КАК СЭМПЛИРОВАТЬ?

X5 Tech

как сэмплировать?

хотим сгенерировать текст, похожий на человека

1. текст осознанный
2. текст разнообразный

жадный метод [greedy]

берём токены с наибольшей вероятностью

beam search

выбираем k наиболее вероятных токенов

для каждого из них выбираем свои k наиболее вероятных токенов

чтобы не порваться от экспоненциальности, отбрасываем k2-k вариантов, на каждом шаге оставляем только топ-k путей

greedy vs beam search

greedy

детерминированный

быстрый

можем выбрать лучшее на каждом шаге

не гарантирует разнообразия реплик

beam search

детерминированный

k = 1 -- жадный

вычислительно сложный

можно генерировать более разнообразные реплики

реплики очень общие

temperature samplings

$$p(X_i) = \frac{e^{x_i/\tau}}{\sum_{j=1}^{|V|} e^{x_j/\tau}}$$

перевзвешиваем веса

параметр τ назовём температурой

top-k sampling

$$p' = \sum_{x \in \mathrm{Vocab}} p(x_i | x_{1:i-1})$$

на каждом шаге отбираем k самых вероятных слов

введём нормировочный коэффициент

шкалируем выходы софтмакса

top-k sampling

$$p' = \sum_{x \in \mathrm{Vocab}} p(x_i | x_{1:i-1})$$

на каждом шаге отбираем k самых вероятных слов

введём нормировочный коэффициент

шкалируем выходы софтмакса

из-за разности распределений хотим иметь разный k на каждом шаге

X5 Tech

nucleus [top-p] sampling

$$\sum_{x \in \text{Vocab}} p(x_i | x_{1:i-1}) \geq p$$

берем минимальное число токенов, сумма вероятностей которых >= p