Задача “Из doc или docx получить иерархический html код”

**Создать микросервис с REST-апи**

Входящие данные json: {url: ‘путь исходного файла’}

Исходящие данные json: {url: ‘путь до полученного архива’}

На выходе должны получить архив с данными:   
файл json иерархические данные

папка library - в которой находятся все изображения документа

Для работы используем готовый вариант кода предоставляемый заказчиком

в папку library сохраняем все изображения, которые есть в документе

в json файле перечисление частей документа

каждая часть, далее по тексту “chapter” состоит из

{

title: название параграфа,

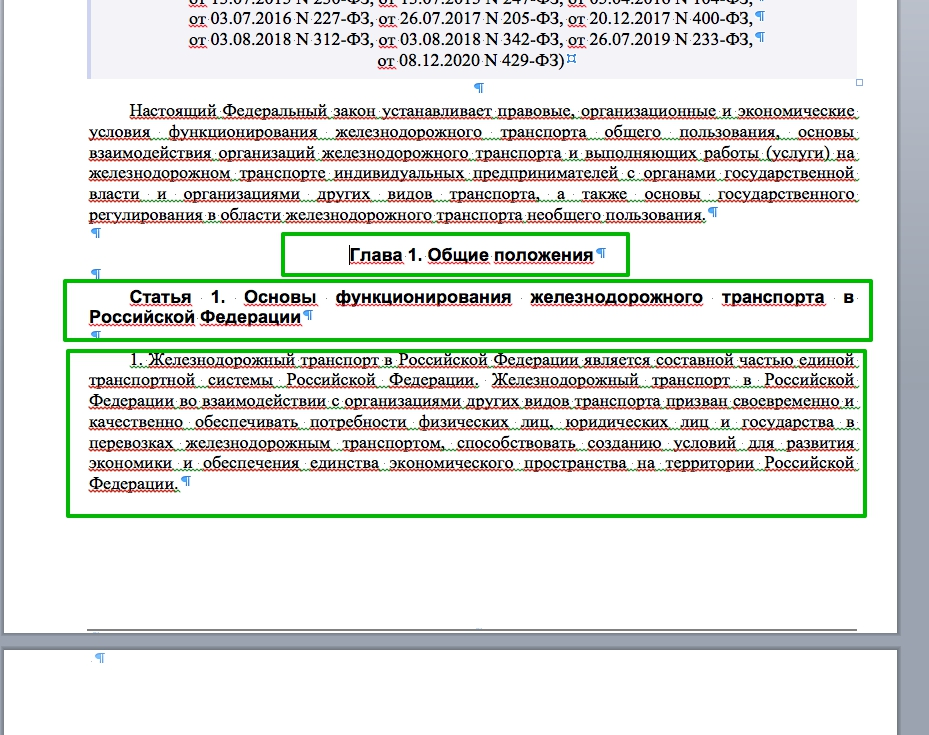
content: преобразованный в html код параграф,

level: уровень параграфа,

inline: это пункт документа

}

Новый блок - это либо заголовок, либо параграф начинающийся с цифры

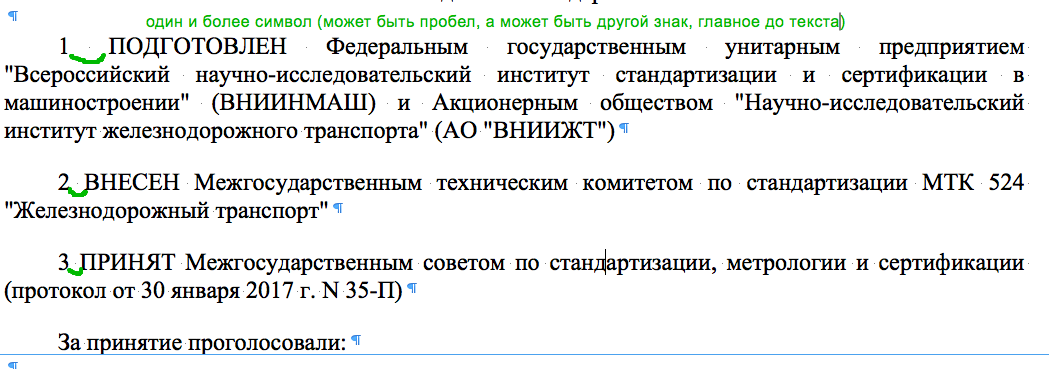


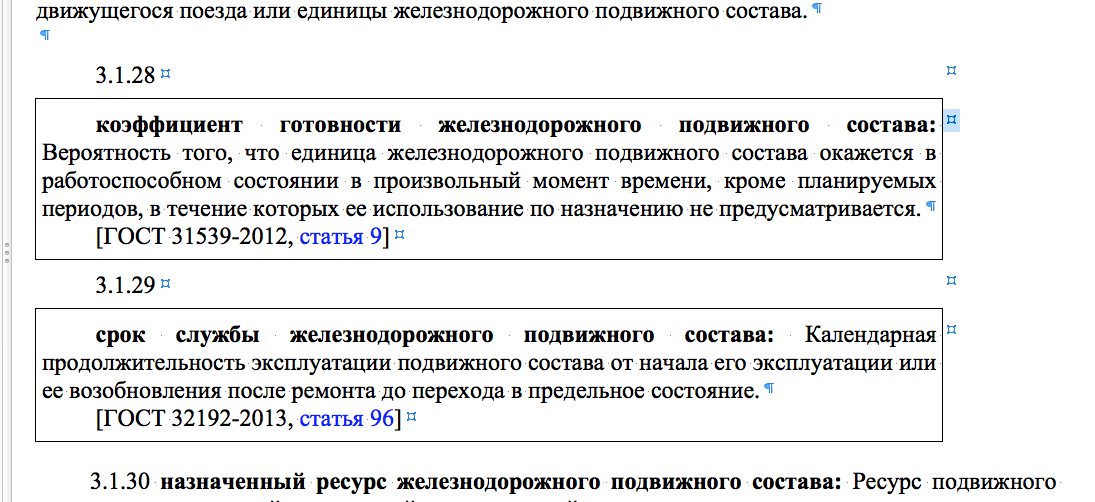
Если новый абзац является заголовком - то значение inline = false

Значение inline = true - если это красная строка начинающаяся с цифры

Eсли inline = false - title - это название заголовка, иначе title - цифра в параграфе. Данная цифра может быть многоуровневой например 1.1

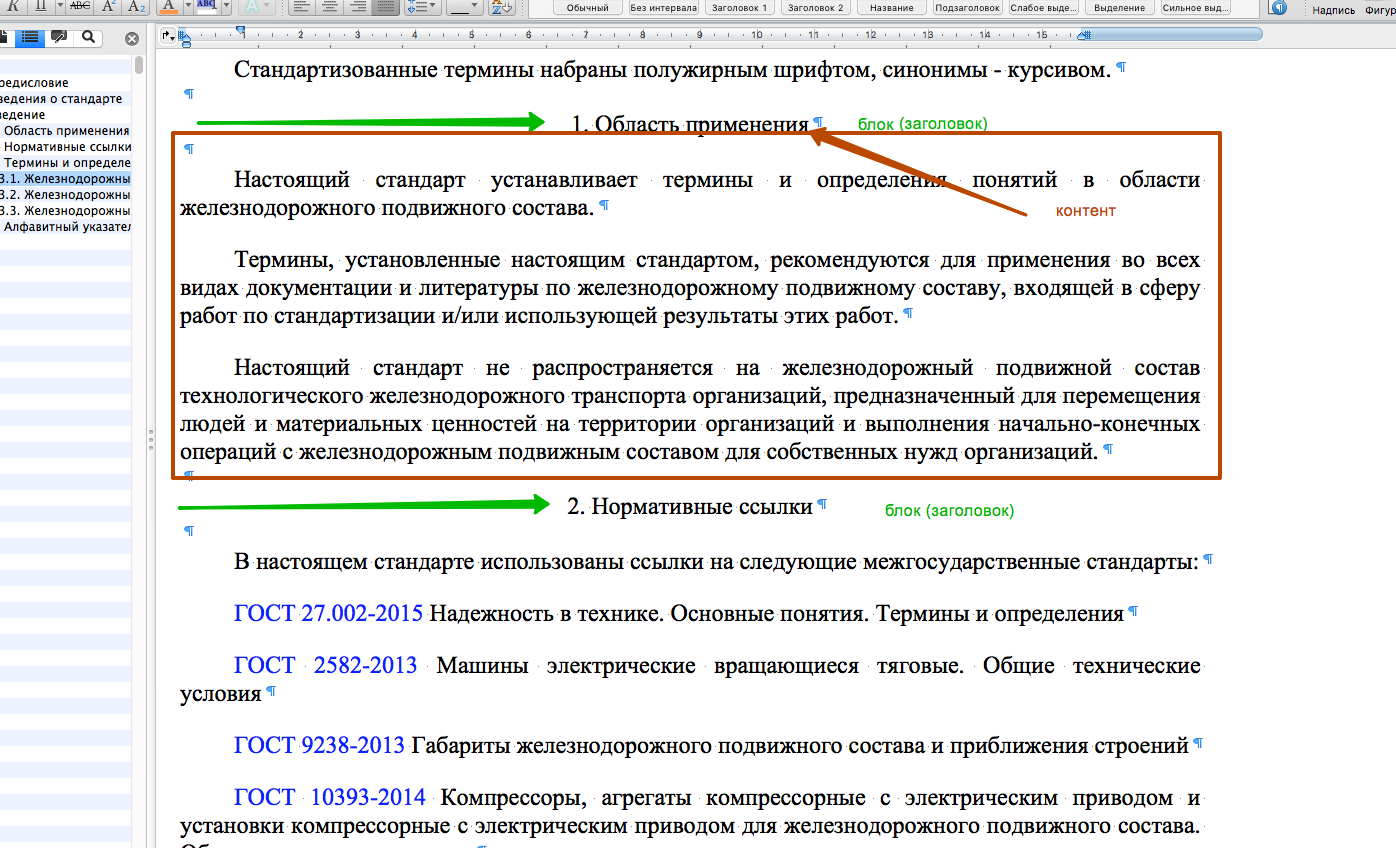
Главное выделять между Цифрой и Текстом именно цифровое значение. Между цифрами может быть в качестве разделителя только точка



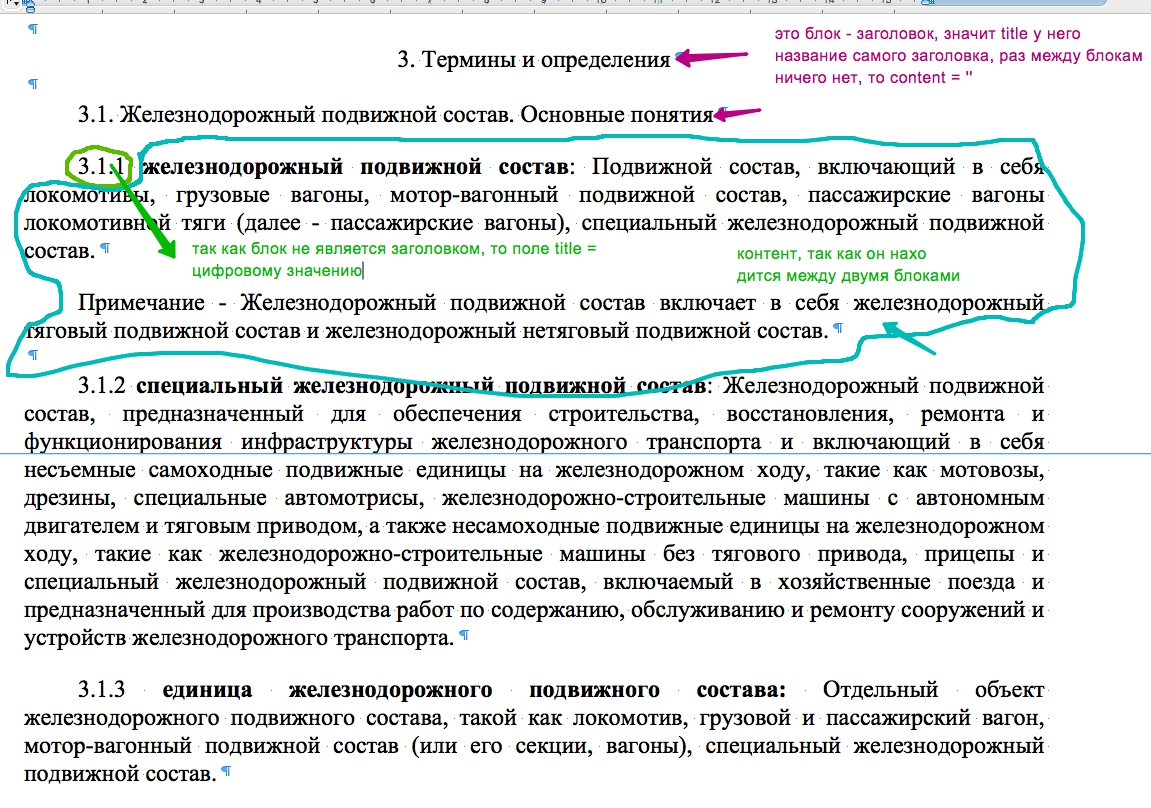


обратить внимание что бывает, так что между цифровым и текстовым значением есть перенос на новую строку

Правило заполнения поля content: входит весь пробразованный html текст от текущего блока, до следующего



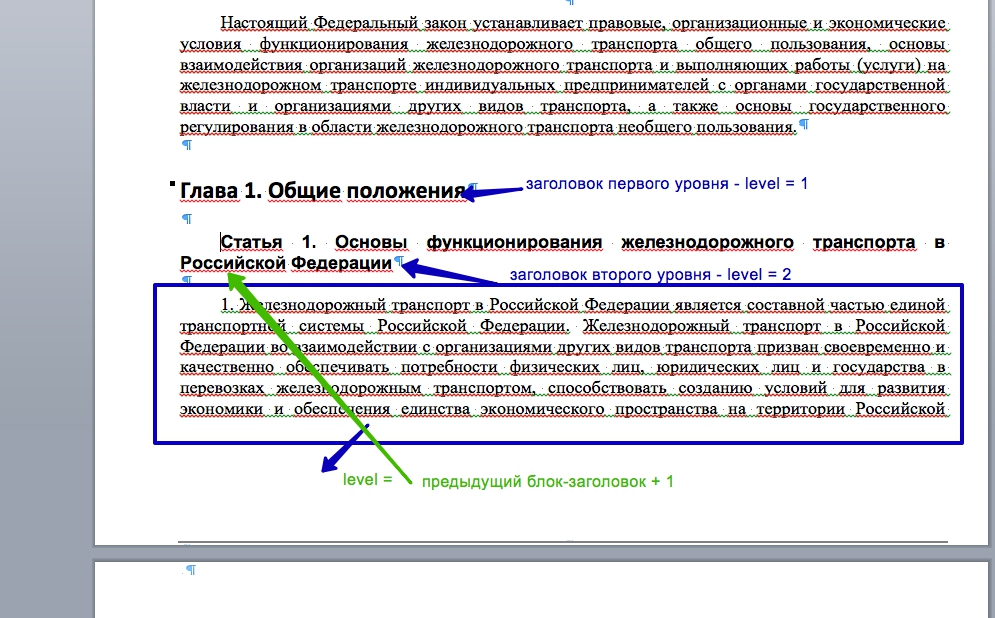
Если между блоками нет текста- значит поле пустое



В начале документа, текст, который находится до первого найденного блока в документе, определяем в блок с {title: ‘Начало документа’, content: тут html текст, level: 1’}

Заполнение поля “level”

Уровень определяется исходя из уровня заголовка. Блок который не является заголовком, имеет уровень +1 от предыдущего уровня-заголовка



**Парсинг параграфов**

Все таблицы, формулы, изображения, списки, специальные символы, параграфы должны быть преобразованы в html код

Все ссылки заменяем на текст

При парсинге - повторяем точно стиль текста, кегль, жирность, шрифт формулы, изображения, таблицы, списки

пример

<https://disk.yandex.ru/i/OoH1AivxvAAn0w> - файл док

html текст