСП РАН

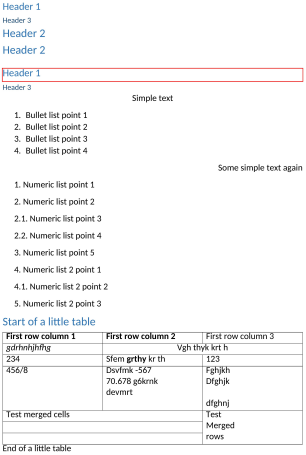
Пример извлечения метаданных для параграфа документа


```
"node_id": "0.0",
"text": "Header 1",
"annotations": [
  {
    "start": 0,
    "end": 8,
    "name": "indentation",
    "value": "0"
  },
  {
    "start": 0,
    "end": 8,
    "name": "alignment",
    "value": "left"
  },
  {
    "start": 0,
    "end": 8,
    "name": "size",
    "value": "16.0"
  },
  {
    "start": 0,
    "end": 8,
    "name": "style",
    "value": "heading 1"
  }
],
```

Разметка

ИСП РАН

Страницы docx документа преобразуются в изображения с параграфами, обведенными в рамку.



Header 1

Header 3

Header 2

Header 2

Header 1

Header 3

Simple text

1. Bullet list point 1
2. Bullet list point 2
3. Bullet list point 3
4. Bullet list point 4

Some simple text again

1. Numeric list point 1
2. Numeric list point 2
2.1. Numeric list point 3
2.2. Numeric list point 4
3. Numeric list point 5
4. Numeric list 2 point 1
4.1. Numeric list 2 point 2
5. Numeric list 2 point 3

Start of a little table

First row column 1	**First row column 2**	First row column 3
gdrhnhjhfhg	Vgh thyk krt h	
234	Sfem **grthy** kr th	123
456/8	Dsvfmk -567 70.678 g6krnk devmrt	Fghjkh Dfghjk dfghnj
Test merged cells		Test Merged rows

End of a little table

Постановка задачи

ИСП РАН

Что еще предстоит сделать:

- Описать структуру, которую необходимо извлечь;
- Провести тестирование реализованного метода обработки docx документов;
- Реализовать метод извлечения структуры и провести экспериментальную проверку реализованного метода;
- Провести оценку качества.