**Сегментация текста**

Сегментация текста – это разделение текста на блоки. На начальном этапе будет рассматриваться сегментация, выраженная при помощи скобок и кавычек. В дальнейшем необходимо будет рассмотреть вариант выделения текста и при помощи запятых.

Сегментация может быть внешней и внутренней. Внутренняя сегментация – это сегментация в пределах одного предложения. Внешняя сегментация охватывает несколько предложений в одном абзаце.

Сегментация может быть вложенной. Например:

*Таким образом, это создает вершину (это один из признаков (свойств) объекта).*

Обработка такой сегментации должна вестись в порядке: изначально обрабатывается основной текст предложения и лишь затем вложенный сегмент, а за ним вложенный во вложенный сегмент. Если рассматривать внешнюю сегментацию, то принцип обработки аналогичный – вначале обрабатывается предложение, которое включает блок сегментации и только потом этот блок. Т.е. у каждого предложения должна быть ссылка на массив блоков сегментации, а блоки сегментации должны рассматриваться как отдельные предложения со своими возможными ссылками на вложенные блоки сегментации.

**Итого:**

Весь текст состоит из абзацев. Один абзац состоит, как минимум, из одного предложения. Это предложение может содержать/не содержать блоки сегментации. Блоки сегментации могут содержать другие предложения и блоки сегментации.

**List<Paragraph> -> List<Sentence> -> List<TextBlock> -> List<Word>**

При анализе текста и разбиении его на лексемы, необходимо учитывать, что идиомы ищутся только в пределах своего блока сегментации. При поиске идиомы можно не обращать внимание только на запятые. На остальные знаки препинания и стоит учитывать.

Для начала реализовать работающий алгоритм без разбиения на параграфы и текстовые блоки!!!

Реализовать поиск Фамилий и инициалов в тексте.

Описание шаблона:

пробел либо какой-то знак пунктуации, либо начало строки, затем одна большая буква или все большие, затем пробел, затем одна большая буква

На каком-то этапе необходимо будет определить все слова, которые пишутся с ЗАГЛАВНОЙ буквы и не являются началом предложения!!! А также определить не являются ли эти слова географическими названиями: к примеру, Московская область, Жовтневый район… И при этом еще могут быть сокращения… Например, Московская обл. или обл. Московская. Поэтому необходимо будет провести поиск по словосочетаниям или поиск по Москве и всем возможным объектам.