Вступ

- Ціль проекту: створити агрегатор новин та технічних статей по тематиці використання науки про дані, машинного навчання та штучного інтелекту в бізнес операціях (маркетингу, фінансах, логістиці, менеджменті).

- Мотивація для вибору теми: у зв'язку з професійною діяльністю мене цікавить інновації у сфері бізнес операцій. Я певний період часу хотіла створити канал з агрегатором інформації у цій сфері, але вручну кожен день збирати статті було б не зручно, тому цей проект допоможе автоматизувати щоденний збір статей.

- Ідея проекту: донавчити LLM для отримання короткого опису статей (summarization), далі кожен день збирати бізнес і технічні статті, отримувати їх короткий підсумок та переводити на українську мову.

Процес роботи

- Збір і підготовка даних.
  Для донавчання моделі я використовувала відфільтровані під свій запит датасети: з технічними статтями та описами "scientific_papers" та "scitldr" зі статтями з "arxiv" та "xsum" зі новинами з BBC по тематиці проекту з Hugging Face, а також датасет Open4Business.

- Навчання та адаптація моделі.
  Для донавчання я використала модель "facebook/bart-large-cnn", яка була навчання на статтях з новинами, тому я слідкувала, щоб в датасеті переважали технічні статті з "arxiv".

- Інтеграція рішення в продукт або прототип.
  Для прогнозу я отримувала підбірку статей з новинного агрегатора, додатково парсила повний текст новини (агрегатор на безкоштовному еккаунті пропоную тільки обрізану версію статті), застосовувала донавчену модель для отримання підсумку та застосовувала іншу модель для отримання перекладу підсумку на українську мову. Результати - в файлі “summary.json”.

Виклики та їх вирішення

- Найбільших викликом стали обмежені ресурси для донавчання моделі. Також цей процес довелося починати декілька раз. Тому весь проект було простіше створювати на Google Colab.

Інструкції з запуску:

- Ноутбук data_preparation містить код для формування датасету.
- Ноутбук model містить код для дотренування моделі.
- Ноутбук inference містить код для отримання генерації результату.