SlavSky team

Разработка программного решения — генератора названий для новостных подборок (сюжетов) Интерфакса

Август, 2021

Проблематика

Задача

Разработать программное решение, которое будет в автоматическом режиме генерировать названия для тематических подборок (сюжетов).

Решение

1. В процессе решения **была выявлена проблема недостатка данных** для обучения моделей, поэтому **мы произвели дополнительный сбор данных путем парсинга** сайта Интерфакс.

2. **Было обучено несколько моделей суммаризации**, чтобы в итоге выбрать ту, которая делает генерацию названий тематических подборок наиболее точным образом.

Стек решения

Python, nltk, sklearn, pytorch

Дальнейшая Реализация Решения

0

Хакатон

- Первичный сбор данных;
- Построение моделей и их сравнение по метрикам качества.

1

Ресерч (1-2 мес./ 150-300 т.р.)

- Дополнительный сбор данных;
- Улучшение текущих моделей, построение новых и сравнение;
- Написание отчета с необходимыми рекомендациями.

2

Пилотирование (1-3 мес./ 150-300 т.р.)

- Тестирование и дообучение итоговой модели: на данном этапе система будет предлагать писателю сюжетов свой (автоматически сгенерированный) вариант названия, писатель будет принимать или отвергать сгенерированное название и отправлять свой ответ обратно на сервер. Это позволит лучше оценить и дообучить модель на новых данных и принять решение о выводе системы в продакшн.

3

Продакшн (1-3 мес./ 100-300 т.р.)

- Внедрение итоговой модели в бизнес процессы заказчика;
- Написание необходимого ПО.

Новаторские идеи/ фичи проекта

В процессе решения задачи хакатона была реализована **система на основе методов машинного обучения** для решения задачи суммаризации.

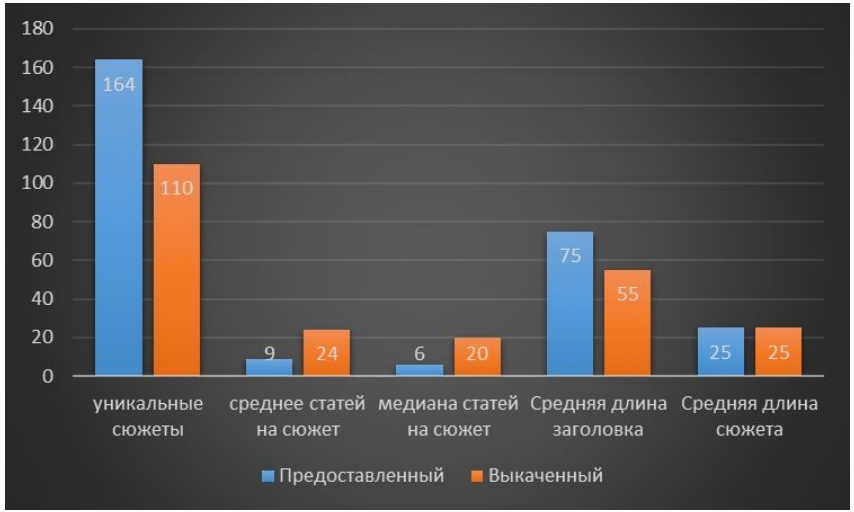
Исходный **датасет был дополнен** текстами, для которого применялся препроцессинг и использовались заголовки и/или тексты новостей.

В ходе решения валидировались результаты **двух независимых моделей**.

Прототип

Аналитика: Данные

Сравнительный анализ датасетов



- 1511 предоставленных новостей в сюжетах
- 2640 дополнительно загруженных новостей в сюжетах

Датасеты объединялись в файл единого формата и происходило разбиение датасета на **трейн**, **валидацию** и **тест**.

Аналитика: Алгоритмы

Baseline

Алгоритм, основанный на n-граммах и максимальной tf-idf матрице.

Finetuned mT5 model

Производилось объединение заголовков в сюжетах в единый текст и mT5 with finetuning модели на новом датасете.

Результаты

Результаты работы модели mT5 на тесте

target_text	preds
Нападения в школах	Нападение на школе в Нагорном Карабахе
Афганистан во власти та	Уход США с Талибами
Аварии в Москве	ДТП с автобусом подмосковье
Евро-2012	Евро
Вакцина от COVID-19	Вакцинны от COVID-19
Кубок Кремля	Кубок Кремле
Долг РФС перед Капелло	Задолженность с Капелло
"Ролан Гаррос"-2016	"Ролан Гаррос"-2016
Нобелевская премия - 20	Нобелевская премия - 2019
Пожар на территории "Н	Пожар на ТЭмскнефтехима
Трамп в соцсетях	Блокировка Twitter
Авиасообщение	Авиаперелеты
Обогащение урана Иран	Оборонение урана
Лукашенко	Белорусская оппозиция
Отношения России и Евр	Навальный кризис в РФ
Заявления Байдена	США с военным кризисом кризису
ДТП под Сызранью	ДТП с автобусом под Новой сел
Поставки нефти	Экономический кризис ОПЕК+

Метрики качества на валидации и тесте

model	rouge-1f	rouge-2f	rouge-lf	bleu
mT5 with fine-tuning	24.2	8.15	22.94	26.35
baseline	19.98	7.0	18.36	23.71

Наша Команда

Андрей Власов

NLP Data Scientist,
Sber

Николай Швецов

Data Scientist,
Helmholtz Association

Анна Клюева

Intern Data Scientist,
Sbermarket

Спасибо за внимание!