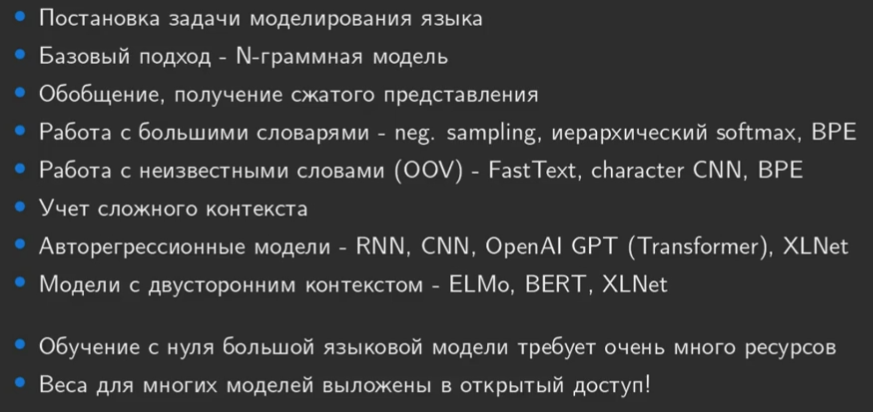
**Общие советы по моделированию языка**



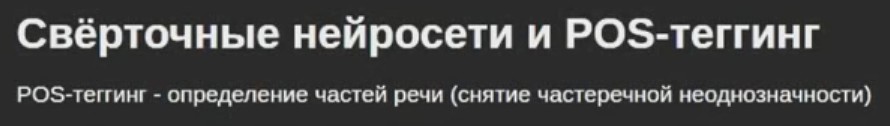
**Сокращение словаря**

**Byte Pair Encoding** - простой алгоритм кодирования, который позволяет сжать длину последовательностей за счёт расширения словаря (то есть можно выбирать разные соотношения длины последовательностей и размера словаря).

Разберём упрощённый алгоритм обучения на примере токенизации "косил косой осой".

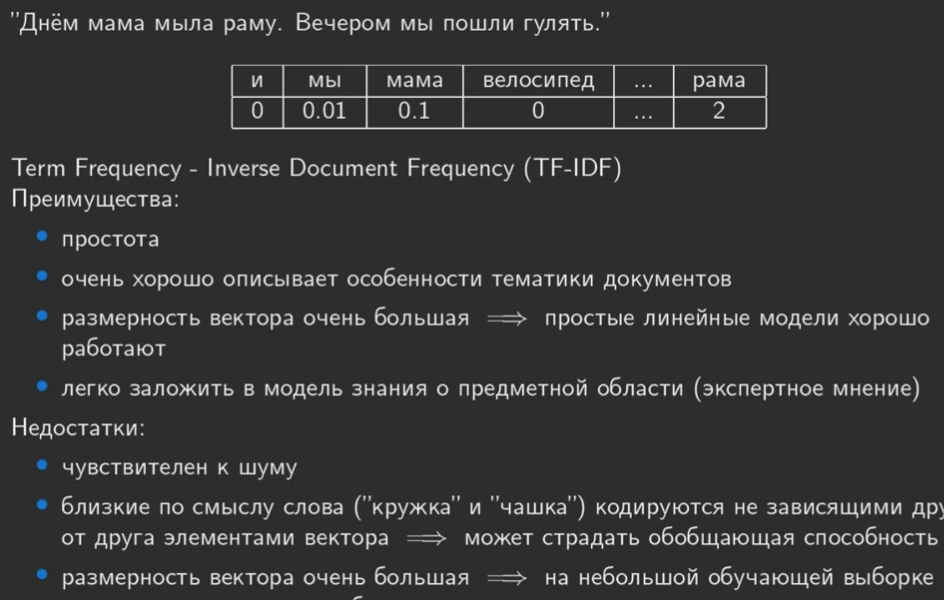
1. Найти самую частотную биграмму - по два употребления имеют "ко", "ос", "со" и "ой" - выбираем "ко".
2. Назначить выбранной биграмме номер. Мы вместо номеров будем использовать заглавные буквы. Назначили: "ко"  -> А.
3. Обновить последовательность с учётом созданного правила подстановки "ко"  -> А: "Асил Асой осой".
4. Повторяем шаги 1-3 для биграммы "со": подстановка "со"  -> Б, последовательность "Асил АБй оБй"
5. Повторяем шаги 1-3 для биграммы "Бй": подстановка "Бй"  -> В, последовательность "Асил АВ оВ".
6. Останавливаемся, когда сделаем заданное количество замен (то есть когда достигнем максимально допустимого размера словаря) или когда заменами не сможем дальше сжать тексты обучающей выборки.

**POS-теггинг**

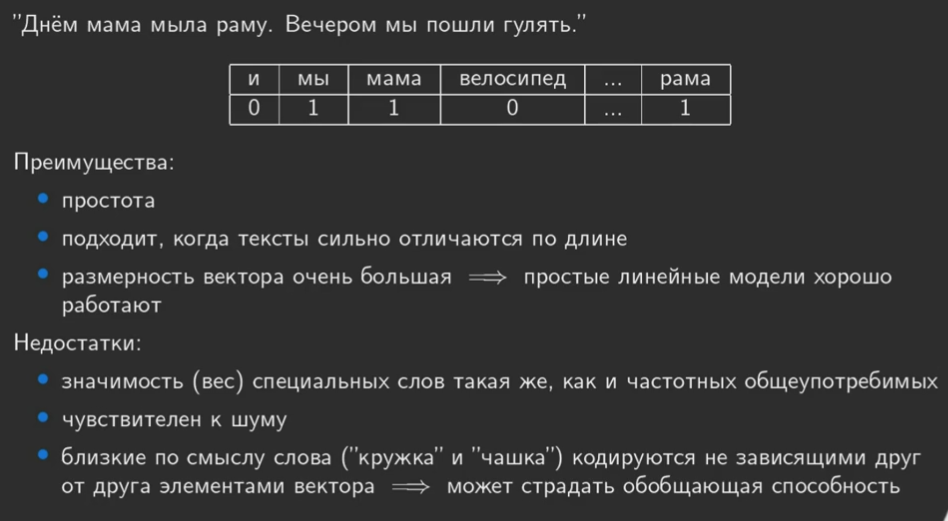


**Извлечение признаков (векторизация)**

**TF-IDF**



**Векторное представление**



**N-граммы**

