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

Введение 2

1 Постановка задачи 3

2 Обзор существующих решений 4
- 2.1 Извлечение структуры из документов на основе оглавления и правил . . . 4
- 2.2 Извлечение структуры из документов на основе машинного обучения . . . 5

3 Описание практической части 7
- 3.1 Описание логической структуры . . . 7
- 3.2 Обоснование выбранного инструментария . . . 7
- 3.3 Подготовка датасета для обучения . . . 7
- 3.4 Выделение признаков . . . 8
- 3.5 Подбор классификатора . . . 10
- 3.6 Анализ важности признаков . . . 11
- 3.7 Настройка параметров модели . . . 13

Заключение 15

Список литературы 16

Приложение 17

Введение

Большое количество текстовой информации представлено в виде pdf-документов, причем эти документы могут быть сканированными копиями 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 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, 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.

1 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.

Ограничим класс документов, с которыми будем работать, сканированными документами без текстового слоя. Будем считать, что в документах не содержатся изображения и таблицы.

2 Обзор существующих решений

По извлечению структуры из документов существуют несколько подходов:

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

2.1 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.

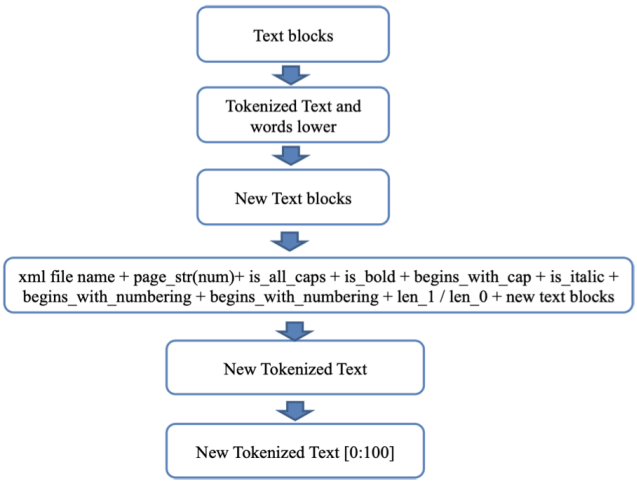
В 2019 году проводились соревнования FinTOC [2], где из финансовых документов извлекалась структура в виде иерархии уровней заголовков документов. Максимальная глубина уровней равна пяти. Одна из команд-участниц [3] извлекала необходимую структуру используя оглавление документов, а также систему правил, 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.

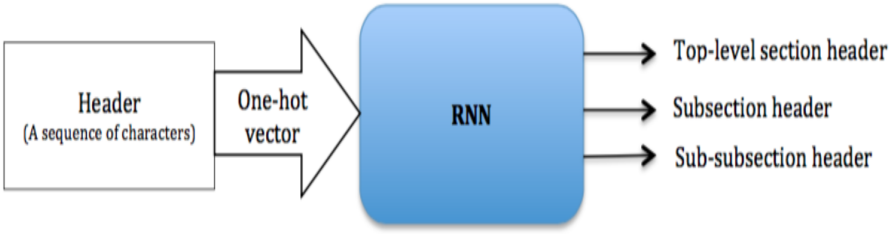
2.2 Извлечение структуры из документов на основе машинного обучения

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 [4] 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 [5] 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.

3 Описание практической части

3.1 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.

3.2 Обоснование выбранного инструментария

В качестве языка реализации был выбран Python, так как для данного языка программирования разработано множество необходимых при решении задачи библиотек, а именно:

- re – библиотека для работы с регулярными выражениями;
- opencv – библиотека для работы с изображениями (библиотека компьютерного зрения и машинного обучения);
- pytesseract – обертка над программой tesseract, предназначенной для распознавания текста на изображении;
- sklearn – библиотека для решения различных задач машинного обучения.

3.3 Подготовка датасета для обучения

Для обучения классификатора строк необходимо иметь размеченный корпус документов. Для этой цели был использован набор документов в виде изображений в формате .jpeg, скачанный из интернета и размеченный с использованием системы для разметки, разработанной в ИСП РАН [6].

Данная система разметки с помощью библиотеки pytesseract позволила выделить в документах без текстового слоя строки текста, а также рамки, заключающие в себе эти текстовые строки (bounding boxes). То есть результатом разметки стал набор .json файлов, каждый из которых содержал название документа и список текстовых строк с координатами заключающих их рамок и правильными метками.

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. Коэффициент согласованности можно посчитать следующим образом:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

где p_o - вероятность согласованности меток, приписанных аннотаторами (точность), p_e - вероятность того, что согласие аннотаторов достигнуто в результате случайной разметки.

После разметки десяти документов значение статистики κ оказалось равным 0.975, после чего было решено размечать остальной корпус документов. В результате было размечено 600 документов и отдельные .json файлы были объединены в один, использующийся для обучения классификатора.

3.4 Выделение признаков

Для каждой размеченной строки документа необходимо определить набор характеризующих её числовых признаков, которые нужно выделить. Далее выделенный вектор признаков подается на вход классификатору.

Среди всех признаков можно выделить следующие группы:

- Признаки, основанные на регулярных выражениях.

  Данная группа признаков основывается на анализе начала и конца каждой строки. Такие признаки очень важны для выявления элементов списков различных типов, а также могут сигнализировать о конце заголовка или начале списка.

  Например, для каждой строки анализировалось следующее:

  - с помощью регулярных выражений специального вида выявлялось, начинается ли строка с цифры или буквы со скобкой или точкой (также анализируются иерархические выражения вида 1.1.1);
  - начинается ли строка с тире (и других символов, характерных для маркированного списка);
  - состоит ли строка целиком из заглавных букв (характерно для некоторых заголовков);
  - начинается ли строка с заглавной (строчной) буквы;

- начинается ли строка с конкретных слов типа «Раздел», «Секция», «Глава» и т. д.;
- оканчивается ли строка символами вида «. , ; :»;
- оканчивается ли строка строчной буквой.

- Текстовые признаки.

  Данная группа признаков связана с подсчетом некоторых строковых характеристик, а именно:

  - количество букв в первом и втором словах строки;
  - количество слов в строке (строка разбивается на слова по пробелам);
  - количество символов в строке (длина строки).

- Визуальные признаки.

  Данная группа признаков связана с графическим представлением текста в документе. То есть при анализе строки рассматривается не ее текст, а следующие признаки:

  - отступ от левого края страницы;
  - высота текста строки (точнее высота ограничивающей ее рамки);
  - отступ от верхнего края страницы;
  - с помощью специальных функций определялась жирность шрифта различных уровней.

Кроме того, к признакам, перечисленным выше, для каждой строки были добавлены аналогичные признаки четырех предыдущих и следующих строк. Это полезно для анализа продолжения блока строк конкретного типа.

Для строк, которые начинаются с нумерации, определялось, есть ли в документе строка, предшествующая данной с нумерацией, меньшей данной на единицу.

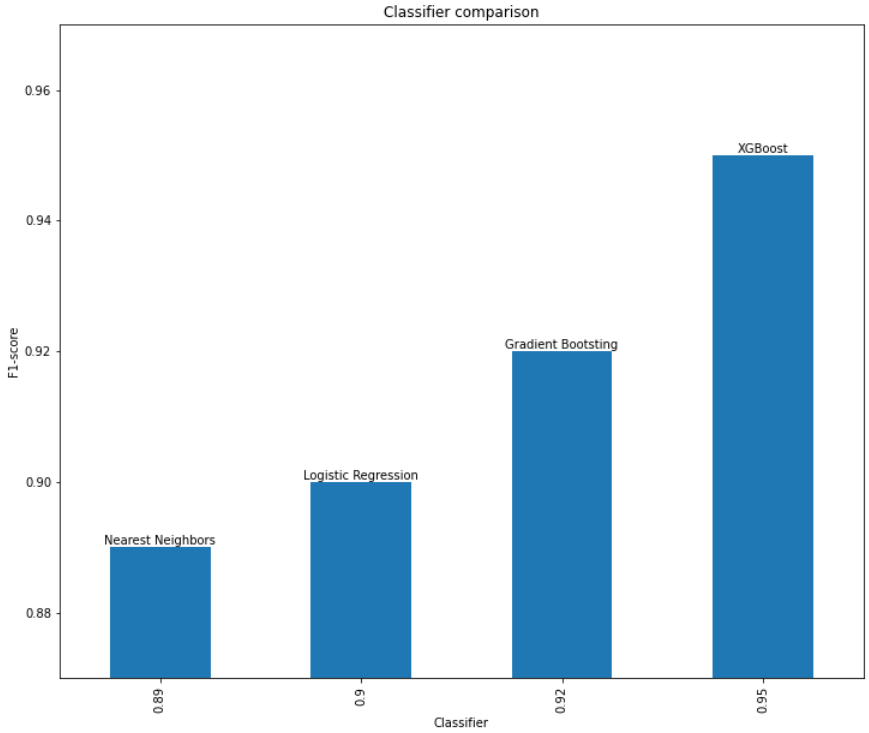
И, наконец, для каждого документа вычислялся средний отступ от левого края страницы, средняя высота шрифта, средняя длина строки, среднее число слов в строках, среднее значение дли жирности шрифта, среднее число букв в первом слове каждой строки. Данные значения добавлялись к признакам каждой строки документа.

3.5 Подбор классификатора

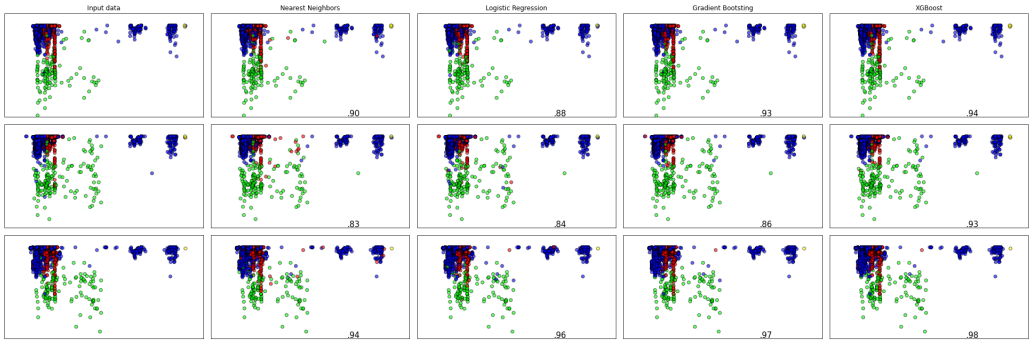
При решении задачи было опробовано множество классификаторов, для лучших из них проведен анализ результатов. В анализе участвовало 4 классификатора:

- алгоритм k ближайших соседей (KNeighborsClassifier);
- логистическая регрессия (LogisticRegression);
- градиентный бустинг (GradientBoostingClassifier);
- экстра-градиентный бустинг (XGBClassifier).

Множество документов тремя способами было разбито на тренировочное и тестовое множества (разбиение по документам), на каждом разбиении было проведено обучение классификаторов и вычисление F1-score. Усредненные значения F1-score для каждого классификатора показаны на графике ниже:



На графиках, изображенных ниже, точками показано распределение множеств строк документов различных типов по двум признакам: горизонтальная ось соответствует величине отступа строки от левого края страницы, вертикальная ось соответствует значению жирности шрифта. При этом красному цвету соответствуют строки-заголовки, зелёному – списки, синему – текст, желтому – всё остальное. В первом столбце изображено правильное распределение строк, остальные столбцы соответствуют результатам предсказаний классификаторов.



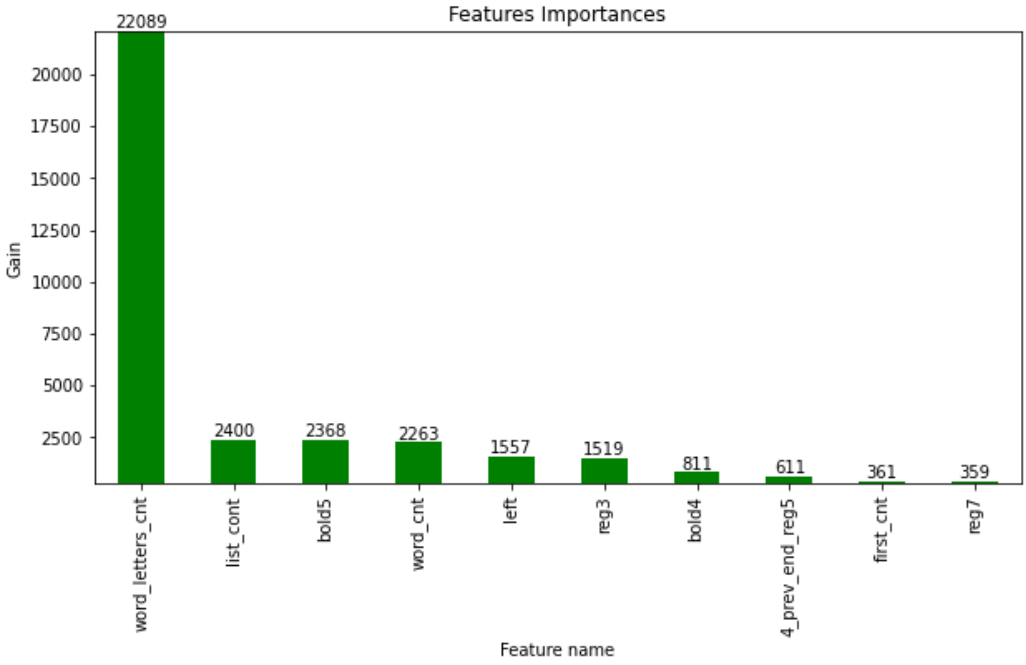
Судя по графикам, по данным двум признакам достаточно легко спутать текст и список по величине отступа, а заголовок и текст по жирности шрифта (однако на графиках не учитываются другие признаки).

Отдельные группы точек справа соответствуют строкам с номерами страниц документов. Как правило, нумерация страниц располагается внизу по центру или в правом нижнем углу страницы. Некоторые классификаторы относят такие строки к типу «Другое».

В целом, все рассмотренные классификаторы показали хороший результат, наилучший результат показал XGBClassifier, поэтому было решено выбрать его.

3.6 Анализ важности признаков

Ниже представлены первые 10 признаков с наивысшей важностью. Это признаки, которые имеют наибольший вес при вычислении предсказания классификатора.



Описание признаков:

- $word\_letters\_cnt$ – число символов первого слова в строке (слова разделяются пробелами);
- $list\_cont$ – признак для строк, являющихся элементами списка, принимающий значение булевого типа: True, если перед данной строкой в документе найдена строка, являющаяся предыдущим элементом данного списка и False иначе;
- $bold5$ – признак жирности шрифта (число 5 означает размер ядра свертки для определения уровня жирности);
- $word\_cnt$ – число слов в строке;
- $left$ – отступ от левого края страницы;
- $reg3$ – признак начала строки с выражения вида 1.1.1 (произвольный уровень вложенности, вместо цифр могут быть буквы);
- $bold4$ – признак жирности шрифта с размером ядра свертки, равным 4;
- $4\_prev\_end\_reg5$ – признак, заканчивается ли предыдущая строка буквой;

- $first\_cnt$ – число букв в начале строки;
- $reg7$ – признак начала строки с тире.

С использованием данных о важности признаков, в признаковое пространство были добавлены новые признаки. Например, вместо одного признака, отвечающего за жирность шрифта, была добавлена целая группа признаков, отвечающая за различные уровни жирности шрифта. Для самых важных признаков к вектору признаков каждой строки документа были добавлены усреднённые значения данных признаков по документу.

На графике выше представлен итоговый список признаков и значения их важности.

3.7 Настройка параметров модели

Градиентный бустинг — это алгоритм машинного обучения, который строит модель предсказания в форме ансамбля слабых предсказывающих моделей (например, деревьев решений).

В нашем случае, при использовании XGBClassifier, для настройки доступны три группы параметров:

- параметры, отвечающие за общее функционирование алгоритма (тип предсказывающей модели, число нитей для запуска);
- параметры для предсказывающей модели, в нашем случае это дерево (шаг обучения, максимальная глубина, параметры регуляризации и т. д.);
- параметры, задающие функцию потерь для минимизации и метрику для оценки качества.

Мы сосредоточимся на подборе параметров второй группы.

Настройка параметров классификатора проводилась по алгоритму, предложенному в [7]:

1. выбирается относительно высокий шаг обучения ($learning\_rate$), для которого подбирается количество деревьев ($n\_estimators$);
2. подбираются параметры деревьев ($max\_depth, min\_child\_weight, gamma, subsample, colsample\_bytree$) для выбранного шага и количества деревьев;
3. настраиваются параметры регуляризации ($lambda, alpha$) для снижения сложности модели и улучшения производительности;

4. понижается шаг обучения и выявляются оптимальные параметры.

На графике ниже представлено изменение F1-score в процессе подбора параметров.

Параметры итоговой модели:

$learning\_rate$

$n\_estimators$

max_depth

min_child_weight

$gamma$

$subsample$

$colsample\_bytree$

$alpha$

Итоговый F1-score, полученный в результате кросс-валидации:

0.98

Заключение

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

Был разработан манифест, описывающий каждый из типов строк, размечен копус 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.

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. 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] Ke Tian and Zi Jun Peng. Finance document extraction using data augmented and attention. In The Second Workshop on Financial Narrative Processing of NoDalida 2019, 2019.

[5] 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.

[6] ParagraphLabelerApp

[7] Complete Guide to Parameter Tuning in XGBoost with codes in Python

Приложение

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.

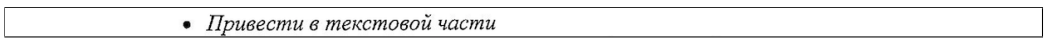
На вход вам будут подаваться документы, в которых выделена прямоугольником одна строка. Вам необходимо для каждой выделенной строки документа определить её тип. Необходимо «Заголовок» пометить цифрой 1, «Список» - 2, «Текст» - 3, «Другое» - 4.

1. Заголовок

   Название главы, секции, подглавы, параграфа. Строка помечается заголовком, если:

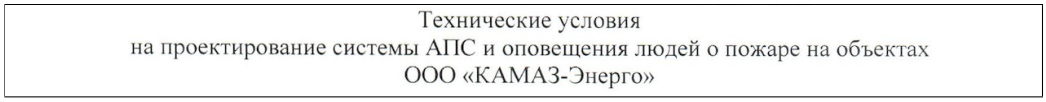
   - текст визуально (полностью) выделяется жирностью;

   - текст полностью выделяется шрифтом (курсив, подчеркнутый, другой шрифт, другой размер шрифта);

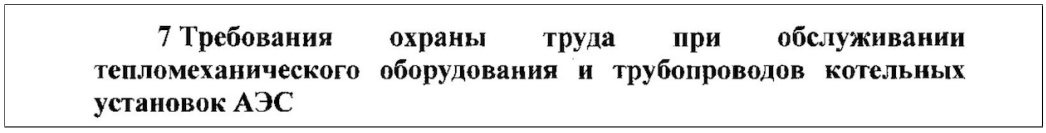


   при этом если текст строки выделен шрифтом частично, то заголовком это не считается;

   - текст выделяется отступом (расположен по центру);



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

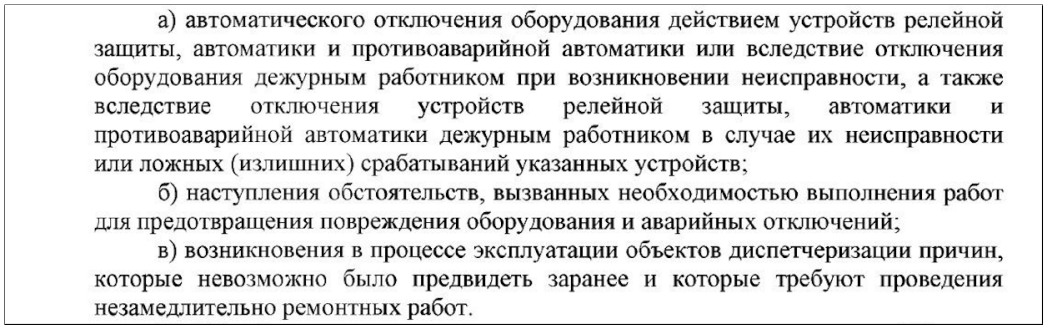


7 Требования охраны труда при обслуживании тепломеханического оборудования и трубопроводов котельных установок АЭС

2. Список

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

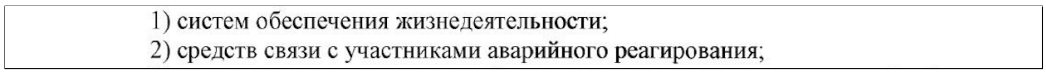
   - строка наряду с несколькими другими строками пронумерована («1. 1) а) 1.1» и т. д. в начале строки);



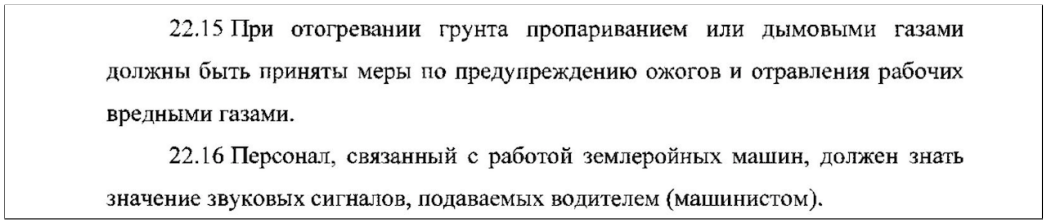
а) автоматического отключения оборудования действием устройств релейной защиты, автоматики и противоаварийной автоматики или вследствие отключения оборудования дежурным работником при возникновении неисправности, а также вследствие отключения устройств релейной защиты, автоматики и противоаварийной автоматики дежурным работником в случае их неисправности или ложных (излишних) срабатываний указанных устройств;

б) наступления обстоятельств, вызванных необходимостью выполнения работ для предотвращения повреждения оборудования и аварийных отключений;

в) возникновения в процессе эксплуатации объектов диспетчеризации причин, которые невозможно было предвидеть заранее и которые требуют проведения незамедлительно ремонтных работ.



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

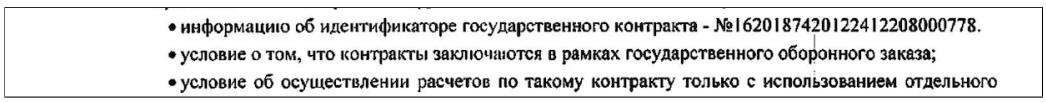


22.15 При отогревании грунта пропариванием или дымовыми газами должны быть приняты меры по предупреждению ожогов и отравления рабочих вредными газами.

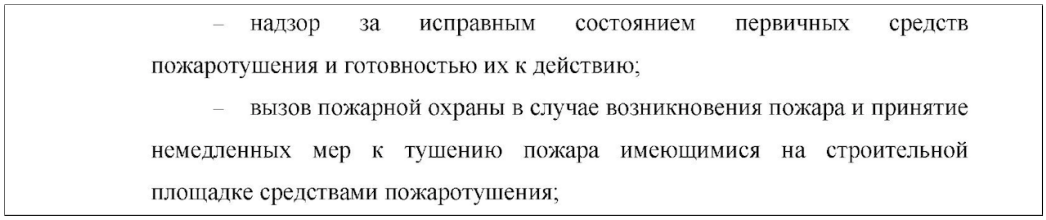
22.16 Персонал, связанный с работой землеройных машин, должен знать значение звуковых сигналов, подаваемых водителем (машинистом).

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

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



- информацию об идентификаторе государственного контракта - №16201874201224122080007778.
- условие о том, что контракты заключаются в рамках государственного оборонного заказа;
- условие об осуществлении расчетов по такому контракту только с использованием отдельного


– 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;

В примере выше 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.

8.9 Аппарат Генерального директора, Департамент планирования производства, модернизации и продления срока эксплуатации, Проектно-конструкторский филиал выполняют закрепленные обязанности по управлению документацией в соответствии с положениями о подразделениях.

8.10 Управление документацией в части финансово-экономической, бухгалтерской и юридической деятельности осуществляют подразделения центрального аппарата по соответствующим направлениям во взаимодействии с финансово-экономическими, юридическими службами и бухгалтериями АС.

9 Контроль проектирования (конструирования)

9.1 Обеспечение качества выбора площадки и проектирования АС осуществляется при организации и контроле выполнения следующих работ:

- исследование потребностей в электроэнергии и строительстве АС;
- выбор новых площадок размещения АС;
- организация разработки коммерческих и технических требований к АС;
- выполнение необходимых проектных, изыскательских и исследовательских работ;
- установленная действующими нормами и правилами экспертиза проектов;
- проектирование и конструирование оборудования и систем для АС;
- организация авторского надзора при сооружении и вводе в эксплуатацию;
- проверка на соответствие установленным требованиям программ обеспечения качества при выборе площадки и проектировании.

9.2 В соответствии с «Положением о порядке назначения и порядке взаимодействия разработчиков проектов РУ и АС между собой и Эксплуатирующей организацией атомных станций, их функциональных обязанностях и ответственности» определяются организации-разработчики проектов АС – Генеральные проектировщики АС, которые разрабатывают и