Часть 1

1. Автоматический дискурсивный анализ - автоматический анализ текста, представленного как сложное коммуникативное явление. При дискурсивном анализе используют дискурсивную связанность, структуру дискурса, извлечение именованных сущностей.
2. Расстояние Левенштейна между двумя строками – это использование минимального количества операций для превращения одной строки в другую. Используются операции вставки одного символа, удаления одного символа или замены одного символа на другой.

Для преобразования словоформы «мошка» в «мускат» потребуется использовать 3 операции – заменить второй и третий символы, а также добавить один символ в конец словоформы.

1. Нормализация текста – процесс автоматической обработки, при котором из текста убирается информация о грамматических категориях, остается только смысловая составляющая текста.

К видам нормализации можно отнести лемматизацию и стемминг.

При лемматизации все словоформы текста приводятся к начальным формам – для глагола – инфинитив, для прилагательного – мужской род, единственное число, именительный падеж, и т.д.

При стемминге алгоритм обработки текста убирает все окончания, оставляя только основы лексем.

Однако при стемминге возникает ряд проблем. Основная из них – это обработка слов, которые при изменении формы меняют также свою основу. Например, словоформа «ложка» в родительном и винительном падежах во множественном числе имеет форму «ложек». В таком случае стеммер может игнорировать такие случаи, что приведет к потере некоторых форм слова, или может урезать форму слова до неизменяющейся основы, в данном случае «лож-», что приведет к ситуациям, в которых стеммер будет путать данную основу с другими.

1. Лексическая вероятность – это вероятность использования определенного тэга для конкретной словоформы без учета контекста употребления.

К допущениям относительно лексической вероятности модели можно отнести следующие положения: вероятность появления какой-либо лексемы в тексте зависит только от ее собственного грамматического тэга, а вероятность какого-либо тэга в тексте зависит только от предыдущего тэга.

1. К проблемам для синтаксического анализа в терминах деревьев зависимостей можно отнести избыточность, эллиптичность.

7. В корпусе НКРЯ вероятность увидеть существительное после прилагательного значительно выше, чем увидеть существительное после наречия. Условная вероятность в первом случае равна 0,199, во втором – 0,009. Вероятность появления существительного равна 1 722 487, вероятность появления нареч+сущ равна 15 388, вероятность появления прил+сущ равна 344 134.