Московский Государственный Университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Системного Программирования

Курсовая работа

Извлечение логической структуры из сканированных документов

Автор:
группа 328
Богатенкова Анастасия Олеговна

Научный руководитель:
Козлов Илья Сергеевич

Москва, 2020

Содержание

1 Введение

Большое количество текстовой информации представлено в виде pdf-документов, причем эти документы могут представлять собой сканированные копии других документов и у них может отсутствовать текстовый слой. При этом размер документов может быть очень большим. Зачастую требуется осуществлять поиск по содержимому таких документов и желательно осуществлять это более эффективным способом.

Как правило, документы имеют логическую структуру и имеют название, разбиение на главы, подглавы и т. д., содержат нумерованные и маркированные списки. Выделение такой логической структуры документа может помочь при решении задач автоматизированного анализа документов, а также при поиске по документам.

Применяется множество разнообразных подходов, которые позволяют выделять в тексте заголовки и распознавать логическую структуру документов. Для эффективного извлечения такой структуры может быть необходима метаинформация, такая как размер и тип шрифта, отступы, междустрочные интервалы и т. д. Поэтому извлечение логической структуры логично делать на этапе анализа сканированных документов.

2 Постановка задачи

Целью моей курсовой работы является разработка метода выделения логической структуры из документов. Рассмотрим структуру документа в виде глав, подглав (и т. д.), элементов нумерованных и маркированных списков. Поставим задачу следующим образом: необходимо классифицировать каждую строчку документа как заголовок, элемент списка или текст.

При решении задачи можно выделить следующие этапы ее выполнения:

1) Описание конкретной логической структуры, которую нужно выделить, т. е. разработка манифеста.

2) Разметка корпуса документов для обучения классификатора. Разметка проводится по правилам, указанным в манифесте.

3) Реализация метода и проведение экспериментальной проверки разработанного метода.

3 Обзор существующих решений

По извлечению структуры из документов существуют несколько подходов:

- на основе оглавления;
- на основе правил;
- на основе машинного обучения.

3.1 Извлечение структуры из документов на основе оглавления и правил

По анализу документов проводится очень много соревнований на ICDAR. В одном из таких соревнований [1] производилось извлечение структуры из книг, содержимое которых было получено с помощью оптического распознавания символов. Структура книг в виде разбиения на страницы, параграфы, главы извлекалась с использованием оглавления, которое присутствовало в большинстве книг.

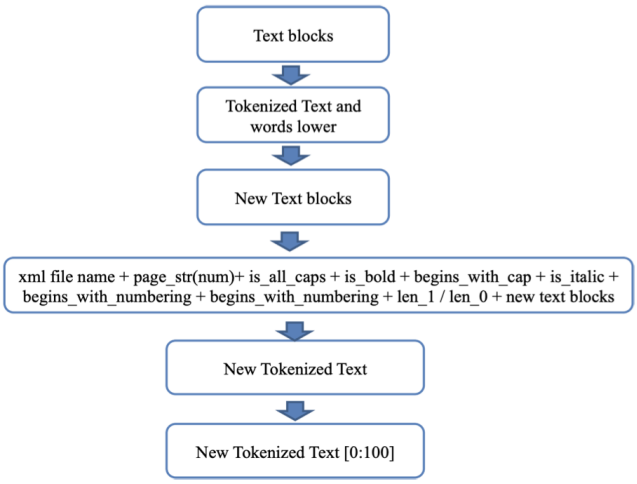
В 2019 году проводились соревнования FinTOC [2], где из финансовых документов извлекалась структура в виде иерархии уровней заголовков документов. Максимальная глубина уровней равна пяти. Одна из команд-участниц [3] извлекала необходимую структуру используя оглавление документов, а также систему правил, которые применялись для определения иерархии заголовков. Сначала идентифицировались страницы, содержащие текст оглавления, затем в документе находились страницы, соответствующие заголовкам, указанным в оглавлении. Последним шагом являлось выделение иерархии найденных заголовков, основанное на применении правил: анализировались такие признаки, как междустрочный интервал, отступ, шрифт, символы нумерации. Использованный подход позволил получить достаточно высокую точность, но низкую полноту, так как некоторые оглавления документов были неполными.

Извлечение структуры документов на основе оглавления имеет ряд недостатков. Во-первых, невозможно обрабатывать документы, в которых нет оглавления. Во-вторых, при использовании этого метода в структуру документа не будут включаться заголовки, которые не вошли в оглавление, например заголовки более низкого уровня. В-третьих, для нашей задачи необходимо извлекать элементы маркированных и нумерованных списков, которые не включаются в оглавление документа.

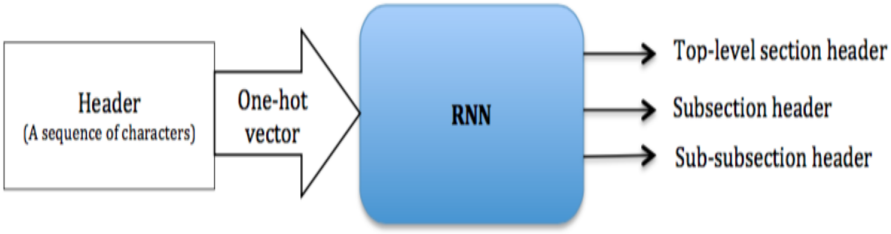
3.2 Извлечение структуры из документов на основе машинного обучения

В соревнованиях [2] кроме извлечения иерархической структуры документов решалась задача определения, является ли конкретный блок документа заголовком. Командам был дан набор pdf-документов, xml-файлов с выделенными блоками документов, а так же набор признаков для каждого блока: является ли шрифт блока жирным, курсивом, состоит ли текст из заглавных букв, начинается с заглавной буквы или с нумерации. Кроме данных признаков, каждая из команд использовала различные дополнительные морфологические, семантические, лингвистические признаки. На основе этих признаков обучались различные классификаторы: SVM, MNB, Extra Tree, Decision Tree, Gradient Boosting. Для оценки результатов использовалась F1-мера, максимальный score в соревновании – 0,982.

Победители соревнования [5] создали новый датасет для обучения с помощью аугментации данных, перевели новые сгенерированные текстовые блоки в векторное представление, а затем использовали рекуррентные нейронные сети LSTM и BiLSTM для решения задачи классификации. Процесс аугментации показан на схеме ниже.



В статье [4] 2017 года структура документа извлекалась с использованием методов машинного обучения, включая глубокое обучение. Цель данной работы – автоматически идентифицировать и классифицировать различные секции документов и понять их смысл в рамках документа (назначить семантическую метку). В рамках моей задачи интересна классификация секций документа.



Классификатор, который был использован при решении задачи, состоит из нескольких частей. Сначала строки документа подаются на вход классификатору (классификатор строк), который определяет, является ли строка заголовоком, затем строки-заголовки классифицируются точнее другими классификаторами (классификаторы секций). В этом решении структура документа имела вложенность 3, то есть предполагалось выделение секций, подсекций, подподсекций. Кроме того, в данной работе был размечен датасет, на котором происходило обучение модели. Метрика качества - F1-мера, при идентификации заголовков итоговый score – 0,96; при классификации секций средний F1-score – 0,81.

Существующие решения использовать сложно, так как, во-первых, документы, с которыми необходимо работать, могут содержать заголовки или элементы нумерованных списков глубокой вложенности, например 1.1.1.1 (уровней вложенности может быть и больше), а внутри этих уровней могут располагаться нумерованные/маркированные списки, также вложенные.

Во-вторых, как правило в существующих решениях основное внимание уделялось выделению структуры в виде заголовков, для нашей задачи требуется также определять элементы списков.

В-третьих, во большинстве примеров, приведенных выше, осуществляется классификация текстовых блоков, в нашей задаче необходимо классифицировать каждую строку документа.

4 Описание практической части

4.1 Описание логической структуры

Для того, чтобы классифицировать каждую строку документа, необходимо определить, по какому принципу конкретная строка будет относиться к тому или иному классу. Это необходимо при осуществлении разметки корпуса документов для обучения классификатора.

С этой целью был разработан манифест, прикрепленный в приложении.

4.2 Подготовка датасета для обучения

5 Заключение

Список литературы

[1] AntoineDoucet, GabriellaKazai, SebastianColutto, Günter Mühlberger. Icdar 2013 competition on book structure extraction. In Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, pages 1438–1443. IEEE, 2013.

[2] Rémi Juge, Najah-Imane Bentabet, Sira Ferradans. FinTOC-2019 Shared Task: Finding Title in Text Blocks.

[3] Gael Lejeune Emmanuel Giguet. Daniel fintoc-2019 shared task: Toc extraction and title detection. In The Second Workshop on Financial Narrative Processing of NoDalida 2019, 2019.

[4] Muhammad Mahbubur Rahman, Tim Finin. Deep nderstanding of a Documents Structure. In 4th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, 2017.

[5] Ke Tian and Zi Jun Peng. Finance document ex- traction using data augmented and attention. In The Second Workshop on Financial Narrative Process- ing of NoDalida 2019, 2019.

6 Приложение

6.1 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*

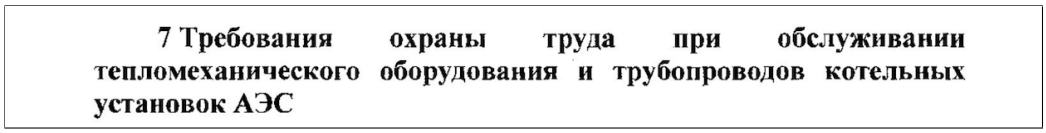
при этом если текст строки выделен шрифтом частично, то заголовком это не считается;

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»

– если заголовок занимает несколько строк, остальные строки тоже относятся к типу «заголовок».

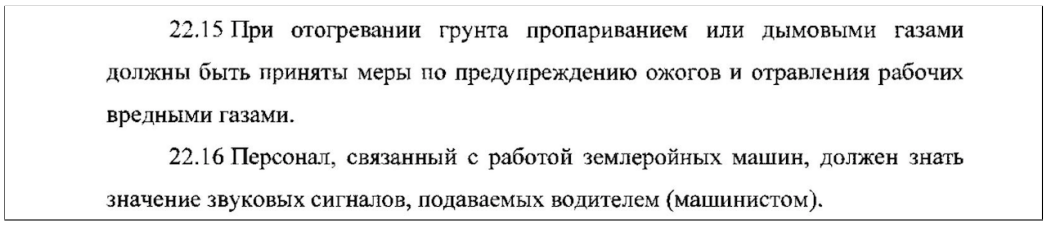


7 Требования охраны труда при обслуживании тепломеханического оборудования и трубопроводов котельных установок АЭС

2) Список

Начало нумерованного или маркированного списка. Строка помечается как элемент списка, если:

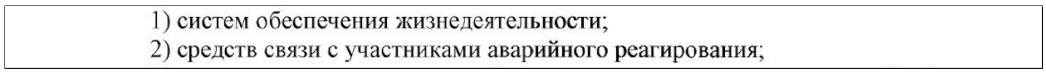
– строка наряду с несколькими другими строками пронумерована («1. 1) а) 1.1» и т. д. в начале строки);



22.15 При отогревании грунта пропариванием или дымовыми газами должны быть приняты меры по предупреждению ожогов и отравления рабочих вредными газами.

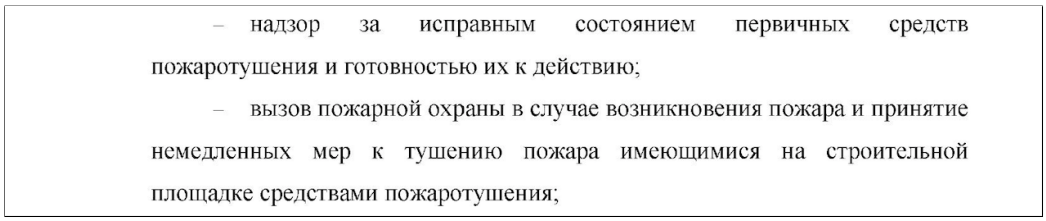
22.16 Персонал, связанный с работой землеройных машин, должен знать значение звуковых сигналов, подаваемых водителем (машинистом).

На картинке выше как элемент списка будут помечены только две строки, остальные помечаются как текст.



1) систем обеспечения жизнедеятельности;
2) средств связи с участниками аварийного реагирования;

– строка наряду с несколькими другими выделена некоторым маркером (точка, тире и т. д.) ;



– надзор за исправным состоянием первичных средств пожаротушения и готовностью их к действию;

– вызов пожарной охраны в случае возникновения пожара и принятие немедленных мер к тушению пожара имеющимися на строительной площадке средствами пожаротушения;

Здесь как элемент списка будут помечены только две строки, остальные помечаются как текст.

– если элемент списка визуально занимает несколько строк, все строки кроме первой помечаются как текст. Также к списку не относятся строки, помеченные как заголовок (выделенные шрифтом, жирностью и т. д.).

3) Текст

Все остальные строки, содержащие текст документа, помечаются как 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. В случае, если прямоугольником выделена строка, не содержащая текст (пустая строка, рукописная подпись, печать), строку необходимо пометить как «Другое» (цифра 4), либо пропустить данную строку при разметке.

Ниже приведен пример разметки. Красными прямоугольниками выделены заголовки, зелёными - элементы списков, а синими - текстовые блоки.