Введение

В главе 3 вы поверхностно ознакомились с библиотекой 🤗 Datasets и увидели три главных шага для использования ее в процессе fine-tuning:

1. Загрузить датасет из Hugging Face Hub.
2. Произвести препроцессинг с помощью `Dataset.map()`.
3. Загрузить и вычислить метрики.

Но это лишь малая часть того, на что способна 🤗 Datasets! В этой главе мы углубимся в библиотеку и попутно мы найдем ответы на следующие вопросы:

- Что делать, когда нужного набора данных нет в Hub?
- Как вы можете разделиить датасет? (Что если вам *действительно* нужно использовать Pandas?)
- Что делать, когда ваш набор данных огромен и «расплавит» оперативную память вашего ноутбука?
- Что, черт возьми, такое «отображение памяти» (memory mapping) и Apache Arrow?
- Как вы можете создать свой собственный датасет и отправить его в Hub?

Принципы, которые вы изучите в этой главе, подготовят вас к более глубокому использованию токенизации и fine-tuning'а моделей в главе 6 и главе 7 – заваривайте кофе и мы начинаем!