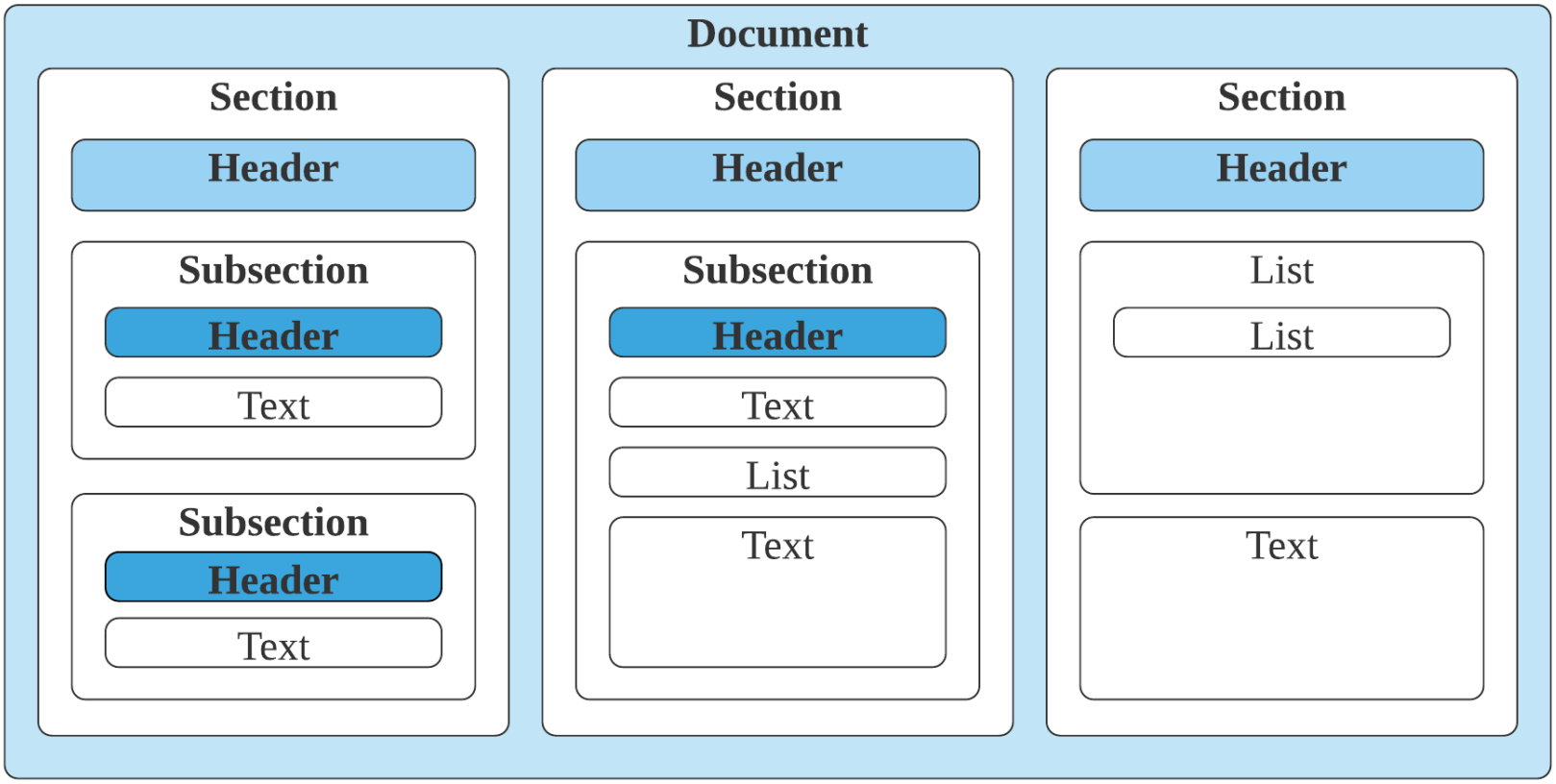
Извлечение иерархической логической структуры из текстовых документов в формате docx

Богатенкова Анастасия

5 ноября 2020 г.

Актуальность темы

Как правило, документы имеют логическую структуру, выделение которой может помочь при решении задач автоматизированного анализа документов.

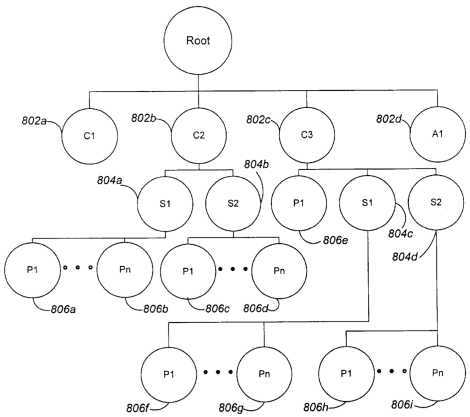


Почему docx?

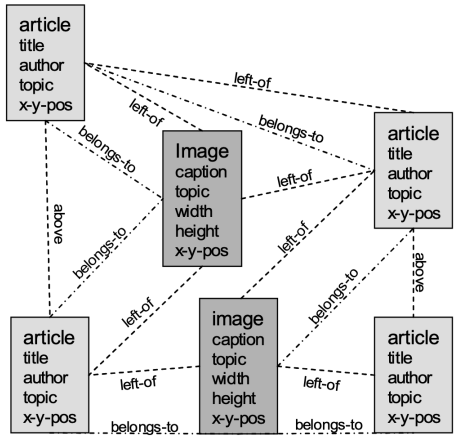
- В настоящее время большое количество документов создаётся и хранится в формате docx.
- Однако данный формат позволяет описывать только физическую структуру документа, то есть описывается то, как **выглядит документ**.
- Хотелось бы уметь выделять из подобных документов логическую структуру

Какую структуру извлекать?

Существуют различные способы представления структуры документа. Мы будем извлекать иерархическую структуру в виде дерева, так как многие документы состоят из последовательности вложенных друг в друга частей.



Tree structure



Graph structure

$[0.5] < START > \rightarrow < TITLE > < COLUMN > < COLUMN >$
$[0.5] < START > \rightarrow < TITLE > < COLUMN >$
$[1.0] < TITLE > \rightarrow < text.line >$
$[1.0] < COLUMN > \rightarrow < TEXT\_BLOCKS >$
$[0.8] < TEXT\_BLOCKS > \rightarrow < TEXT\_BLOCK > < space > < TEXT\_BLOCKS >$
$[0.2] < TEXT\_BLOCKS > \rightarrow < TEXT\_BLOCK >$
$[1.0] < TEXT\_BLOCK > \rightarrow < TEXT\_LINES >$
$[0.9] < TEXT\_LINES > \rightarrow < text.line > < newline > < TEXT\_LINES >$
$[0.1] < TEXT\_LINES > \rightarrow < text.line >$

(a)

$< START > \rightarrow < TITLE > < COLUMN >$
$\rightarrow < text.line > < COLUMN >$
$\rightarrow < text.line > < TEXT\_BLOCKS >$
$\rightarrow < text.line > < TEXT\_BLOCK > < space > < TEXT\_BLOCKS >$
$\rightarrow < text.line > < text.line > < newline > < TEXT\_BLOCK > < space > < TEXT\_BLOCKS >$
$\rightarrow < text.line > < text.line > < newline > < text.line > < space > < TEXT\_BLOCKS >$
$\rightarrow < text.line > < text.line > < newline > < text.line > < space > < TEXT\_BLOCK >$
$\rightarrow < text.line > < text.line > < newline > < text.line > < space > < text.line >$

(b)

Fig. 6. Representing a document in terms of formal grammars: (a) example of a stochastic context free grammar that derives a text document with a title. Upper case symbols refer to non-terminal symbols, while lower case symbols show terminal symbols. (b) a sample document with a title and three lines, derived using the grammar in (a)

Formal grammar representation

План работы

- Провести обзор некоторых форматов документов;
- Провести обзор способов представления логической структуры документа;
- Описать особенности формата docx;
- Описать структуру, которую необходимо извлечь;
- Реализовать метод извлечения структуры и провести экспериментальную проверку реализованного метода;
- Провести оценку качества.