Генерация текста: выбор токенов и температура

Вероятности следующих слов уже посчитаны. Почему один и тот же вопрос даёт то одинаковые, то разные ответы?

Вероятности готовы. Осталось превратить их в текст. Для этого программа выбирает один токен, добавляет его к началу и снова вызывает модель. Способ выбора помогает управлять тем, насколько разнообразными будут продолжения.

Ранее

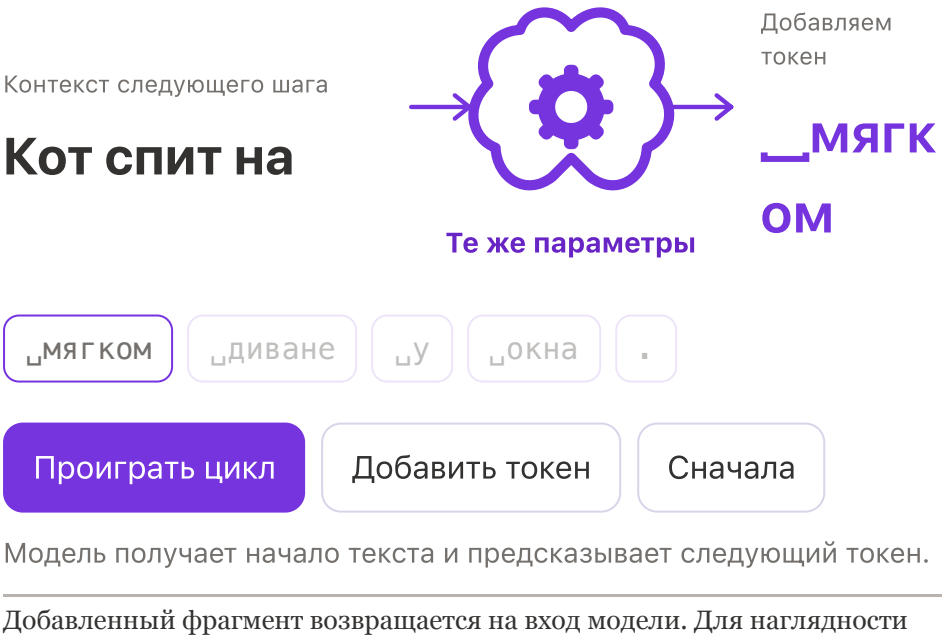
Ранее мы разобрали: токены и токенизация. Здесь используем эти идеи в новой задаче.

← Токены и токенизация

1. От распределения к последовательности

После получения распределения программа выбирает следующий токен, добавляет его к тексту и повторяет вычисление. Каждый выбранный фрагмент становится частью входа следующего шага.

Предсказать → выбрать → добавить → повторить



Добавленный фрагмент возвращается на вход модели. Для наглядности фрагменты показаны целыми словами.

Интерактивный эксперимент — в электронной версии урока.

При **жадном выборе** программа берёт токен с максимальной вероятностью. Это простое правило, но лучший локальный выбор не гарантирует лучший полный текст: после него контекст изменится, а вместе с ним изменятся дальнейшие варианты.

При **сэмплировании** токен выбирают случайно по распределению. Доля 70 % означает большую вероятность выбора, чем доля 20 %. Это не равномерная жеребьёвка. В десяти попытках вариант с 70 % не обязан выпасть ровно семь раз. Стратегии генерации

Разделим работу модели и программы выбора. Сеть вычисляет оценки следующих токенов. Затем программа применяет правило: взять самый вероятный вариант или провести случайный выбор с учётом долей. После этого выбранный токен присоединяется к тексту, и сеть снова получает обновлённый вход.

Поэтому два разных продолжения не обязательно означают две разные модели. Одни параметры и одно начало могут дать одинаковое распределение, а случайный выбор разойдётся. Уже после первого различия контексты станут разными, и следующие распределения тоже могут различаться.

> У токенов вероятности 70 %, 20 %, 10 %. Как работает случайный выбор по этому распределению?
>
> ○ Все три выбираются одинаково часто по правилу
>
> ○ Всегда выбирается первый
>
> ○ В десяти попытках первый обязан выпасть семь раз
>
> ○ Первый выбирается вероятнее остальных, но не обязательно в каждой попытке

2. Температура и случайный выбор

В некоторых задачах мы бы хотели, чтобы на один и тот же запрос модель могла давать разные ответы. Например, преподаватель просит придумать вопрос для обсуждения на семинаре. При повторном запуске ему может пригодиться новая формулировка или другой ракурс. Если каждый раз выбирать только самый вероятный токен, мы будем получать одно и то же продолжение.

Чтобы управлять разнообразием ответов, введём в правило выбора следующего токена специальный параметр — температуру. При нулевой температуре программа выбирает токен с наибольшей вероятностью. При положительной температуре появляется случайный выбор: более вероятные токены выпадают чаще, но шанс есть и у других. Чем выше температура, тем равномернее становятся шансы и тем чаще выбираются менее вероятные варианты. Сама модель при этом остаётся той же: мы меняем способ выбора продолжения.

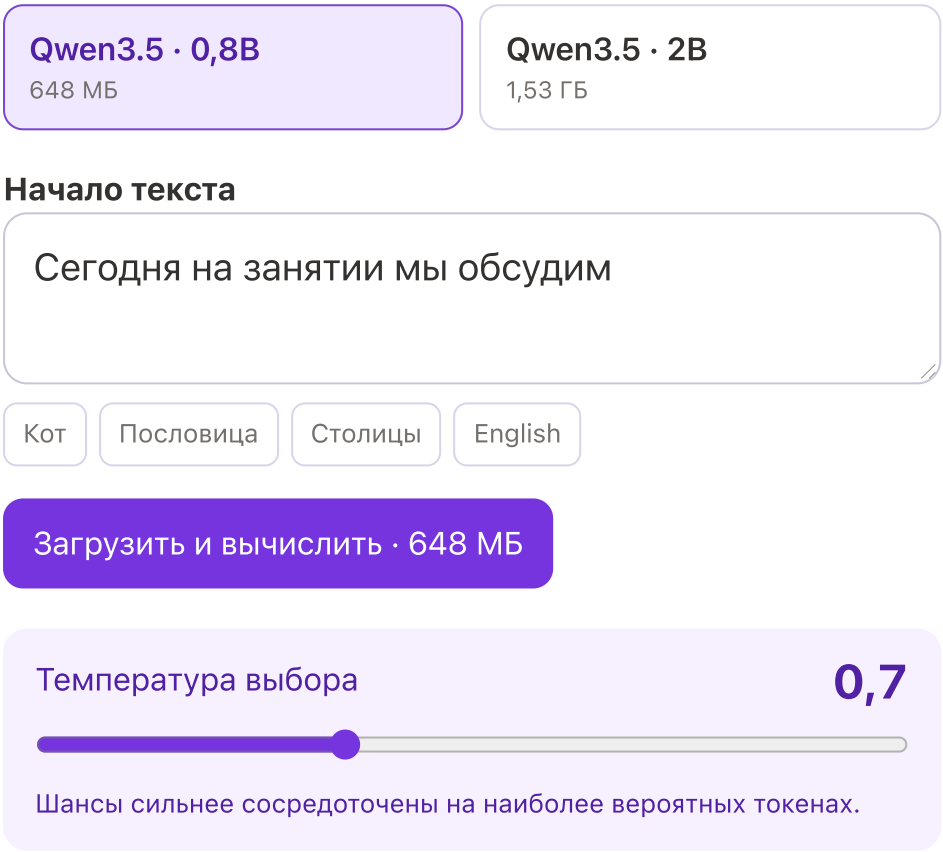
Заметка

Температура управляет распределением выбора, но не достоверностью; каждый новый токен меняет контекст.

Разделим две стадии. Сначала сеть вычисляет оценки следующих токенов. Затем программа превращает эти оценки в шансы выбора с учётом температуры. При том же входе и тех же параметрах исходные оценки сети сохраняются.

В демо два столбца: **вероятность модели** и **шанс выбора**. Не меняйте начало текста и двигайте ползунок. Левый столбец остаётся прежним; правый показывает действие температуры. При T = 1 они совпадают. При T = 0 программа выбирает максимум — ждать вероятности 100 % не нужно.

Температура на настоящей модели



Интерактивный эксперимент — в электронной версии урока.

Понизьте температуру: выбор сильнее сосредоточится на самых вероятных токенах. Поднимите её: у менее вероятных продолжений станет больше шансов. Нажмите «100 выборов без изменения текста», чтобы увидеть случайный разброс. Повторите серию с той же температурой: счётчики могут отличаться, хотя распределение осталось прежним.

Теперь добавьте токен к тексту — вручную или по шансам. Контекст изменился, поэтому модель вычислит новые исходные вероятности. Ползунок температуры меняет правило выбора, а добавление токена — условия следующего предсказания.

Температура не добавляет знаний и не проверяет логику. Если модель высоко оценивает неверное продолжение, низкая температура лишь усилит его преимущество.

> Температуру уменьшили при тех же оценках токенов. Что обычно произойдёт с распределением?
> ○ Оно сильнее сосредоточится на высоких оценках
> ○ Параметры модели переобучатся
> ○ Верные факты автоматически получат 100 %
> ○ Все доли станут одинаковыми

3. Разнообразие и достоверность

В настоящей авторегрессионной генерации выбранный токен добавляют к контексту, получают новое распределение и продолжают. На каждом шаге кандидаты и их доли могут быть другими. Поэтому длинный текст не получается повторением одного неизменного распределения.

Для остановки часто используется специальный токен конца. Также программа может остановить ответ по лимиту длины или команде пользователя. Обычная точка не обязательно завершает генерацию: после неё может начаться следующее предложение.

Управлять разнообразием полезно, но оно не равно качеству. При оценке ответов отдельно смотрят на выполнение задания, связность и достоверность. Для сравнения двух настроек нужен набор одинаковых запросов и заранее понятный критерий; один удачный пример легко вводит в заблуждение.

Нулевая температура уменьшает случайность выбора следующего токена. Она не проверяет факты и не гарантирует, что разные вычислительные системы дадут совершенно одинаковый результат.

> Модель написала точку в конце предложения. Обязана ли генерация остановиться?
> ○ Нет, остановка вообще невозможна
> ○ Нет, может появиться следующее предложение
> ○ Да, если температура больше единицы
> ○ Да, любая точка означает конец ответа

4. Выбор следующего токена

Вернитесь к живому опыту выше. Сохраните одно и то же начало и сравните два пути: сначала берите наиболее вероятное продолжение, затем выберите менее вероятный

токен вручную. Уже после первого различия контексты расходятся — вместе с ними меняются и дальнейшие прогнозы.

Попробуйте свой короткий запрос. Если требуется придумать несколько идей для обсуждения, разнообразие может быть полезно. Если нужно указать дату или проверить вычисление, случайный выбор не заменяет проверку источника или расчёт.

Для содержательного сравнения настроек используйте одинаковые запросы и заранее выбранные критерии. Один удачный или неудачный ответ ещё мало говорит о качестве модели.

Температура управляет распределением выбора, но не достоверностью; каждый новый токен меняет контекст.

Доли задают шансы, а не расписание результатов.

Остановка определяется служебным сигналом и настройками программы.

Домашняя работа

1. Лимит ответа — 15 новых токенов. Уже добавлено 11. Сколько ещё разрешено добавить до лимита?

Источники

- Стратегии генерации (https://huggingface.co/docs/transformers/llm_tutorial)

Повторить эту тему