Семинарское задание №5
N-граммная языковая модель

Мосиенко Константин Викторович

2017

Необходимо написать две программы: make_model и generate.py. Первая должна строить 3-граммную языковую модель и сохранять её на диск, вторая - генерировать текст в соответствии с построенной моделью. Необходимо обратить внимание на следующие моменты:

- В качестве сглаживания необходимо воспользоваться «Jelinek-Mercer interpolation» с $\lambda_{t_{i-1}} = \lambda = 0.95$.
- Если не получается всё реализовать в оперативной памяти, make_model можно сделать в виде bash-скрипта, в котором можно запускать ваши *.py скрипты, сохраняющие что-то на диск, и стандартные linux-утилиты (sort, uniq).
- generate.py должен работать следующим образом: после загрузки модели он ждёт от пользователя ввода двух слов на одной строке через пробел, которые будут использоваться в качестве «сида». Затем он итеративно выбирает следующее слово, на основе 2-х предыдущих, случайно сэмплируя его из 20-и наиболее вероятных. Необходимо вывести предложение длиной 15-20 слов. После генерации generate.py переходит к ожиданию следующего ввода от пользователя.
- Подумайте, что необходимо предрассчитать и сохранить в модели.

Коллекцию документов можно взять со страницы курса: txt.tar.gz.