МИНОБРНАУКИ РОССИИ САНКТ-ПЕТЕРБУРГСКИЙ
ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Факультет КТИ
Кафедра АМ

Создание рефератов текста

По дисциплине «Обработка естественных языков»

Работу выполнил:

Серафим Кириллович Иванов

Группа: 0385

Преподаватель:

Малютин Егор Владимирович

Санкт-Петербург

2025

Содержание

1. Постановка задачи

Автоматически построить рефераты текстовых документов. Ввод: Массив текстов в формате JSON. Примеры текстов прикрепляю. Вывод: Массив рефератов в формате JSON (порядок рефератов соответствует порядку текстов во входных данных).

Максимальный размер каждого из рефератов — 300 символов (включая пробельные). Если размер реферата превышает указанный порог, то будут оцениваться только первые 300 символов. Тривиальное решение (первые 300 символов документа) допускается, но не приветствуется.

Дополнительные требования к оцениванию: ROUGE-2 — близость набору вручную составленных рефератов на основе биграмм слов (значение от 0 до 1).

Входные данные. массив вида

["первый текст", "второй текст"]

Выходные данные. Для каждого текст требуется вывести реферат:

["реферат первого текста", "реферат второго текста"]

2. Ход работы

2.1. Предусловия

Было принято решение вспрользоваться уже готовыми моделями "sarahai/ru-sum" и "cointegrated/rut5-base-absum", взятыми с hugging face. Моделм хорошм тем, что уже предобучены на поставленную задачу и умеет работать с русским языком. Все настройки модели можно увидеть в прикрепленном к отчету репозитории (папка LAB2).

2.2. Описание работы скрипта

Скрипт состоит из следующих логичесих частей:

1. Импортирование необходимых библиотек;
2. Инициализация модели;
3. Обработка введенного текста;

4. Оценка полученного результата с помощью априорно считающегося правильным текста.

Для тестирования был составлен набор текстов, состоящий из небольшмх рассказов, которые вручную были записаны в виде реферата в отдельную переменную.

Выходные тексты для модели "sarahai/ru-sum":

1. "В избушке в избушке живёт пёс, который звать Дик. Он любит смотреть, как я курю. По словам Дика, он не понимал, откуда берутся ягоды, чтобы не кусали комары и морошку.",
2. "Скворец Петруша любит подражать голосам и звукам. Поэт сочинял стихотворение и печатал их в журналах. Кроме того, у него есть трели, свисты, рулады, сморчки, половодье, серёжки на вербе.",
3. "Вообще-то зайцы не собирают букетов, а заячий хвост называется «пых» или «цветок». Кроме того, заяц, который собрал букет, должен быть кому-то вручён. По мнению охотников, он может и по шее накостылять.",
4. "В лесу у костра напал страх. По его словам, кто-то шептался обо мне из темноты, а также пытался поднять голову. Несмотря на то, что это шевелится там и шепчется листобой.",
5. "В ночь в печной трубе задул листобой — холодный октябрьский ветер. Несмотря на то, что она была полна берёзовых листьев, уже раскрыта настежь. По словам хозяина, они были сбрасывали листья, растущей под окном, а сам хозяин ее уже скрылся куда-то."

Выходные тексты для модели "cointegrated/rut5-base-absum":

1. "У меня в избушке живёт пёс, которого звать Дик. Он любит смотреть, как я курю. Раз на болоте я нашёл черничную поляну. Подмигнул ему и стал есть ягоды прямо с куста.",
2. "У одного поэта жил скворец Петруша. Поэт сочинял стихи и печатал их в журналах. За стихи платили деньги: рубль и сорок копеек за строчку. Скворец любит подражать голосам и звукам.",
3. "Зачем лисе или волку заячий букет? Все полевые цветы над ушами у русаков, все лесные за хвостами у беляков. Но вот объявился, глядите, заяц, который собрал букет. Не знает, кому его подарить.",
4. "Ночью в лесу у костра напал страх. Я глядел в огонь и боялся поднять голову.

Только вдруг в тишине слышался тоненький треск и короткий шёпот. Кто-то шептался обо мне.”,

5. ”Холодный октябрьский ветер пришёл с севера, из тундры, уже прихваченной льдом, с берегов Печоры. Листобой завывал в печной трубе, шевелил на крыше осиновую щепу, бил, трепал деревья.”

Невооруженным взглядов видны проблемы:

1. Тексты зачастую имеют стилистические и грамматические ошибки
2. Не всегда предложения корректны

тем не менее, везде понятно о чем идет речь и понятна суть исходного текста.

Сравнивая полученные рефераты и эталонными (рукописными) порлучены метрики, представленные в табл. 2.1 и табл. 2.2

Таблица 2.1: Результаты оценки ROUGE метрик для модели ”sarahai/ru-sum”

Пара	**Метрика**	**Recall**	**Precision**	**F1-score**
1	ROUGE-1	0.211	0.286	0.242
	ROUGE-2	0.045	0.069	0.055
	ROUGE-L	0.211	0.286	0.242
2	ROUGE-1	0.516	0.593	0.552
	ROUGE-2	0.394	0.481	0.433
	ROUGE-L	0.516	0.593	0.552
3	ROUGE-1	0.159	0.233	0.189
	ROUGE-2	0.021	0.034	0.026
	ROUGE-L	0.159	0.233	0.189
4	ROUGE-1	0.250	0.300	0.273
	ROUGE-2	0.056	0.069	0.062
	ROUGE-L	0.250	0.300	0.273
5	ROUGE-1	0.800	0.615	0.696
	ROUGE-2	0.633	0.487	0.551
	ROUGE-L	0.733	0.564	0.638

3. Выводы

Проведённое исследование показало, что применение предобученных моделей ”sarahai/ru-sum” и ”cointegrated/rut5-base-absum” для автоматического построения рефератов русскоязычных текстов даёт результаты удовлетворительного качества. На основании метрик ROUGE:

Таблица 2.2: Результаты оценки ROUGE метрик для модели "cointegrated/rut5-base-absum"

Пара	**Метрика**	**Recall**	**Precision**	**F1-score**
1	ROUGE-1	0.316	0.400	0.353
	ROUGE-2	0.091	0.133	0.108
	ROUGE-L	0.316	0.400	0.353
2	ROUGE-1	0.452	0.519	0.483
	ROUGE-2	0.333	0.379	0.355
	ROUGE-L	0.452	0.519	0.483
3	ROUGE-1	0.227	0.323	0.267
	ROUGE-2	0.063	0.097	0.076
	ROUGE-L	0.227	0.323	0.267
4	ROUGE-1	0.567	0.431	0.294
	ROUGE-2	0.528	0.036	0.031
	ROUGE-L	0.467	0.431	0.294
5	ROUGE-1	0.167	0.192	0.179
	ROUGE-2	0.067	0.077	0.071
	ROUGE-L	0.167	0.192	0.179

1. Средние значения F1-оценок для ROUGE-1 и ROUGE-L варьируются от 0.18 до 0.69, что свидетельствует о частичном совпадении автоматически сгенерированных и эталонных рефератов;
2. Наилучшие результаты наблюдаются для текстов, обладающих более чёткой синтаксической структурой и минимальным количеством грамматических ошибок;

Выбранные модели справляется с задачей выделения основной мысли текста, однако требует дополнительной постобработки или адаптации для повышения согласованности и точности формулировок. Для улучшения качества рекомендуется дообучение моделей на корпусе русскоязычных текстов с вручную составленными аннотациями и применение более современных архитектур.