ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

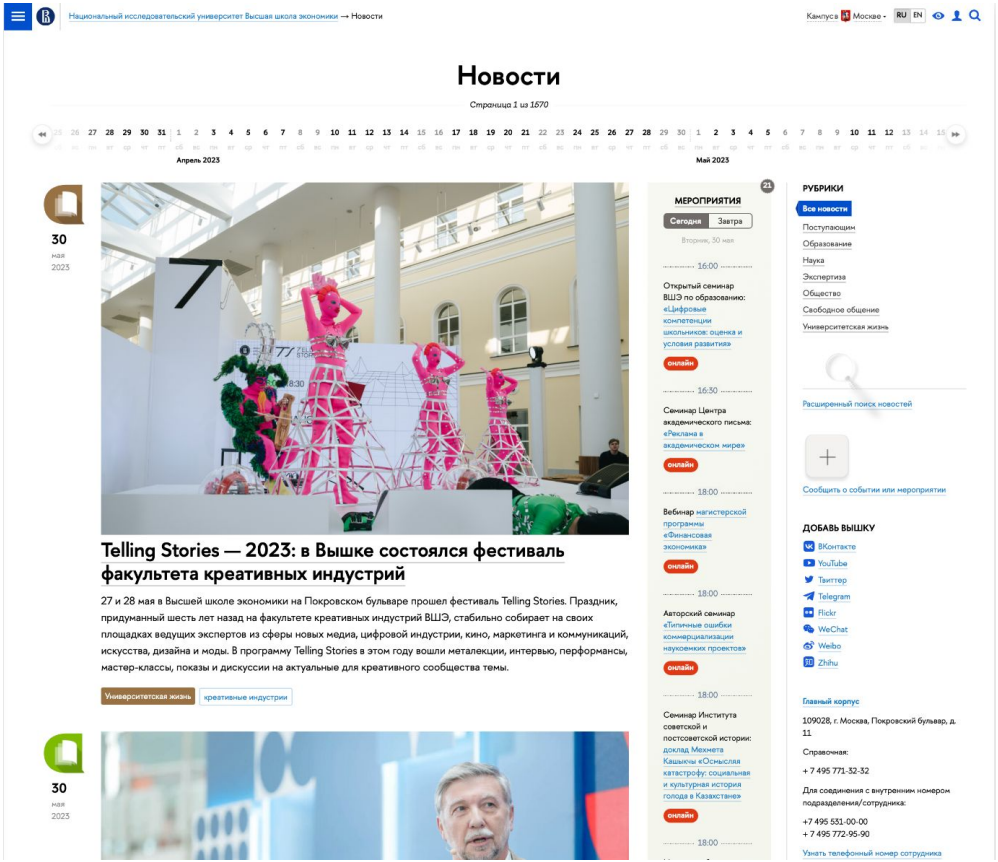
Программный проект на тему:

Применение Методов Автоматического Сокращения Текста для Аннотирования Статей Новостного Портала

Выполнил: Кульпин Павел Леонидович

Руководитель ВКР: Лобачевский Семен Михайлович

Ожидаемые результаты

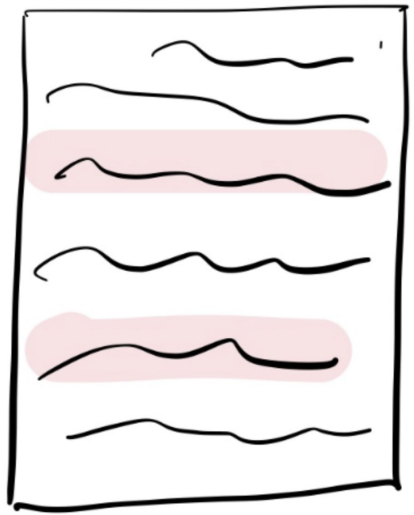


Онлайн-сервис, в котором редакторы смогут генерировать аннотации свежих статей.

Рисунок 1. https://www.hse.ru/news/

Задача суммаризации текста

- **Экстрактивный подход**



- **Абстрактивный подход**

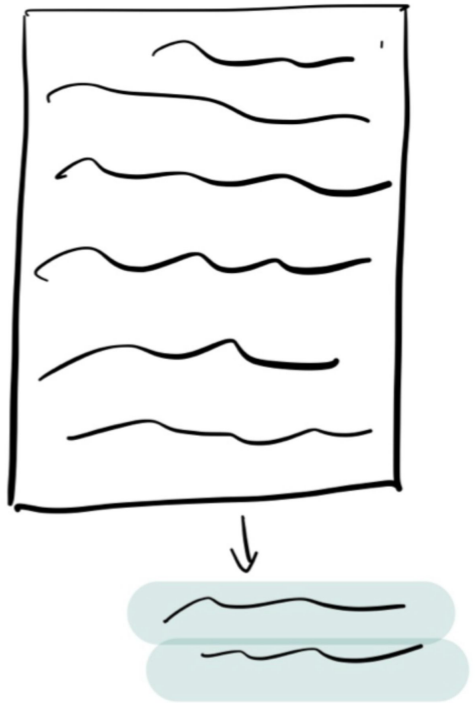


Рисунок 2. Иллюстрация

Набор данных - hse.ru/news

Директор Института медиа Эрнест Мацкявичюс стал лауреатом Премии Первых

28 марта в «Сириусе» состоялась церемония запуска четвертого сезона самого масштабного проекта для детей и подростков — Всероссийского конкурса «Большая перемена». Директору Института медиа Эрнесту Мацкявичюсу была вручена Премия Первых в номинации «За вклад в развитие «Движения Первых».

© Пресс-служба «Большой перемены»

Журналист, ведущий программы «Вести в 20:00» телеканала «Россия 1» Эрнест Мацкявичюс — один из главных наставников «Большой перемены». Его лекция «Простые способы быть убедительным» открыла двухдневный марафон, посвященный старту конкурса.

«Большая перемена» — флагманский проект Российского движения детей и молодёжи «Движение первых». Начало нового сезона собрало победителей трех предыдущих конкурсов, педагогов-наставников, деятелей искусства и представителей медиа, олимпийских чемпионов, ректоров вузов, руководителей детских общественных и образовательных проектов, федеральных министерств и ведомств. Более 2000 школьников и студентов, а также 1000 педагогов, чьи ученики стали победителями «Большой перемены», поделились историями успеха, рассказали о своих проектах и вспомнили самые запоминающиеся моменты состязания. В первый день участники встретились с директором департамента информации и печати МИД РФ Марией Захаровой, начальником управления Президента РФ по общественным проектам Сергеем Новиковым, министром просвещения РФ Сергеем Кравцовым, генеральным директором АНО «Большая Перемена» Наталией Мандровой, директором киностудии «Союзмультфильм» Борисом Машковцевым.

Рисунок 3.
https://www.hse.ru/ba/media/news/823682579.html

Заявить экспедицию на конкурс проекта «Открываем Россию заново» можно до 15 июня

© Алексей Евдокимов

Проект «Открываем Россию заново» президентской платформы «Россия — страна возможностей», Высшей школы экономики и программы «Больше, чем путешествие» приглашает университеты, коммерческие и некоммерческие организации подать заявку на организацию и проведение студенческих экспедиций в 2023 году. Подать заявку можно до 15 июня, победившие экспедиции получат финансовую и организационную поддержку.

В заявке определяются тема, концепция и примерная программа экспедиции. Экспедиция должна соответствовать одному из 6 направлений: образование, экология, социальная сфера, урбанистика и развитие территорий, культура и технологии.

«Открываем Россию заново» — всероссийская межуниверситетская программа студенческих экспедиций. Учащиеся лучших вузов участвуют в социально значимых проектах, меняющих жизнь в регионах к лучшему. Организаторы программы — НИУ ВШЭ и платформа «Россия — страна возможностей». Партнеры — программа «Больше, чем путешествие», сеть Точек кипения. В рамках программы «Открываем Россию заново» студенты проявляют профессиональные компетенции, отправляются в экспедицию для работы над реальными проектами, а региональные команды получают поддержку лучших молодых специалистов страны.

Рисунок 4. https://www.hse.ru/news/edu/836677308.html

Игры, в которые играют женщины

13 марта Лаборатория экономико-социологических исследований ГУ-ВШЭ (ЛЭСИ) провела семинар "Игры, в которые играют женщины" (по материалам к/ф "Прогулка" А. Учителя). Мероприятие посвящалось Международному женскому дню

13 марта Лаборатория экономико-социологических исследований ГУ-ВШЭ (ЛЭСИ) провела семинар "Игры, в которые играют женщины" (по материалам к/ф "Прогулка" А. Учителя). Мероприятие посвящалось Международному женскому дню, и как бы это ни было непривычно после недавней серии семинаров "Герои нашего времени", в центре дискуссии на этот раз оказалась Героиня.

В роли докладчика по традиции выступил руководитель ЛЭСИ д.э.н., проф. Вадим Радаев. В качестве оппонентов были приглашены к.и.н., проф. кафедры прикладной политологии Вышки Сергей Медведев и к.и.н. доцент кафедры истории философии Школы Виталий Куренной.

В начале своего выступления В. Радаев попытался определить жанр "Прогулки". Он остановился на том, что фильм представляет собой романтическую городскую молодежную комедию. Вкратце напомним, что по сюжету главная героиня (Оля Малахова) заключает пари со своим немолодым женихом ("папаней") о том, что она целый день проведет, гуляя по городу, при этом ни разу не присядет и еще приведет с собой двух свидетелей, которые это подтвердят. О мотивах и сути этого пари и рассуждали участники семинара. В. Радаев предложил четыре возможных интерпретации поведения главной героини:

1. В "феминистической версии" пари рассматривалось как самоутверждение, желание героини доказать всем, что "я на самом деле могу, особенно если захочу").

2. В "поисковой версии" прогулка героини по Питеру объяснялась как "поиск недостающего" - ведь каждый из трех Олиных "кавалеров" по характеру дополняет двух остальных.

3. В "игровой версии" мотивы героини свелись к созданию искусственных рамок, выдуманного мира, который контролируется самой женщиной.

4. В "манипулятивной версии" поведение героини было позиционировано как "оттачивание мастерства" соблазнения.

Судя по всему, "манипулятивная версия" была особенно дорога сердцу докладчика: ей В. Радаев уделил наибольшее внимание. В частности, он предложил рассматривать пять основных уровней овладения "мастерством манипуляции". Главным критерием "классификации" было количество мужчин, которым женщина способна манипулировать одновременно. По такой шкале главная героиня фильма "Прогулка" получила наивысшую отметку. Впрочем, в общей атмосфере Международного женского дня В. Радаев настоятельно отрицал циничность манипулятивного поведения, ведь "действительно манипулировать

Рисунок 5. https://www.hse.ru/news/communication/22282.html

Набор данных - hse.ru/news

Очистить от:

- HTML-кода и сопутствующих объектов
- ссылок
- номеров телефонов
- адресов почт
- спецсимволов
- референсов и аппендикса

Отфильтровать объекты:

- Слишком короткие и слишком длинные тексты
- Отношение длины аннотации к длине статьи > 0.5

	Статья	Аннотация
Число объектов	176428	
Среднее число слов	535.7	36.5
Среднее число предложений	24.4	1.8
Минимальное число слов	53	4
Максимальное число слов	3023	232
Число уникальных слов	1893027	304447
Число уникальных лемм	611524	127881
Среднее число уникальных слов	290.0	31.7
Среднее число уникальных лемм	221.8	28.5

Таблица 1. Статистики очищенного набора данных

Архитектуры и метрики

Базовое решение:

- TextRank + TF-IDF (300-dim navec)

Sequence-to-Sequence модели:

- T5
- BART
- GPT

Метрики качества:

- BLEU
- ROUGE-N
- ROUGE-L
- ROUGE-Lsum

Требования и ограничения

- Наиболее важны **качество суммаризации** и **связность генерации**
- Ограничение веса модели: GPU на **6 гб**
- **Не можем попробовать все** существующие модели

Результаты обученных под задачу моделей

	R-1	R-2	R-L	R-Lsum	BLEU
TextRank + TF-IDF (navec)	**15.9**	6.3	11.4	11.0	3.8
mT5_multilingual_XLSum	7.1	2.3	6.6	6.7	4.0
mT5_m2o_russian_crossSum	3.4	0.5	3.4	3.4	2.3
mbart_ru_sum_gazeta	13.5	5.7	13.1	13.2	**12.8**
rut5_base_sum_gazeta	13.7	5.7	**13.4**	**13.5**	11.9
rugpt3medium_sum_gazeta	10.1	**7.5**	9.7	9.8	6.8
rut5-base-absum	9.9	4.0	9.5	9.6	7.9

Таблица 2. Результаты доступных обученных моделей

Сравнение моделей, обучавшихся без учителя

	R-1	R-2	R-L	R-Lsum	BLEU
rugpt3large_based_on_gpt2	4.6	2.2	4.4	4.4	3.7
ruT5-large	5.4	2.6	5.2	5.5	**4.8**
FRED-T5-large	**5.7**	**4.0**	**5.4**	**5.7**	**4.8**

Таблица 3. Результаты обучения на 4000 объектов

Дообучение выбранных моделей

AdamW:

- LR = 0.0001
- weight_decay = 0.01

Cosine Annealing Scheduler:

- num_cycles = 0.5
- warmup_steps - 2000 объектов

	R-1	R-2	R-L	R-Lsum	BLEU
TextRank + TF-IDF (navec)	15.9	6.3	11.4	11.0	3.8
mbart_ru_sum_gazeta	29.7	17.2	29.3	29.4	26.1
rut5_base_sum_gazeta	**30.4**	**19.3**	**30.1**	**30.2**	**29.6**
rugpt3medium_sum_gazeta	14.2	12.6	13.7	14.1	6.9
FRED-T5-large	29.7	17.9	29.5	29.5	9.4

Таблица 4. Результаты моделей после дообучения

Примеры генераций

«Экология предоставляет лучшие возможности для профессионального и личностного развития молодежи»

Фото: пресс-центр Международной детско-юношеской премии «Экология — дело каждого»

В начале мая Институт экологии НИУ ВШЭ и Международная детско-юношеская премия «Экология — дело каждого» провели в Дагестане совместный семинар, где обсудили подготовку молодежных экологических проектов для федеральных и международных конкурсов. На встрече эксперты института представили методики организации проектной деятельности в сфере экологии и устойчивого развития для педагогов и молодежи Дагестана. Участие в мероприятии приняли преподаватели и учащиеся более 50 школ, колледжей и вузов республики.

Семинар состоялся при поддержке Министерства образования и науки Республики Дагестан, Северо-Кавказского межрегионального управления Росприроднадзора по Республике Дагестан и Дагестанского государственного университета.

Кинаят Юнусова, сотрудница Министерства образования и науки Республики Дагестан, отметила, что приоритетной деятельностью ее коллег является создание условий для всестороннего развития и обучения детей, а преференции для абитуриентов при поступлении в вузы мотивируют старшеклассников на участие в федеральных конкурсных мероприятиях с инициативными проектами в сфере экологии и устойчивого развития.

Гурам Султанов, начальник отдела государственного экологического надзора и надзора на море по Республике Дагестан Северо-Кавказского межрегионального управления Росприроднадзора, отметил важность выполнения стратегии экологической безопасности в регионе и актуальность экологического

Рисунок 6. https://www.hse.ru/news/life/834095118.html

mBART	Ирина Резанова, заместитель директора Института экологии, руководитель Ассоциации экошкол "Лидеры устойчивого развития" рассказала об опыте ЭкоШколы НИУ ВШЭ.
ruT5-base	В НИУ ВШЭ состоялся семинар Международной детско-юношеской премии "Экология - дело каждого", в котором приняли участие представители федеральных и региональных партнеров.
FRED-T5-large	Семинар состоялся при поддержке Министерства образования и науки Республики Дагестан. На семинаре обсуждались вопросы взаимодействия образовательных организаций НИУ ВШЭ и Республики Дагестан для реализации проектов в области экологии и устойчивого развития. В качестве спикеров выступили сотрудники Института экологии Ирина Резанова, Наталья Поротникова, Гурам Султанов, Кинаят Юнусова, Наталья Коротаева, Наталья Шакина. В мероприятии также приняли участие эксперты Международной детско-юношеской премии "Экология - дело каждого" Светлана Радионова, Наталья Рощина, Наталья Юсуфова.

Таблица 5. Примеры генераций 1

Примеры генераций

Большие вызовы: в Вышке прошла проектная сессия, посвященная Ямалу

© Высшая школа экономики

В стенах Высшей школы экономики состоялась проектная сессия «Большие вызовы Ямала». Ее спикерами стали губернатор Ямало-Ненецкого автономного округа Дмитрий Артюхов, ректор НИУ ВШЭ Никита Анисимов и проректор НИУ ВШЭ Дмитрий Земцов. Студенты Вышки спроектировали концепции экспедиций в регион, в которые они смогут отправиться в 2023–2024 годах в рамках проекта «Открываем Россию заново».

«Для нас сегодняшняя встреча — действительно событие знаковое. Дмитрий Артюхов возглавляет комиссию по молодежной политике. Для него общение с молодежью — это возможность понимать еще лучше запросы молодых людей. ВШЭ — это 50 тысяч студентов, средний балл которых при поступлении — 96. Это лучшие студенты нашей страны. То, что сегодня наши студенты приглашены на Ямал, — для нас большая честь и большая радость. В прошлом году была организована одна поездка на Ямал, в этом году будет две. Это свидетельство того, что наши отношения с регионом развиваются», — отметил ректор НИУ ВШЭ Никита Анисимов.

Рисунок 7. https://www.hse.ru/news/life/833947535.html

mBART	Губернатор Ямало-Ненецкого автономного округа (ЯНАО) Дмитрий Артюхов встретился студентами НИУ ВШЭ, которые планируют поехать в экспедицию <на> Ямал.
ruT5-base	В рамках проекта студенческих экспедиций «Открываем Россию заново» состоялась встреча студентов НИУ ВШЭ с губернатором ЯНАО Дмитрием Артюховым.
FRED-T5-large	В рамках проекта студенческих экспедиций «Открываем Россию заново» состоялась встреча студентов НИУ ВШЭ с губернатором Ямало-Ненецкого автономного округа Дмитрием Артюховым. В ходе встречи студенты представили свои идеи экспедиций, которые будут реализованы в ЯНАО в 2023 году, а также презентовали регион будущим участникам экспедиций. Студенты представили свои предложения концепций экспедиций губернатору, который ответил на несколько вопросов. В конце встречи Дмитрий Артюхов вручил студентам сертификаты участия проекте. В рамках экспедиции студенты смогут посетить города Салехард, Муравленко, Лабытнанги, Новый Уренгой и Ноябрьск. В течение двух месяцев студенты будут проектировать концепции экспедиций регионам, которые войдут состав экспеди

Таблица 6. Примеры генераций 2

Примеры генераций

НИУ ВШЭ и «Самолет» подготовят специалистов анализа данных в девелопменте

© iStock

Факультет компьютерных наук НИУ ВШЭ и одна из крупнейших российских девелоперских компаний «Самолет» открывают новую магистерскую программу «Анализ данных в девелопменте». «Самолет» профинансирует обучение 30 лучших абитуриентов, поступивших на программу.

Что такое анализ данных в девелопменте сегодня

Девелопмент — та область, в которой качественный скачок благодаря использованию data-driven-методов принятия решений еще впереди. Строительство входит в перечень ключевых отраслей экономики РФ, в которых запланировано масштабное внедрение искусственного интеллекта. В девелопменте будущего конкурентное преимущество получат компании, обладающие хорошей технической базой и принимающие решения на основе аналитики и данных, что позволит быстро реагировать на изменения рынка.

НИУ ВШЭ и «Самолет», лидеры в своих областях, решили объединить усилия в рамках образовательного проекта. Уже сегодня компания создает решения на основе аналитики данных для планировок и проектов на сайте, отделки и дизайна, прогнозов намерений клиентов приобрести продукт, оценки NPS пользователей, стратегии коммуникации с клиентами с задолженностями по ЖКХ и т.д.

Рисунок 8. https://www.hse.ru/news/edu/833915250.html

mBART	НИУ ВШЭ и «Самолет», лидеры в своих областях, решили объединить усилия в рамках образовательного проекта. Уже сегодня компания создает решения на основе аналитики данных для планировок и проектов на сайте, отделки и дизайна, прогнозов намерений клиентов приобрести продукт, оценки NPS пользователей, стратегии коммуникации с клиентами с задолженностями по ЖКХ и т.д.
ruT5-base	Магистерская программа "Анализ данных в девелопменте" НИУ ВШЭ совместно компанией "Самолет" запускает новый образовательный проек
FRED-T5-large	Магистерская программа факультета компьютерных наук НИУ ВШЭ "Анализ данных в девелопменте" создана в партнерстве с компанией "Самолет - одним из крупнейших девелоперов в России. На программе будут готовить специалистов в области анализа больших данных, которые смогут работать самых разных сферах бизнеса: строительстве, логистике, управлении логистической инфраструктурой, управлении цепями поставок, маркетинге, консалтинге, IT-разработке. Обучение программе бесплатное. Финансирование со стороны компании покроет 100% стоимости обучения 30 лучших абитуриентов, поступивших программу. Остальные места на программе будут платными. Прием на программу будет вестись по результатам конкурса портфолио, собеседованиям, в оценке которых примут участие эксперты факультета компьютерной науки НИУ Высшая школа экономики.

Таблица 7. Примеры генераций 3

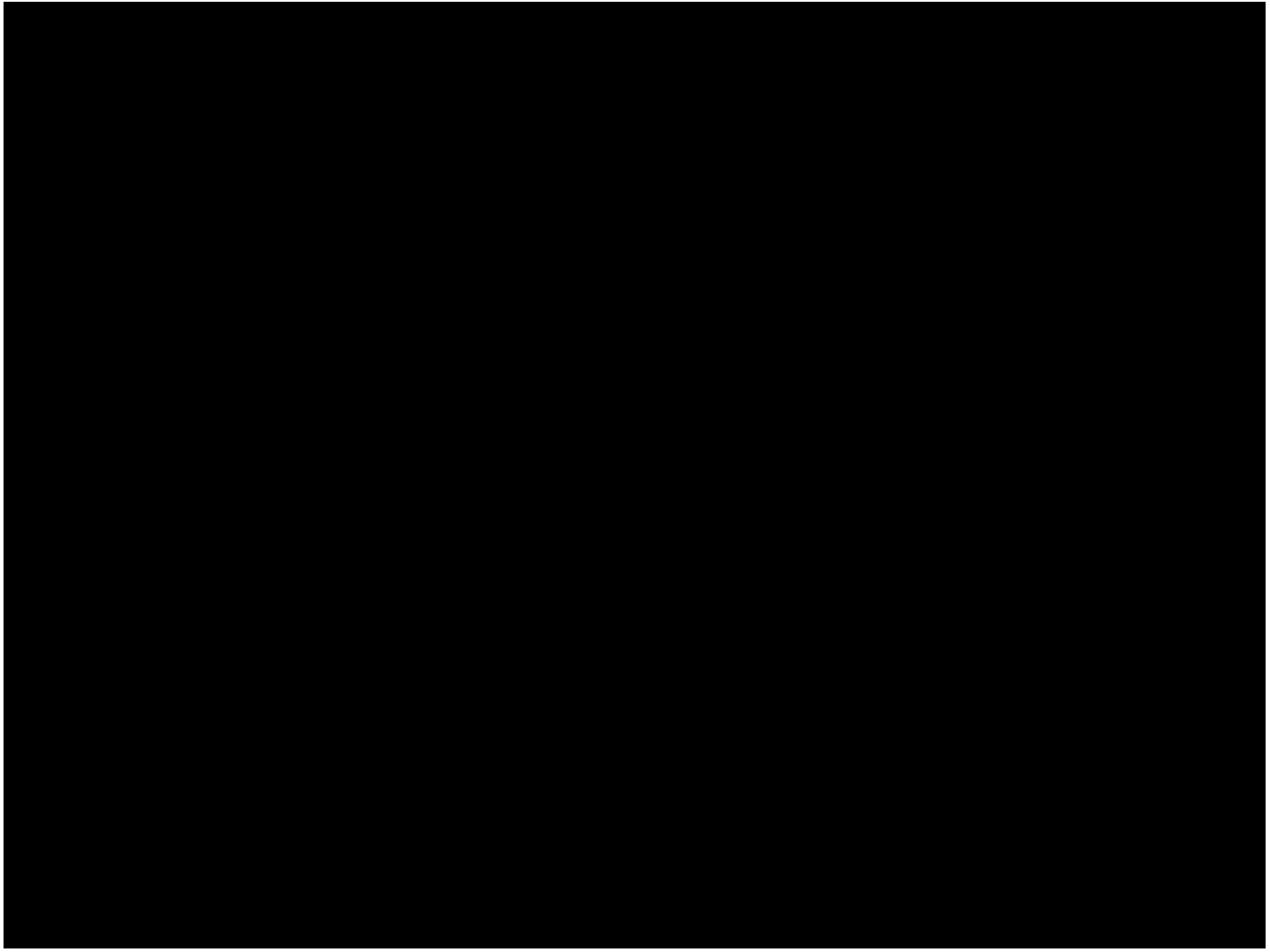
Имплементация сервиса

Инструменты:

- Streamlit
- FastAPI
- requests

Параметры генерации:

- Модель
- length_penalty
- repetition_penalty
- no_repeat_ngram_size
- num_beans
- temperature
- min_length, max_length
- top_k, top_p



Дообучение моделей и план дальнейших работ

Скрипты:

- Обрабатывающий данные описанным образом
- Обучающий/тестирующий любую из описанных моделей на переданной выборке

План:

- Имплементация в новостной портал
- Большие мощности ⇨ большие языковые модели
- Масштабирование ⇨ личные кабинеты, базы данных

Заключение

- Подготовка размеченного набора данных
- Анализ и обучение языковых моделей
- Имплементация онлайн-сервиса и скриптов, совершенствующих систему

F.A.Q.

Есть ли ограничения на входной текст?

- Да, для моделей mBART и FRED-T5-large - 1500 слов, для ruT5-base - 3000. Ограничение искусственное, иначе память GPU может переполниться.

Сколько времени занимает генерация?

- Это сильно зависит от длины текста и модели. Для средней длины текста (535 слов): ruT5-base - **7-12 сек**, FRED-T5-large – **15-20 сек**, mBART – **17-23 сек**.

“Крутится” ли скрипт в ожидании поступления данных (то есть инициализированы ли модели на постоянной основе)?

- Нет, модели инициализируются в теле асинхронной функции при поступлении HTTP-запроса. Это занимает больше времени, но позволяет использовать любое количество моделей на GPU с низкой производительностью.

Спасибо за внимание!