1. В автоматический дискурсивный анализ входят задачи дискурсивной разметки, выявление дискурсивных явлений. Пример – автоматическое построение анафорических цепочек. Дискурсивный анализ может потребоваться для sentiment analysis – определение, к чему именно относится характеристика (анализ анафоры).

2. Расстояние Левенштейна – минимальное количество операция удаления, вставки и замены, которое требуется для превращения одного слова в другое.

мошка – мускат

о -> y, ш -> c, вставка т => расстояние Левенштейна = 3

3. Нормализация – приведение слова к нормальной форме. Есть лемматизация и стемминг. Лемматизация – приведение всех словоформ к одной форме (лемме), например, все существительные в им.п. ед.ч., глаголы в инфинитив. Стемминг – сведение всех словоформ к общей для них основе. Проблема стемминга – выпадающие гласные – пень-пня, основа п-. А еще супплетивизм (человек-люди), но тогда, вероятно, лучше считать супплетивизм двумя разными словоформами?

4. Лексическая вероятность – это вероятность появления леммы в корпусе (не словоформы). Используется для того, чтобы выкидывать разборы с маловероятными леммами. Как пример – **при** (может быть предлогом, глаголом в императиве и каким-то очень редким существительным). Вероятность словоформы высока (из-за предлога), а вот вероятность леммы существительного совсем мала. И это поможет отбросить маловероятный разбор как существительного.

5.

1) сочинение – непонятно, что от чего зависит (впрочем, для составляющих это тоже некоторая проблема)

2) единственность вершины как принцип построения структуры зависимостей иногда вызывает проблемы

7.

Adv&S = 15 388

Adj&S = 344 134

Total = 359 522

P(S|Adj) = 0,957

P(S|Adv) = 0,043

Вероятность встретить существительное после прилагательного выше, чем после наречия.

8.

|  |  |  |
| --- | --- | --- |
| S, VP (S -> VP, a VP -> V NP) |  |  |
| 0 (не может быть V Det) | NP (Det N) |  |
| V (не может быть VP) | Det | N, NP |
| build | a | house |