



Извлечение логической структуры из сканированных документов

- 
- Как правило, документы имеют логическую структуру: название, разбиение на главы, подглавы и т. д., нумерованные и маркированные списки.
 - Мы занимаемся извлечением такой структуры.
 - Произвольную структуру извлекать довольно сложно, поэтому необходимо задать определенные правила, ограничить то, что мы собираемся извлекать.



Мы будем выделять следующие типы строк:

- **Заголовок**

Главы, секции, подглавы, параграфы ... -
раздел документа с названием

- **Список**

Начало нумерованного или
маркированного списка. К списку
относятся все пронумерованные любым
способом строки.

- **Текст**

Все остальное считается текстом.

Постановка задачи

- Будем решать задачу как многоклассовую классификацию - классифицировать каждую строчку документа как заголовок, элемент списка или текст и определять её уровень вложенности.

2. Цель выполнения работы, наименование и индекс системы

2.1. Целью работы является доработка Исполнителем конструкторской

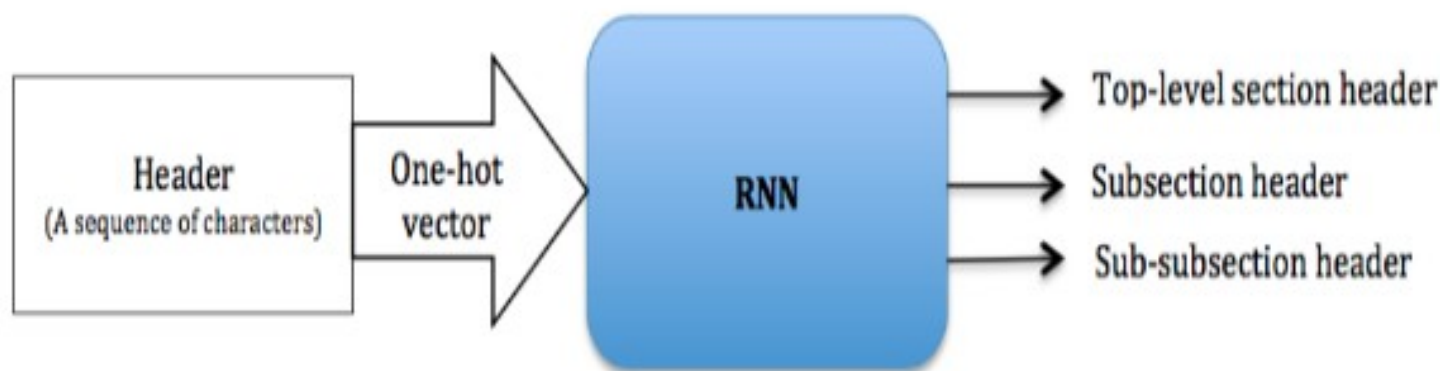
документации интегрированной мостиковой системы ходового командного пункта по техническим условиям ИСПОЛНИТЕЛЯ для заказа 11442М (зав. №802), решение вопросов электрического и конструктивного сопряжения с корабельными системами, определение состава (комплектации) доработанной аппаратуры для ее установки на заказе, а также обеспечение ОАО «Северное ПКБ» необходимой документацией для разработки технического проекта, рабочей конструкторской, приёмо-сдаточной и эксплуатационной документации.

2.2. Наименование: Интегрированная мостиковая система ходового командного пункта.

2.3. Индекс: ИМС-11442М.

Существующие подходы

- по извлечению структуры из документов существуют несколько подходов:
 - на основе оглавления
 - на основе правил
 - на основе машинного обучения.



- Существующие решения использовать сложно, так как может попасться глубокая вложенность, например 1.1.1.1 (уровней может быть и больше), а внутри этих уровней могут располагаться нумерованные/маркированные списки, также вложенные.

3.2.10. АРМ флагмана должно обеспечивать выполнение следующих функций:

3.2.10.1. Отображение информации об обстановке и решение задач на АРМ-КС из состава БИУС «Сигма-11442М»;

3.2.10.2. Осуществление телефонной радиосвязи.

Что было сделано

- Пока проводилась только классификация строк по типам без определения уровня вложенности.
- Выделение структуры на основе правил. С помощью регулярных выражений анализировалось начало каждой строки.
- Pipeline на основе машинного обучения. Ручная разметка. Метрика — f-measure (macro).

<http://10.10.17.71:8888/>



Планы на январь-апрель

- Разметка дополнительного корпуса
- Улучшение pipeline машинного обучения
 - выделение дополнительных признаков на основе отступов, шрифта и т. д.
- Уточнение описания уровней вложенности
- Определение уровня вложенности

- по анализу документов проводится очень много соревнований на ICDAR, например, в 2013 году они проводили соревнования по извлечению структуры из книг (страницы, параграфы, главы и т. д.) - что-то похожее на нашу тему.

The screenshot shows the EOD (Electronic Document) interface. The main window displays a scanned document page with handwritten notes in the top right corner: "F 162 B972 1904". The document is titled "CONTENTS." and lists the following sections:

Section	Page
Editor's Foreword (ms)	7
Preface to the Third Edition	15
Introduction	21
Travel	29
Appendix: No. 1.	
No. 1. Catalogue of Trees, Plants, Birds, Fishes, Animals, &c. mentioned in the Course of this Work; with their common Names, and the Names given them by Catesby and Linnæus.	257
No. 2. Tables and Statements relating to the Commercial Situation of the United States, both before and since the American War	305
No. 3. Anecdotes of the Indians	389
No. 4. — of several Branches of the Fairfax Family, now domiciled in Virginia	397
No. 5. Diary of the Weather	416

The right-hand pane shows the "APPENDIX, N° 1." section, which is a "CATALOGUE of several TREES, PLANTS, BIRDS, FISHES, ANIMALS, &c. mentioned in the course of this Work; with their Common Names, and the names given them by CATESBY and LINNÆUS." The table below lists common names, Catesby names, and Linnæan names.

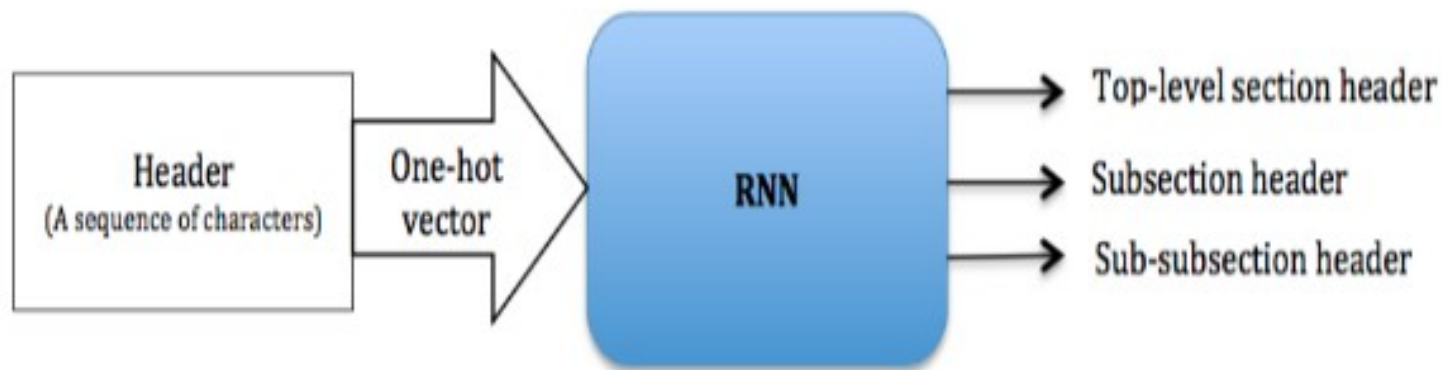
COMMON NAMES.	CATESBY.	LINNÆAN.
Tobacco	Nicotiana.
Maize or Indian Corn	Zea Mays.
Rattle-Snake Root	Aristolochia Pterodactylus, Polygala Senega.
Figwort, Plum or Berry	Cercaria lachrymifera.
Hierony	Nux juglans alba Virginica, Juglans alba, etc.

- в 2019 году проводились соревнования FinTOS, где извлекали структуру из финансовых документов и было два задания - определение заголовков и извлечение оглавления (table of contents).

The screenshot shows a web-based interface for editing document structure. It features a sidebar on the left with a 'Parts' section containing a 'Collapse' button. The main area displays 'Part 1' with its own 'Collapse' and 'Delete Part' buttons. Below this, there are input fields for 'Title' (containing 'IMGP GLOBAL MACRO'), 'Start page' (1), 'End page' (2), 'Topic' (a dropdown menu showing 'Icd'), and 'Confidence' (a dropdown menu showing 'high'). At the bottom, there is a 'Chapters' section with a 'Collapse' button, followed by 'Chapter 1' with 'Collapse' and 'Delete Chapter' buttons. Below Chapter 1, there are input fields for 'Title' (containing 'OBJECTIFS ET POLITIQUE D'INVESTISSEMENT'), 'Start page' (1), and 'End page'.

- В этих соревнованиях структуру извлекал только один участник. Структура документа извлекалась на основе оглавления + использовались правила для нахождения строк в документах (междустрочный интервал, отступ, шрифт и т. д.)

- Есть еще статья 2017 года, структура документа извлекалась с помощью машинного обучения, сначала строки подавались на вход одному классификатору (он определял - заголовок/не заголовок), затем строки-заголовки классифицировались точнее другими классификаторами. В этом решении структура документа имела вложенность 3, то есть предполагалось выделение секций, подсекций, подподсекций.



Уровни вложенности

Заголовки

Ограничимся тремя уровнями вложенности:

- 1 - Глава /параграф /секция /подзаголовки /раздел + нумерация
- 2 - Подглава /подсекция /подпараграф /подраздел + нумерация
- 3 - Подподглава и т. д. + нумерация

Списки

- Вложенные списки с нумерацией 1, 1.1, 1.1.1 и т.д. Выделим такие списки как отдельный класс с неограниченным уровнем вложенности. Уровень таких списков определяется отдельно.



Все остальные виды списков (ограничимся третьим уровнем вложенности)

- Если один список вложен в другой или список относится к блоку текста с заголовком n -го уровня, то уровень вложенности такого списка равен $n+1$.

Просто строки

- Уровень вложенности определяется блоком текста, к которому относится данная строка. Например, если строка - продолжение элемента списка, то ее уровень вложенности соответствует уровню вложенности элемента списка. Если строка располагается после подзаголовка n -го уровня, то ее уровень также равен n .