**Отчёт по задаче обобщения текста. ГП1.**

**Метрики**

Мы выбрали довольно стандартные метрики для задачи seq2seq, они хорошо оценивают качество модели как внешне, то есть относительно задачи (это метрики BLEU, ROUGE и METEOR), так и внутренне (Perplexity). Было решено оставить все 4 метрики для надёжного отслеживания трейна и инференса. Коротко о каждой из них:

1. **BLEU** измеряет совпадение между автоматически сгенерированным текстом (гипотезой) и эталонным текстом (референсом) на основе n-грамм. Основная идея - оценить, насколько автоматически сгенерированный текст близок к эталонному тексту. BLEU учитывает не только полное совпадение n-грамм, но и частичное совпадение, что делает его более гибким и универсальным для оценки качества текстов. Результатом метрики BLEU является число от 0 до 1, где близкое к 1 значение указывает на более совпадающий автоматически сгенерированный текст с эталонным текстом.
2. **ROUGE** — это набор метрик, разработанных для оценки качества генерации текста. ROUGE оценивает схожесть между автоматически сгенерированным текстом и эталонным текстом, используя различные методы сопоставления текстов. Существует множество вариантов, но мы будем использовать ROUGE-N, возможно добавим ещё метрики из этого набора, если поймём, что они нужны. Как и BLEU, значения этой метрики находятся от 0 до 1, где 1 — это совпадение с эталонным текстом.
3. **METEOR** оценивает качество перевода, учитывая семантическую близость между автоматически сгенерированным текстом (или переводом) и эталонным текстом. Она учитывает слова и их порядок для определения сходства. Также эта метрика учитывает синонимы и метрики семантической близости. Кроме того, METEOR использует весовые коэффициенты для различных компонентов оценки, таких как сходство слов, фраз и порядка слов, чтобы предоставить общую метрику качества перевод. Значения метрики – от 0 до 1, где 1 – совпадение с эталонным текстом.
4. **Perplexity** — это показатель оценки качества языковой модели, который измеряет насколько хорошо модель предсказывает новые данные. Perplexity измеряет неопределенность или "удивление" модели при прогнозировании новых данных. Чем ниже значение perplexity, тем лучше модель способна предсказывать новые данные.

**Датасет**

Первоначально мы нашли около 5-7 хороших хороших датасетов с объектами для обучения модели суммаризации на русском языке. В этот список входили суммаризации новостей, вики-страниц, научных статьей, а также синтетически сгенерированных данных. Однако вместе с ментором мы приняли решение создать суммаризатор новостей, чтобы попробовать выбить качество получше. В дальнейшем мы планируем попробовать обучить модель на более широком спектре данных, но это планируется оставить за рамками выполнения проекта.

На данный момент нами было выбрано 3 датасета: [Gazeta](https://huggingface.co/datasets/IlyaGusev/gazeta), [XLSUM](https://huggingface.co/datasets/csebuetnlp/xlsum) и [MLSUM](https://huggingface.co/datasets/reciTAL/mlsum). Все они являются результатом суммаризации самых различных новостей с российских новостных порталов. Все данные были выгружены с сайта [huggingface.co](http://huggingface.co) посредством их собственной библиотеки.

Также мы провели предварительный анализ датасетов, посмотрели на разные распределения и статистики (соответствующие выкладки можно увидеть в файле data\_analyze.ipynb). Пришлось и немного почистить данные от аномалий.

По итогу мы получили один большой датасет суммаризаций российских новостей. Общее количество объектов - 160к.

**Токенизация**

В первой итерации работы над проектом мы также решили токенизировать наши данные, используя BPE токенайзер со словарем в 30к токенов. Но, возможно, в дальнейшем мы выберем другой вариант, если поймём, что он нам не подходит.