Слайд 1
Здравствуйте, меня зовут Богатенкова Анастасия, я являюсь студенткой кафедры Системного программирования Московского государственного университета. Темой нашей работы является «Извлечение логической структуры из сканированных документов».

Слайд 2
Существует большое количество документов, являющихся сканированными документами без текстового слоя, с которыми необходимо работать. Это означает, что по таким документам нельзя осуществлять поиск и в них нельзя выделять или копировать текст. Извлечение структуры из подобных документов может быть полезно для их эффективного анализа, например, при поиске по документам. На слайде представлен вариант структуры документа, которую можно выделить.

Слайд 3
Целью данной работы является извлечение логической структуры из документов. Будет описан метод, основанный на многоклассовой классификации строк документа. Набор классов состоит из заголовков, элементов списков и текстовых строк. На слайде представлен пример классификации строк: в красную рамку обведена строка-заголовок, зелеными рамками помечены элементы списков, в синюю рамку обведены текстовые строки.

Слайд 4
Существует несколько подходов к решению задачи извлечения структуры из документов:
1) первый способ основан на использовании оглавления, если такое присутствует в документе;
2) методы, основанные на применении правил, которые прописываются вручную;
3) методы, основанные на машинном обучении.

Слайд 5
В 2013 году на одном из соревнований извлекалась структура из книг на основе оглавления.

Слайд 6
В соревнованиях 2019 года по извлечению структуры из финансовых документов одна из команд использовала оглавление документов и некоторый набор правил (отступ, жирность, нумерация).

Слайд 7
В одной из работ структура в виде иерархии заголовков извлекалась с помощью композиции классификаторов.

Слайд 8

Перечисленные решения нам не очень подходили, так как в них ограничивался уровень вложенности извлекаемой структуры, однако документы могут содержать глубоко вложенные элементы. Кроме того, мы рассматриваем также извлечение элементов списков, в большинстве работ уделяется внимание заголовкам. И, наконец, наша задача подразумевает классификацию строк, в упомянутых решениях производится классификация текстовых блоков.

Слайд 9

Приступим к описанию набора данных, с которым проводилась работа. Данные представляют собой набор изображений сканированных документов в формате JPEG, скачанный с сайта государственных закупок. Документы содержат тексты договоров госзакупок. Каждое изображение мы рассматривали как отдельный документ. Мы исключили из рассмотрения документы, содержащие таблицы, графики и другие нетекстовые элементы. На слайде представлен пример документа.

Слайд 10

Процесс создания размеченного набора данных можно разбить на несколько этапов:

1) Спецификация задания, то есть формальное определение задания и формата данных.
2) Составление манифеста - специальной инструкции для аннотаторов.
3) Непосредственная разметка данных в соответствии с манифестом.
4) Измерение согласованности аннотаторов. Если значение согласованности является удовлетворительным, можно перейти к следующему шагу. Если это уровень согласованности недостаточно высок, необходимо вернуться к предыдущим этапам и пересмотреть манифест, уточнив спорные моменты.
5) Вынесение решения по аннотациям - объединение результатов разметки несколькими аннотаторами.

Слайд 11

Первый этап разметки - спецификация задания. Мы поставили задачу как многоклассовую классификацию строк документа. Аннотатору последовательно показывались изображения документа со строкой, которую нужно разметить, обведенной в синюю рамку. Аннотатор должен отнести выделенную строку к одному из заданных классов: заголовок, элемент списка, текст или другое (не текст).

Слайд 12

Второй этап разметки - написание манифеста. Это описание правил, которые предписывают аннотатором, по какому принципу им необходимо разметить данные. В манифесте допустимо использование нестрого определенных понятий (жирный текст, выравнивание по центру). Важно, чтобы аннотаторы понимали описанные правила одинаково. Для этого вычисляется согласованность, о которой будет сказано далее.

Слайд 13

Третий этап - непосредственная разметка данных. Для этого использовалась специальная система разметки, разработанная в Институте Системного

Программирования. На слайде показан процесс разметки - для очередной выделенной строки документа необходимо указать ее тип.

Слайд 14

Следующим этапом является вычисление уровня согласия между аннотаторами. После разметки 10 документов двумя аннотаторами была посчитана специальная статистика Каппа Коэна, значение которой оказалось равным 0.975, что считается высоким уровнем согласованности. Было решено размечать остальной корпус документов, в результате чего было размечено 600 документов (около 20 тысяч строк). Таким образом, был получен размеченный набор данных.

Слайд 15

Далее приступим к описанию реализованного метода. На схеме представлены основные шаги в реализации метода.

1) Первым этапом является извлечение текста и координат рамок строк документа с помощью методов оптического распознавания символов. Для этого была использована специальная программа Tesseract.
2) Вторым этапом является составление вектора признаков для каждой строки документа.
3) Третьим этапом является обучение классификатора на полученных векторах признаков.

Опишем второй и третий этапы более подробно.

Слайд 16

На основе выделенного текста строк, координат ограничивающих их рамок и на основе самого изображения документа извлекались признаки, которые можно разбить на три группы:

1) Признаки, основанные на регулярных выражениях. К таким признакам относятся индикаторы того, что строка подходит под определенный шаблон: например, начинается с нумерации или буквы, заканчивается запятой, точкой, буквой и т. д.
2) Текстовые признаки. Данная группа признаков связана с подсчетом некоторых строковых характеристик, например, число букв в первом слове строки, число символов в строке, число слов в строке.
3) Визуальные признаки. Данная группа признаков связана с графическим представлением текста в документе. Выделялись такие признаки, как отступ от левого края страницы, жирность шрифта, высота текста.

Кроме того, к вектору признаков для каждой строки добавлялись признаки для четырех предыдущих и четырёх следующих строк а также усредненные (в пределах документа) значения некоторых признаков.

Слайд 17

При решении задачи было опробовано множество методов машинного обучения, для тех алгоритмов, которые показали наилучшие результаты было проведено сравнение. На слайде представлены результаты для четырех классификаторов: алгоритм k

ближайших соседей, логистическая регрессия, градиентный бусины и xgboost. Наилучший результат показал xgboost, поэтому было решено выбрать его.

Слайд 18
С помощью специальной библиотеки был проведен анализ значимости признаков (мы рассматривали Information gain). На слайде представлена диаграмма признаков, которые имеют наибольший вес при вычислении предсказания классификатора. Самым значимым признаком оказалось число букв в первом слове строки. Следующим по значимости является признак-индикатор того, что строка является продолжением некоторого списка. На третьем месте - жирность шрифта. Далее идут число слов в строке и отступ от левого края.

Слайд 19
После настройки параметров классификатора (итоговые значения параметров представлены на слайде), F-мера на кросс-валидации оказалась равной 0.98995.

Слайд 20
Далее был проведен анализ ошибок классификатора. На слайде представлена матрица ошибок: по вертикальной оси расположены правильные классы, по горизонтальной - те классы, которые предсказал классификатор. В клетках на пересечении расположены значения количества строк, у которых совпали данные классы. Так, например, 8 строк, являющихся элементами списков, классификатор отнес к классу «Заголовок». 4 строки, напротив, являются заголовками, но были отнесены к классу «Список». Аналогичную статистику можно посмотреть и для других пар классов. Как можно заметить, классификатор чаще всего путает заголовки и элементы списков. Это можно объяснить тем, что заголовки часто начинаются с нумерации, а элементы списков имеют большой отступ от левого края страницы.

Слайд 21
Таким образом, был разработан метод выделения логической структуры документа, основанный на классификации строк документа.
Метод состоит из трех шагов - изображения документов обрабатывались с помощью программы Tesseract для получения текста и рамок строк, далее выделялись векторы признаков и обучался классификатор.
Кроме того, с помощью ручной разметки получен размеченный набор данных, доступный для изучения.
Спасибо за внимание!