

Извлечение логической структурь из сканированных документов

328 группа ВМК МГУ

Автор:

Богатенкова Анастасия Олеговна

Научный руководитель:

Козлов Илья Сергеевич

28 апреля 2020 г.

Введение

Большое количество текстовой информации представлено в виде pdf-документов, причем у них может отсутствовать текстовый слой.

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



Статья 9. Сроки оказания услуг.

9.1. Общий срок оказания услуг: с даты заключения Договора по 31.12.2019,
но не ранее получения Заказчиком заключения органа государственного
строительного надзора о соответствии реконструированного объекта капитального
строительства требованиям технических регламентов и проектной документации
(ЗОС).

Постановка задачи

Целью моей курсовой работы является разработка метода выделения логической структуры из документов. 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.

Обзор существующих решений

По извлечению структуры из документов существуют
несколько подходов:

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

1. Извлечение структуры из документов на основе оглавления

По анализу документов проводится очень много соревнований на ICDAR. В одном из таких соревнований [1] производилось извлечение структуры из книг в виде разбиения на страницы, параграфы, главы, структура извлекалась с использованием оглавления, которое присутствовало в большинстве книг.

E
162
B972
1904

CONTENTS.	
	Page
EDITOR'S FOREWORD (new)	7
PREFACE to the Third Edition	15
INTRODUCTION	21
TRAVELS	29
APPENDICES; viz:	
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 Linnaeus . . .	157
Nº 2. Tables and Statements relating to the Commercial Situation of the United States, both before and since the American War	162
Nº 3. Anecdotes of the Indians	189
Nº 4. ————— of several Branches of the Fairfax Family, now domiciliated in Virginia	197
Nº 5. Diary of the Weather	215

2. Извлечение структуры из документов на основе правил

В 2019 году проводились соревнования FinTOS [2], где из финансовых документов извлекалась структура в виде иерархии уровней заголовков документов.

Одна из команд-участниц [3] извлекала необходимую структуру используя оглавление документов и правила (междустрочный интервал, отступ, шрифт, символы нумерации).

Parts Collapse

Part 1 Collapse Delete Part

Title: iMGP GLOBAL MACRO

Start page: 1

End page: 2

Topic: kiid

Confidence: high

Chapters Collapse

Chapter 1 Collapse Delete Chapter

Title: OBJECTIFS ET POLITIQUE D'INVESTISSEMENT

Start page: 1

End page: 2

3. Извлечение структуры из документов на основе машинного обучения

В статье 2017 года [4] структура документа извлекалась с использованием методов машинного обучения, включая глубокое обучение. Сначала строки подавались на вход одному классификатору (он определял - заголовок/не заголовок), затем строки-заголовки классифицировались точнее другими классификаторами.



Сложности в использовании существующих решений

- 1) Документы могут содержать заголовки или элементы нумерованных списков глубокой вложенности.
- 2) Как правило в существующих решениях основное внимание уделялось выделению структуры в виде заголовков, для нашей задачи требуется также определять элементы списков.
- 3) В большинстве примеров, приведенных выше, осуществляется классификация текстовых блоков, в нашей задаче необходимо классифицировать каждую строку документа.

B.2.2 Пожарно-техническая комиссия для осуществления поставленных задач должна:

 B.2.2.1 Ежеквартально проводить детальный осмотр территории, всех зданий, сооружений, установок, складов, лабораторий, мастерских и т.п. в целях выявления нарушений противопожарного режима.

Практическая часть

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. Стока 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»

1) Описание логической структуры

Для описания логической структуры был разработан манифест. Он необходим при осуществлении разметки корпуса документов для обучения классификатора.

Практическая часть

2) Подготовка датасета для обучения

Был использован набор документов в виде изображений в формате .jpeg, скачанный из интернета. Документы содержат тексты договоров различных предприятий (СТО, РД ЭО и др.).

Документы были размечены с использованием системы для разметки, разработанной в ИСП РАН [5].

Для проверки логичности манифеста и правильности разметки была посчитана специальная статистика Cohen's kappa.

Данная статистика позволяет оценить меру согласия между аннотаторами при решении задачи классификации.

STO 1.1.1 03.003.1552-2018

организуют и проводят совместно с администрацией сооружаемых объектов общественные смотры противопожарного состояния закрепленных территорий за подрядчиками, пожарно-технические конференции, соревнования боевых расчетов противопожарных пожарных, участвуют в работе пожарно-технических комиссий, консультируют персонал сооружаемых объектов по вопросам пожарной безопасности, анализируют противопожарное состояние сооружаемых объектов ОИАЭ и разрабатывают мероприятия по улучшению противопожарного состояния объектов строительства, организуют и обеспечивают решение других вопросов, предусмотренных условиями заключенного договора и возложенных обязательств

header Организация тушения пожара при строительстве ОИАЭ

header 7.1 Контроль за техническим состоянием наружного и внутреннего противопожарного водопровода, первичных средств пожаротушения и других систем, установок и средств противопожарной защиты

list 7.1.1 Наружный противопожарный водопровод

list 7.1.1.1 Наружные сети противопожарного водопровода должны находиться в исправном состоянии и обеспечивать требуемый нормативный расход и напор воды на нужды пожаротушения. Проверка их работоспособности должна осуществляться не реже 2 раз в год (весной и осенью) совместно с объектовой (территориальной) пожарной охраной в районе выезда которой находитсяплощадка строительства. По результатам проверки составляется соответствующий акт (приложение Р)

list 7.1.1.2 Испытание сети противопожарного водоснабжения должно производиться после каждого ремонта или подключения новых потребителей

text Комиссия по проверке сети противопожарного водопровода назначается приказом руководителя работ генподрядчика

list 7.1.1.3 Вскрытие колодцев, осмотр пожарных гидрантов и пуск воды должен осуществляться совместно представителями генподрядчика и подразделения пожарной охраны

text Датой оплаты поставленного товара считается дата списания денежных средств с отдельного счета Покупателя

text Расчеты по настоящему контракту осуществляются после поступления средств от государственного заказчика (головного исполнителя) на отдельный счет Покупателя в уполномоченном банке

text В случае поступления денежных средств от головного исполнителя на иной расчетный счет, расчеты между Поставщиком и Покупателем осуществляются с использованием иных расчетных счетов

text 6.5. Об исполнении контрактных обязательств между сторонами подписывается акт подтверждения исполнения обязательств по контракту по форме (Приложение №5 к контракту), свидетельствующий об исполнении контракта

header VI. ОТВЕТСТВЕННОСТЬ СТОРОН

list 7.1. За неисполнение либо ненадлежащее исполнение своих обязательств по контракту Стороны несут ответственность, установленную действующим законодательством РФ

text 7.2. При исполнении контракта, по согласованию Покупателя с Поставщиком, допускается пролонгация контракта на необходимый срок; при неполной выборке предельной суммы по контракту с сохранением номенклатуры и цен; при нарушении сроков поставки товара, выполнения работ, не по вине Поставщика; при необходимости проведения расчетов по контракту с использованием отдельного счета в рамках Федерального закона №275-ФЗ

text 7.3. При нарушении сроков поставки товара, выполнения работ, оказания услуг по вине Поставщика (исполнителя, подрядчика) пролонгация контракта на необходимый срок возможна по решению Заказчика с предъявлением поставщику штрафных санкций в соответствии с условиями заключенного контракта

text 7.4. В случае нарушения сроков поставки Покупатель вправе предъявить Поставщику требование об уплате пени в размере 0,5% от цены контракта за каждый день нарушения срока поставки

text 7.5. Убытки, причиненные Покупателю вследствие ненадлежащего оформления и выставления счета-фактуры, подлежат возмещению Поставщиком в полном объеме

text 7.6. Возмещение Поставщиком убытков в случае неисполнения своего обязательства и выплата неустойки за его неисполнение, не освобождает Поставщика от исполнения обязательства по поставке

text 7.7. В случае нарушения сроков поставки более чем на 10 дней, Покупатель вправе в одностороннем порядке расторгнуть настоящий контракт. Контракт считается расторгнутым с момента направления Покупателем Поставщику уведомления об одностороннем расторжении контракта

text 7.8. Стороны обязуются соблюдать антикоррупционное законодательство. Подписание антикоррупционной оговорки (приложение №4 к контракту) является обязательным условием контракта

text 7.9. В случае невыполнения каких-либо обязательств по контракту, в том числе по несанкционированному открытию (не открытию) отдельного счета в рамках Федерального закона №275-ФЗ, Поставщик обязан возместить Покупателю все убытки в полном объеме, в том числе штрафы, неустойки и издержки, выплаченные Покупателем за неисполнение требований законодательства при исполнении настоящего контракта

text 7.10. Поставщик предоставляет Покупателю право вскрытия Товара для проведения профилактических работ и устранения мелких дефектов, не связанных с ремонтом Товара

text 7.11. В период действия гарантии все затраты по вызову оборудования для ремонта или замены и доставке отремонтированного оборудования несет Поставщик

text 7.12. Поставщик гарантирует, что поставляемое по контракту ПО, свободно от обременений на него со стороны третьих лиц и Поставщик обладает всеми законными правами и/или передачи прав на ПО, поставляемое по контракту. В случае предъявления Покупателю претензий со стороны третьих лиц в связи с использованием Покупателем

Практическая часть

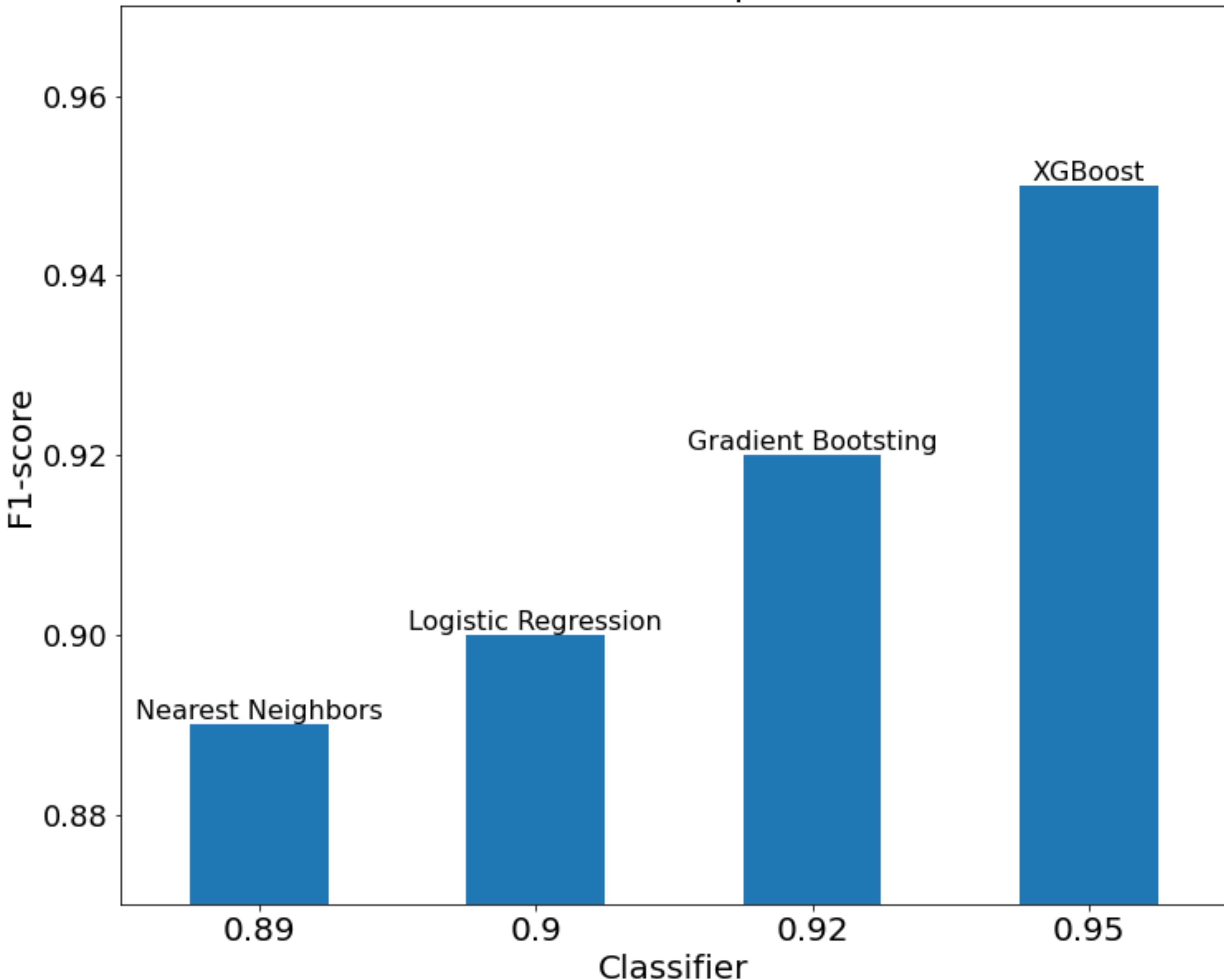
3) Выделение признаков

- Признаки, основанные на регулярных выражениях (строка начинается с заглавной / строчной буквы, цифры, тире и т. д., заканчивается точкой, запятой, буквой).
- Текстовые признаки (количество букв в первом слове, длина строки, число слов в строке).
- Визуальные признаки (отступ от левого края, жирность, высота текста).
 - + признаки четырех предыдущих и следующих строк
 - + усредненные (в пределах документа) значения некоторых признаков

Classifier comparison

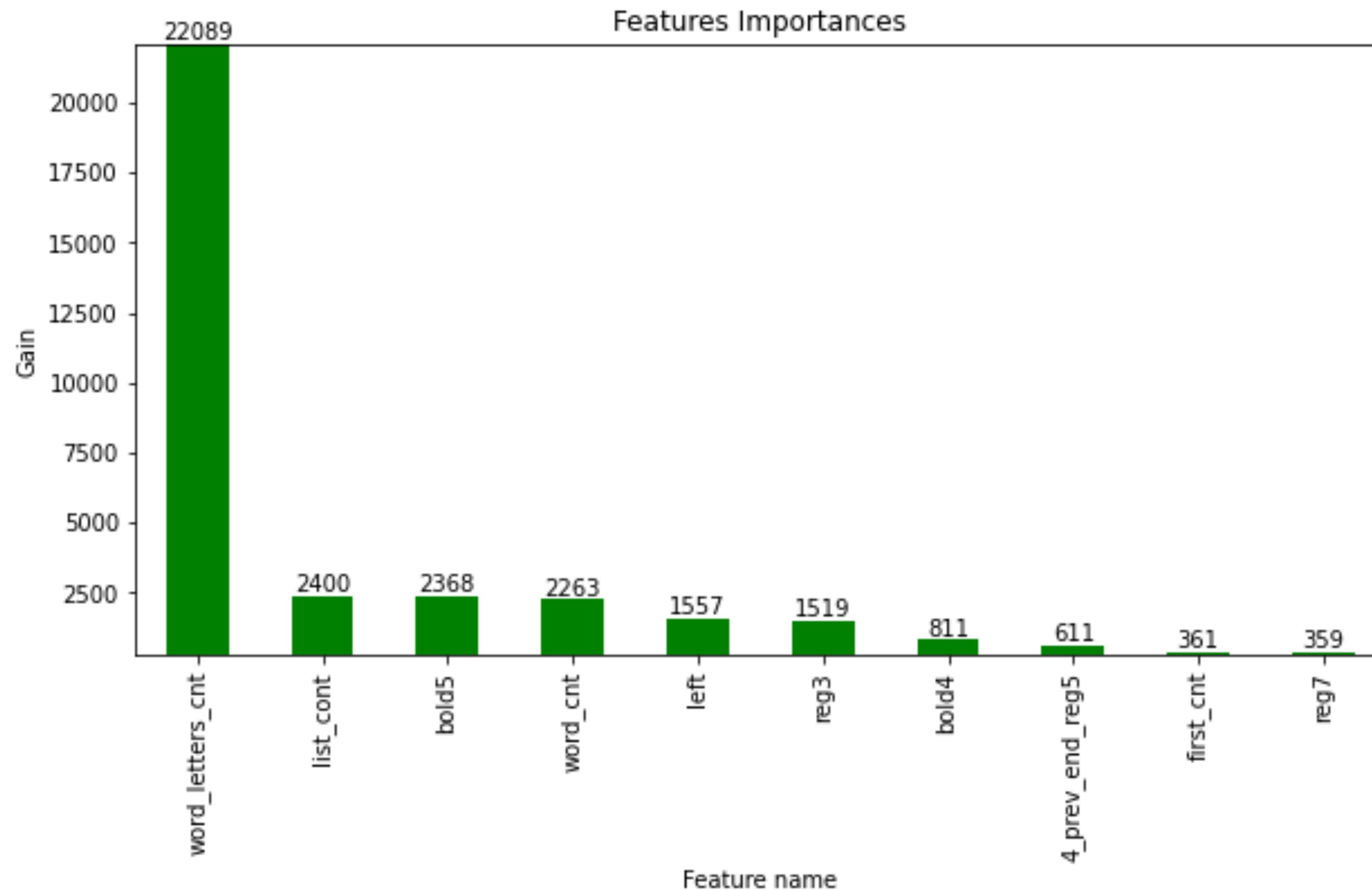
Практическая часть

4) Подбор классификатора



Практическая часть

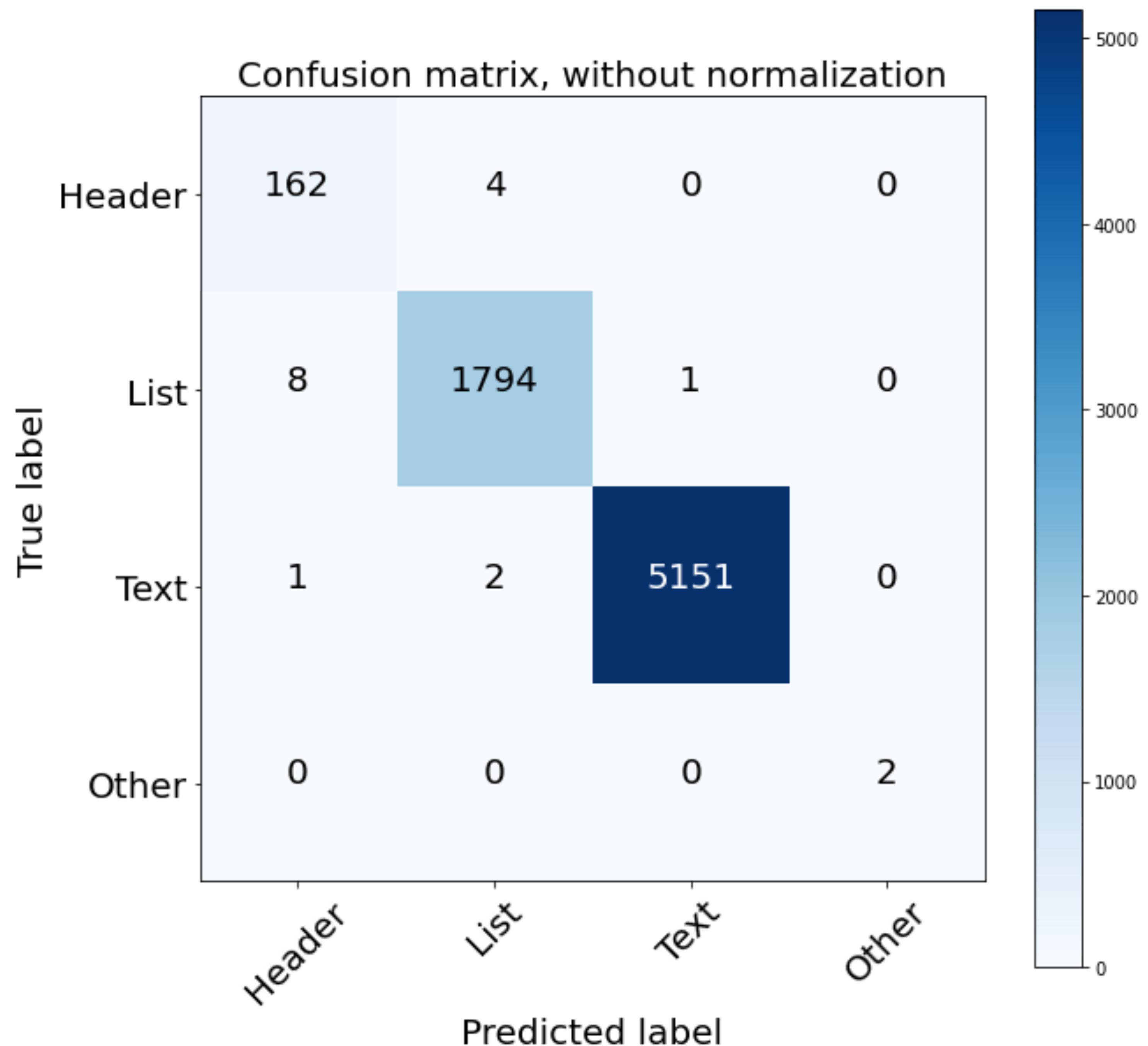
5) Анализ значимости признаков



Практическая часть

6) Настойка параметров
XGBClassifier [6]

7) Анализ ошибок
классификатора



Заключение

В рамках моей курсовой работы был реализован алгоритм классификации строк документа, определяющий в документах заголовки и списки разных уровней вложенности.

- Разработан манифест, описывающий каждый из типов строк, размечен корпус документов, используемый в качестве тренировочных данных.
- Выявлены признаки, наиболее подходящие для обучения классификатора, проанализирована их важность.
- Выбран наиболее подходящий классификатор, для которого были подобраны оптимальные параметры.
- Проведен анализ, на каких строках классификатор ошибается чаще всего.
- Дальнейшие исследования могут быть направлены на определение уровней вложенности строк различных типов.

Список литературы

- [1] Antoine Doucet, Gabriella Kazai, Sebastian Colutto, Günter Mühlberger. Icdar 2013 competition on book structure extraction. 12th International Conference on Document Analysis and Recognition (ICDAR), 2013
- [2] R̄emi 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 understanding of a Documents Structure. In 4th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, 2017
- [5] ParagraphLabelerApp
- [6] Complete Guide to Parameter Tuning in XGBoost with codes in Python