



[МОВС 23, Годовой проект] Классификатор новостей

Куратор проекта:
Павлова Арина

Выполнили студенты магистратуры НИУ ВШЭ:
Денис Чужмаров
Чайников Константин

Описание проекта

Цели проекта проекта

Автоматизация Процесса Анализа Новостей:

Разработка алгоритмов для автоматического обработки новостных данных.

Автоматизация Процесса Поиска Похожих Новостей:

Разработка алгоритмов для поиска похожих новости.

Обеспечение Удобного Доступа к Результатам Работы:

Создание асинхронного бота в Telegram для удобного доступа пользователей к результатам работы

Создание Основы Для Будущего Продукта:

В следующей главе нашего проекта мы планируем уточнить тему и использовать методы глубокого машинного обучения и для достижения результата. Текущая работа станет основой для будущих исследований

Задачи проекта проекта

Подбор и Сбор Данных:

- Выбор источников новостных данных.
- Разработка методов сбора и хранения данных.

Анализ и Обработка Данных:

- Проведение статистического анализа собранных данных.
- Предварительная обработка данных для подготовки к моделированию.

Разработка Моделей Машинного Обучения:

- Обучение модели логистической регрессии для классификации новостей.
- Обучение модели случайного леса для классификации новостей.
- Создание алгоритма для определения схожести между различными новостными статьями.

Разработка и Интеграция Телеграм-бота:

- Программирование асинхронного бота для общения с пользователями.
- Интеграция бота с разработанными моделями машинного обучения.
- Разработка удобного пользовательского интерфейса для взаимодействия с ботом.

Ресурсы проекта

Все результаты работы можно найти на GitHub:

https://github.com/konstantinator/hse_project - часть о машинном обучении

Включая:

- Ноутбук с EDA (Exploratory Data Analysis)
- Обученные модели (.pkl)

<https://github.com/Res0nanceD/hse-project-bot.git> - продакшн часть

Включая:







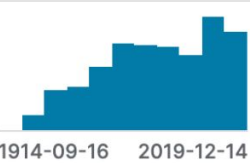
- Исходники телеграм-бота
- Скрипты для установки и docker файл
- Тесты бота и функций

Источники данных

<https://www.kaggle.com/datasets/yutkin/corpus-of-russian-news-articles-from-lenta>

lenta-ru-news.csv (2.08 GB)

Detail Compact Column

 url	 title	 text	 topic	 tags	 date
URL of the news article	Title of the news article	Body of the news article	Topic of the news article	Tags of the news article	Date of the news article
800964 unique values	797832 unique values	800038 unique values	<div><div>Россия20%</div><div>Мир17%</div><div>Other (503909)63%</div></div>	<div><div>Все57%</div><div>Политика5%</div><div>Other (306497)38%</div></div>	
https://lenta.ru/news/1914/09/16/hungarn/	1914. Русские войска вступили в пределы Венгрии	Бои у Сопоткина и Друскеник закончились отступлением германцев. Неприятель, приблизившись с севера к...	Библиотека	Первая мировая	1914/09/16

Информация о датасете

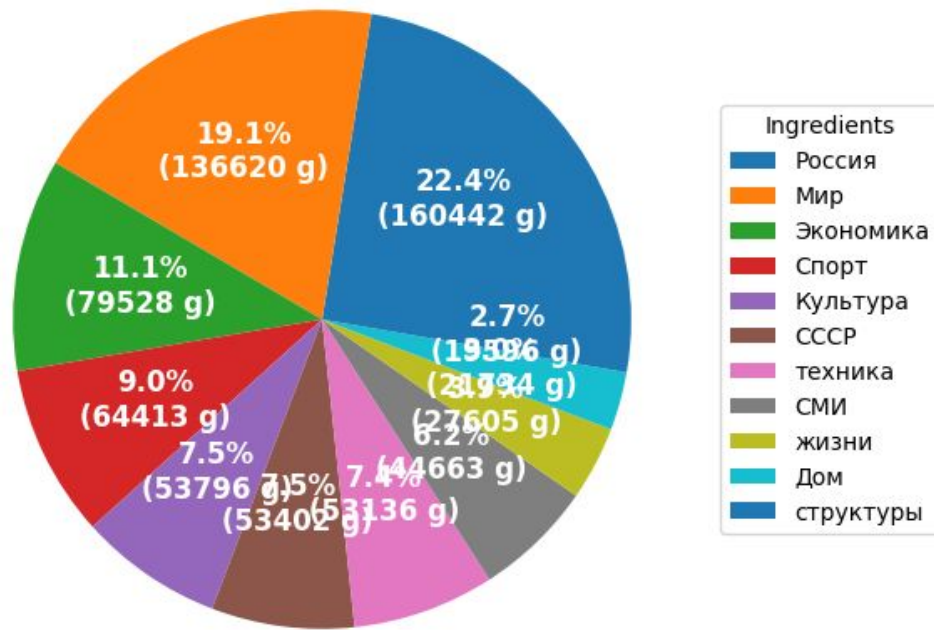
Работа над данными

- Из изначального дата сета из lenta.ru мы отобрали новости, принадлежащие только тем темам, по которым написано более 10000 постов
- Из изначального дата сета были удалены ссылки, поскольку являются излишней информацией
- Из изначального дата сета были удалены тэги, поскольку являются излишней информацией

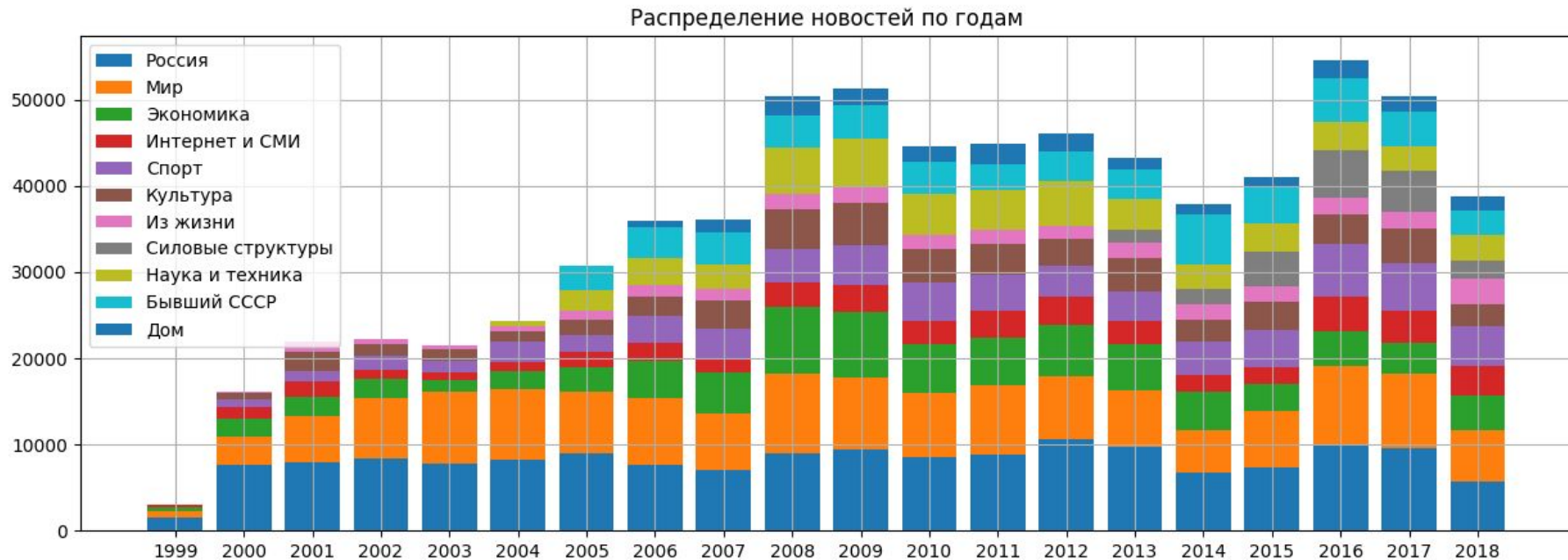
EDA

Распределение тем новостей

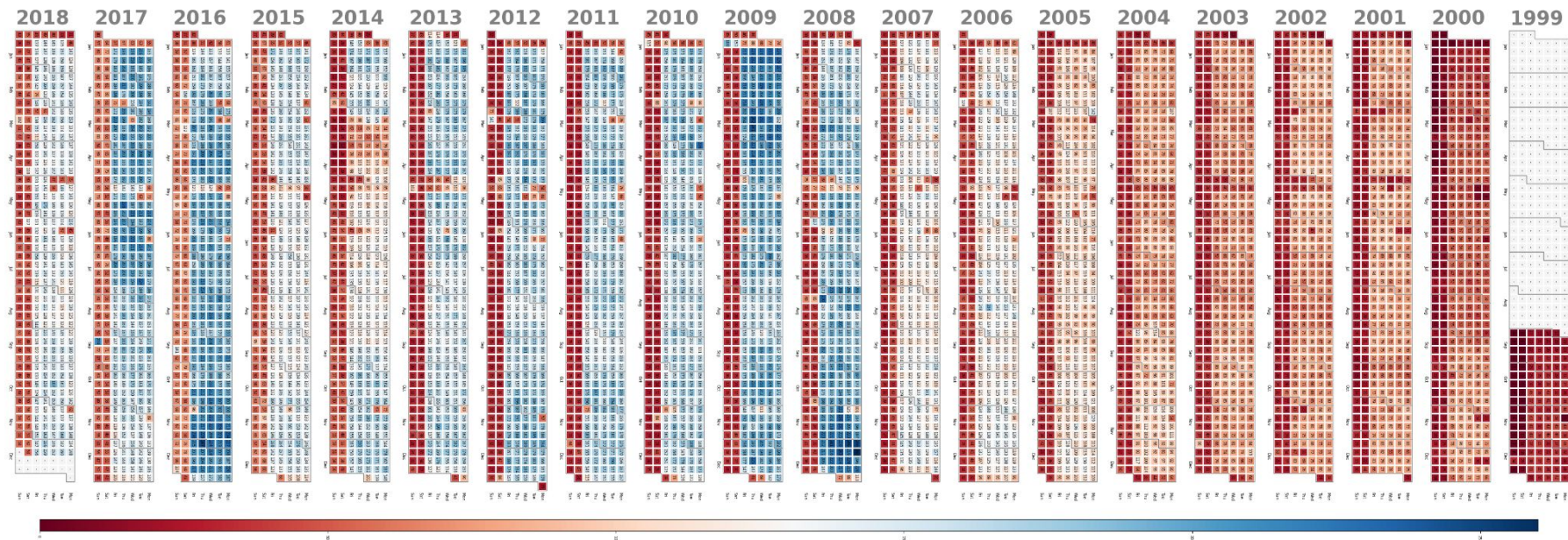
Общее количество тем



Распределение тем новостей по годам

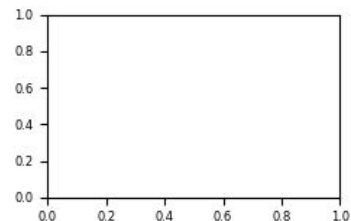
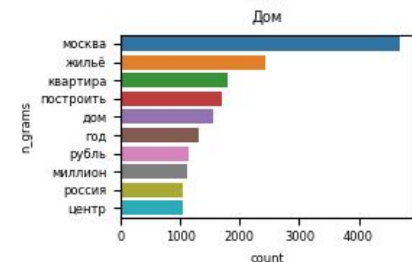
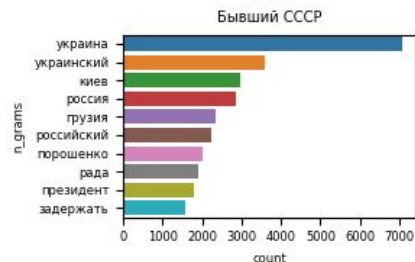
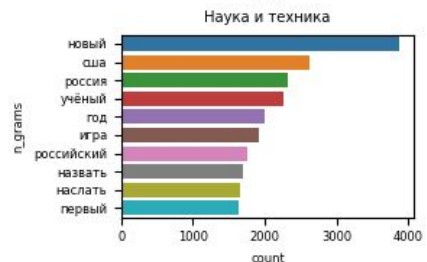
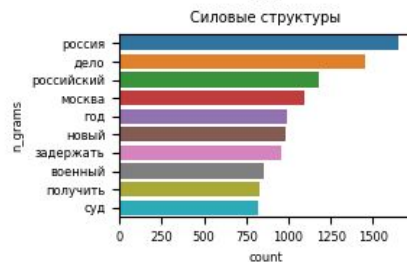
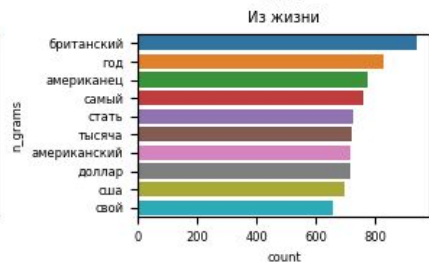
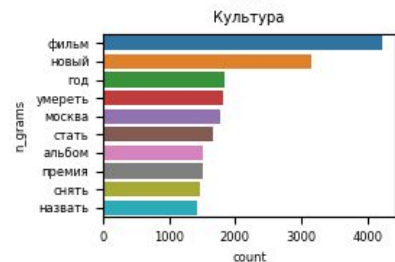
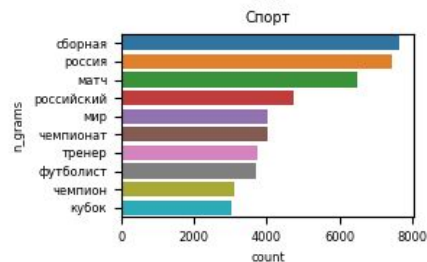
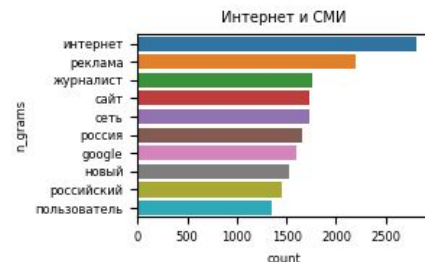
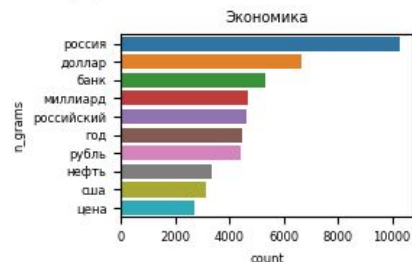
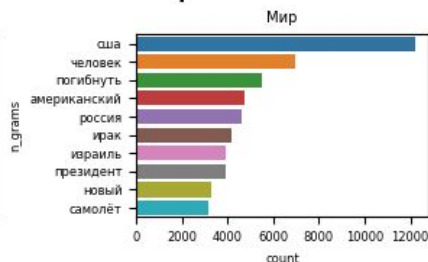
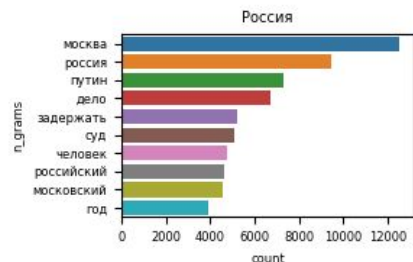


Распределение тем новостей по дням



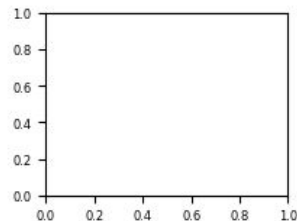
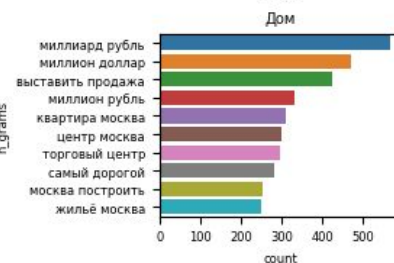
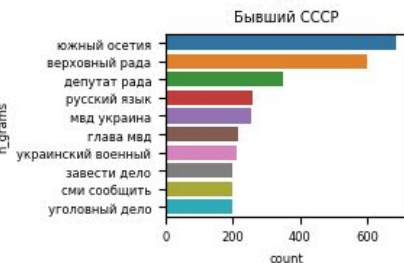
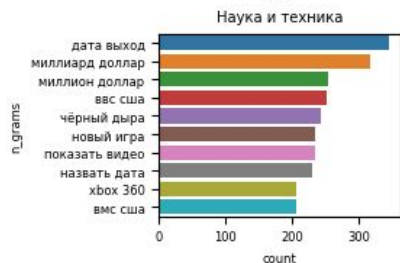
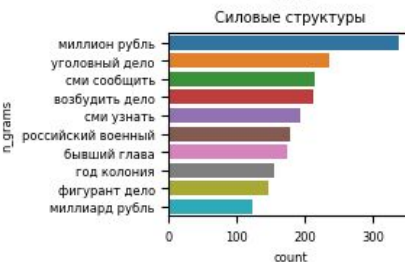
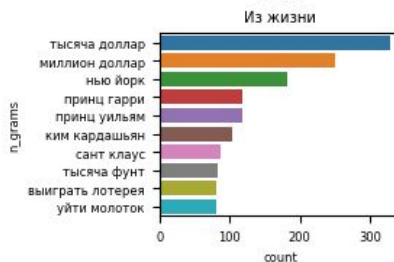
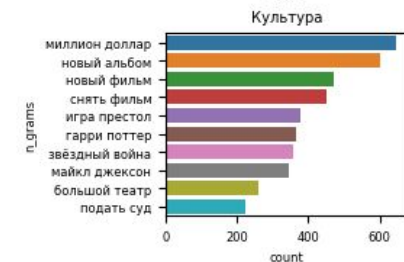
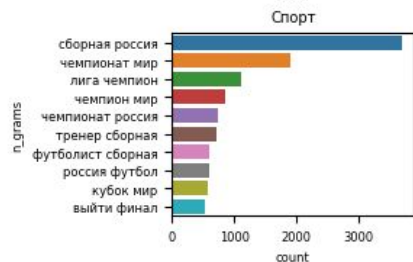
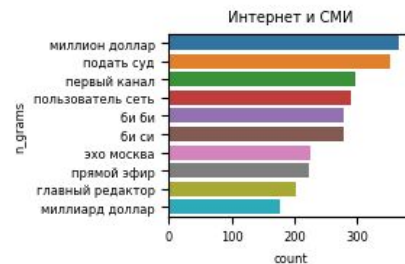
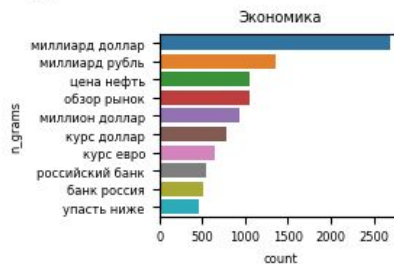
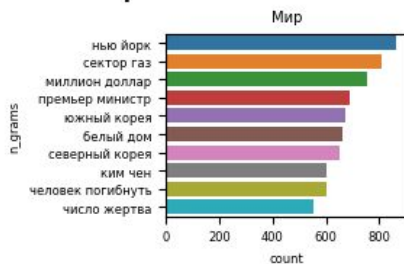
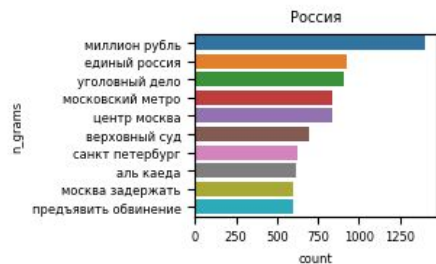
1-2-3 граммы для заголовка

Топ 10 юниграм по классам для заголовка статьи



1-2-3 граммы для заголовка

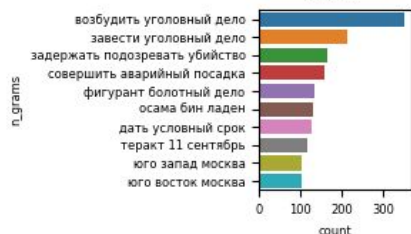
Топ 10 биграмм по классам для заголовка статьи



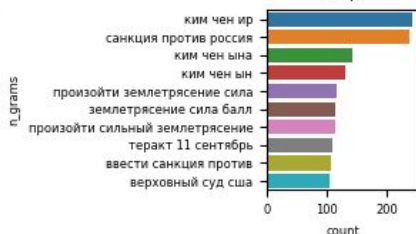
1-2-3 граммы для заголовка

Топ 10 триграмм по классам для заголовка статьи

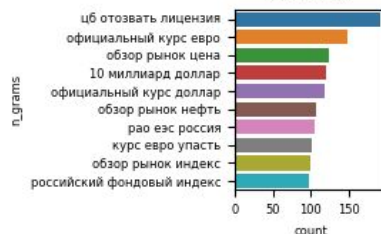
Россия



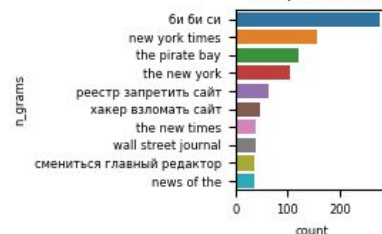
Мир



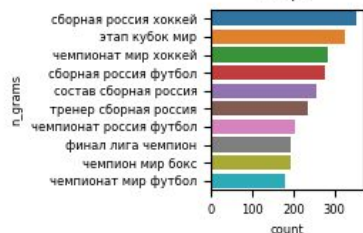
Экономика



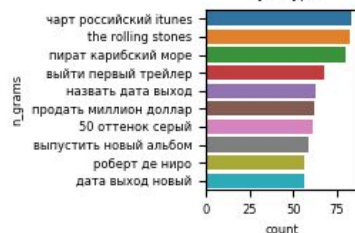
Интернет и СМИ



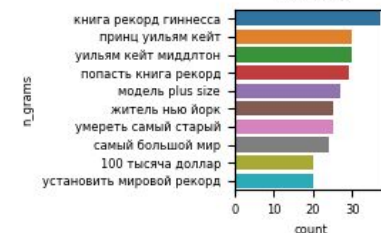
Спорт



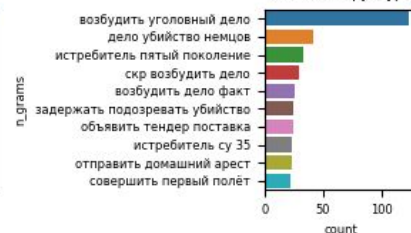
Культура



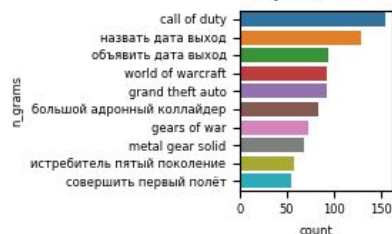
Из жизни



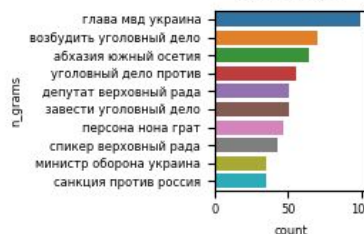
Силовые структуры



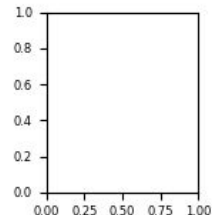
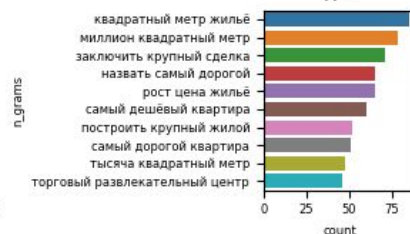
Наука и техника



Бывший СССР

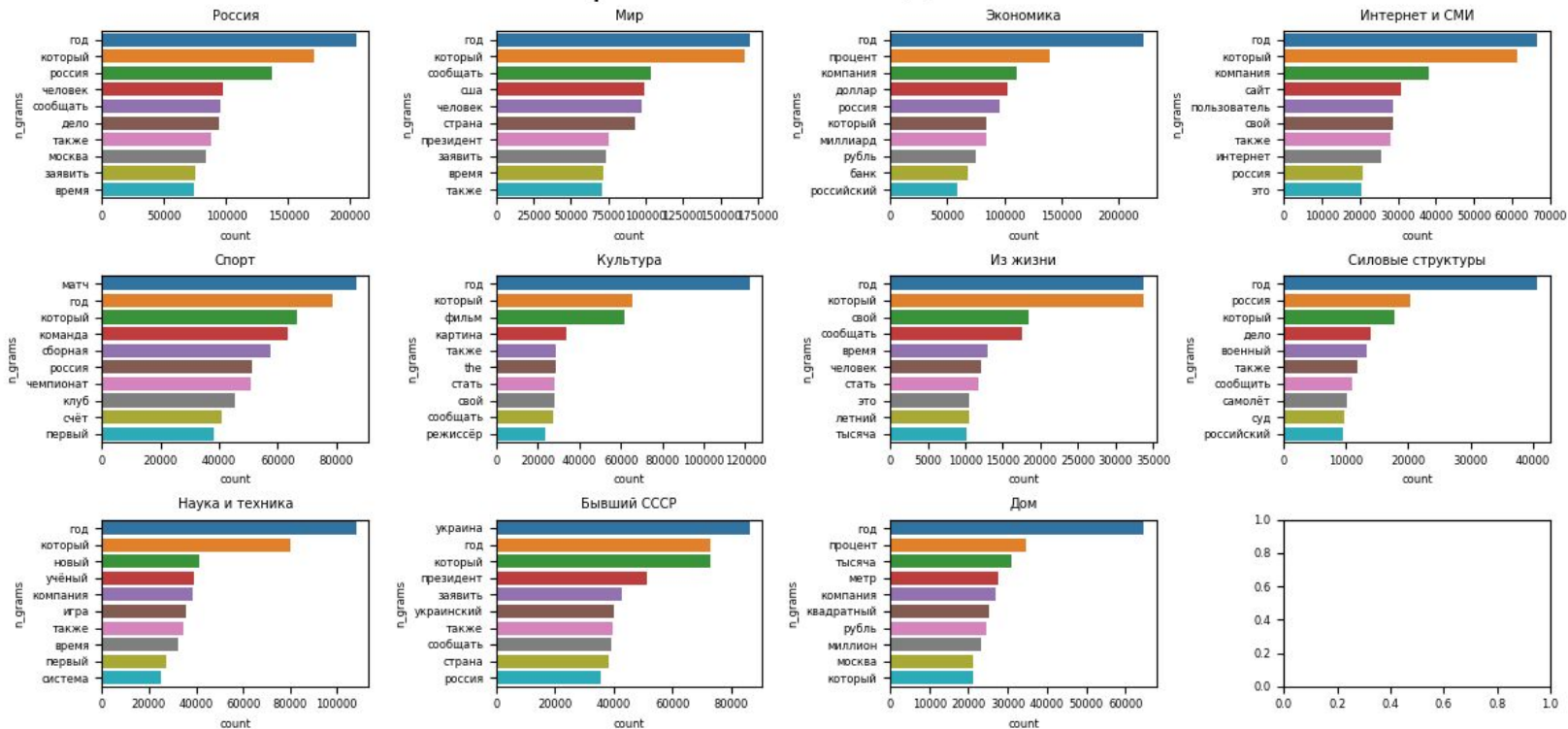


Дом



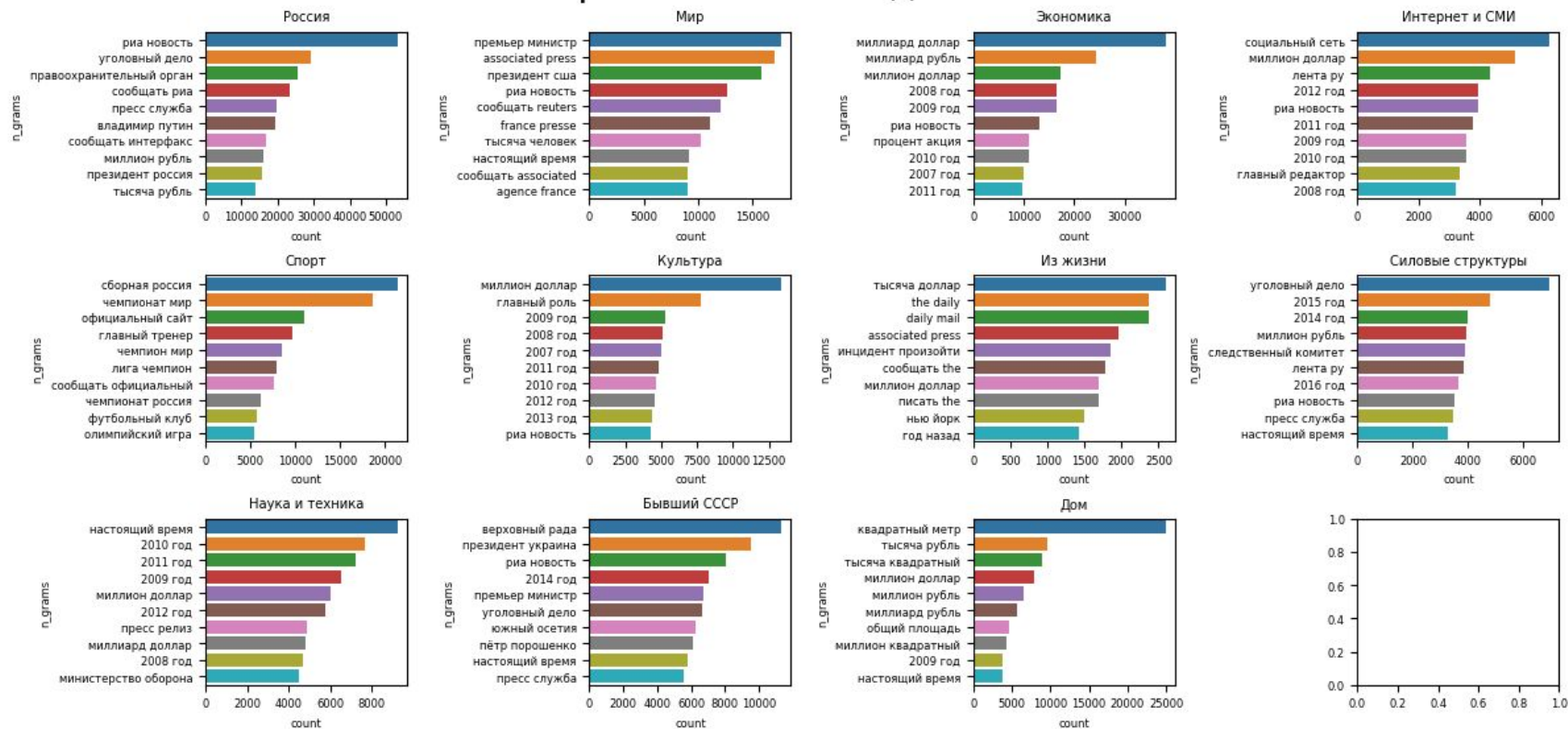
1-2-3 граммы для тела

Топ 10 юниграм по классам для тела статьи



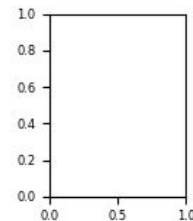
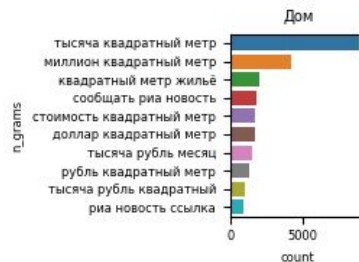
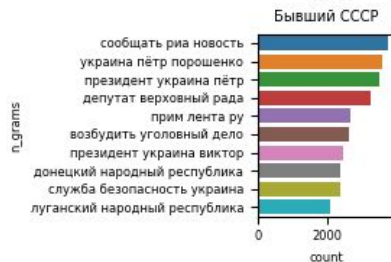
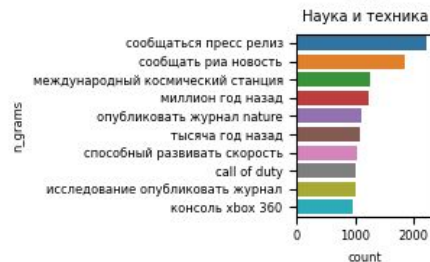
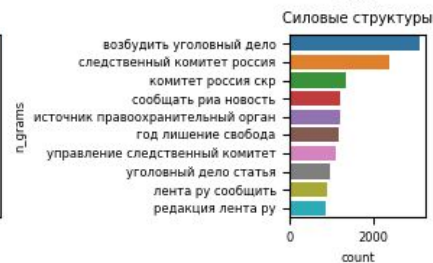
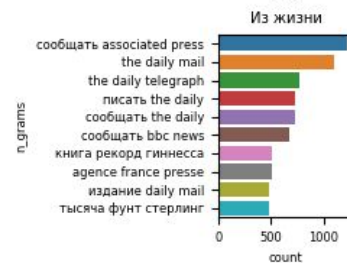
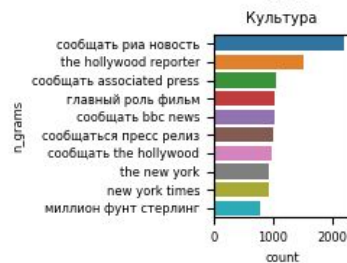
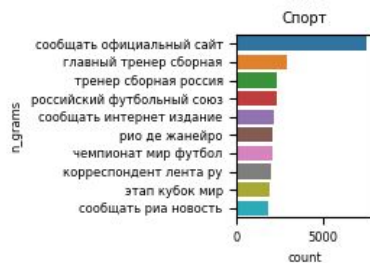
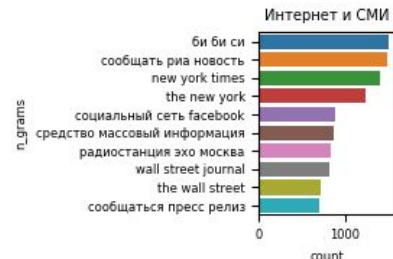
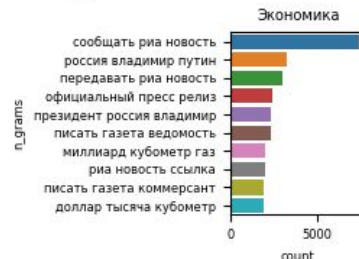
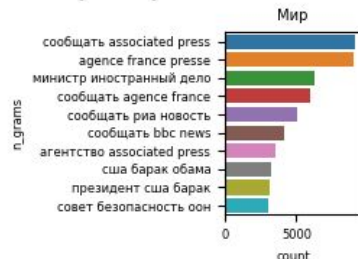
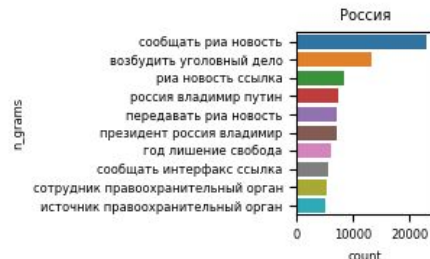
1-2-3 граммы для тела

Топ 10 биграмм по классам для тела статьи



1-2-3 граммы для тела

Топ 10 триграмм по классам для тела статьи



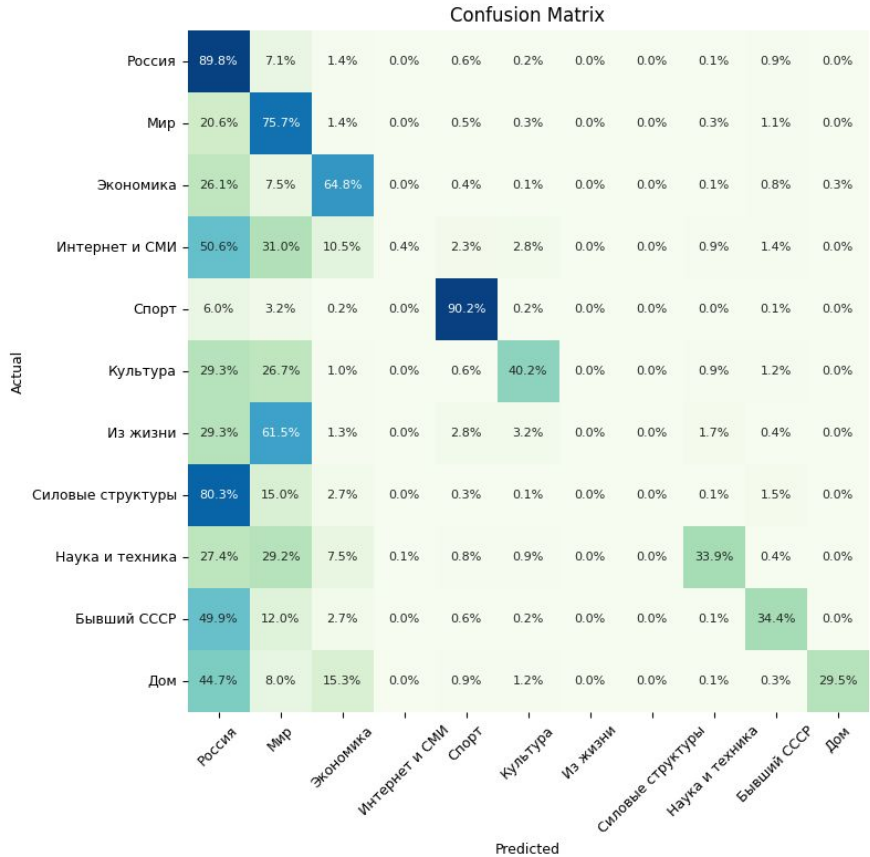
Извлечение признаков

1. Удаляем спец символы (оставляем только цифры и буквы)
2. Приводим к нижнему регистру
3. Лемматизируем слова
4. Удаляем стоп слова
5. 2 энкодера tf-idf: для заголовка и тела статьи
6. Ограничение по числу признаков 1100

Логистическая регрессия vs случайный лес

Случайный лес

	precision	recall	f1-score	support
Россия	0.46	0.90	0.61	32088
Мир	0.53	0.76	0.62	27324
Экономика	0.73	0.65	0.69	15906
Интернет и СМИ	0.75	0.00	0.01	8933
Спорт	0.92	0.90	0.91	12883
Культура	0.85	0.40	0.55	10759
Из жизни	0.00	0.00	0.00	5521
Силовые структуры	0.00	0.00	0.00	3919
Наука и техника	0.90	0.34	0.49	10627
Бывший СССР	0.77	0.34	0.48	10680
Дом	0.96	0.29	0.45	4347
accuracy			0.59	142987
macro avg	0.62	0.42	0.44	142987
weighted avg	0.63	0.59	0.54	142987



Логистическая регрессия

	precision	recall	f1-score	support
Россия	0.76	0.81	0.78	32088
Мир	0.77	0.81	0.79	27324
Экономика	0.83	0.85	0.84	15906
Интернет и СМИ	0.72	0.66	0.69	8933
Спорт	0.96	0.95	0.96	12883
Культура	0.84	0.82	0.83	10759
Из жизни	0.62	0.53	0.57	5521
Силовые структуры	0.67	0.47	0.55	3919
Наука и техника	0.81	0.81	0.81	10627
Бывший СССР	0.78	0.74	0.76	10680
Дом	0.84	0.77	0.80	4347
accuracy			0.79	142987
macro avg	0.78	0.75	0.76	142987
weighted avg	0.79	0.79	0.79	142987

Confusion Matrix

	Россия	Мир	Экономика	Интернет и СМИ	Спорт	Культура	Из жизни	Силовые структуры	Наука и техника	Бывший СССР	Дом
Россия	80.7%	7.8%	2.5%	1.5%	0.4%	0.9%	0.6%	1.3%	1.0%	2.6%	0.7%
Мир	8.0%	81.5%	1.4%	1.1%	0.3%	1.0%	2.2%	0.5%	1.4%	2.5%	0.1%
Экономика	5.9%	2.9%	84.7%	1.6%	0.3%	0.4%	0.3%	0.2%	1.0%	1.1%	1.5%
Интернет и СМИ	8.5%	5.8%	5.2%	65.9%	1.1%	3.7%	3.3%	0.2%	4.5%	1.6%	0.2%
Спорт	1.3%	0.8%	0.2%	0.6%	95.5%	0.3%	0.7%	0.1%	0.2%	0.3%	0.1%
Культура	4.9%	3.8%	0.5%	2.6%	0.4%	81.9%	3.0%	0.1%	1.4%	0.8%	0.5%
Из жизни	7.3%	19.1%	1.2%	5.4%	1.7%	7.3%	53.3%	0.1%	2.7%	1.5%	0.4%
Силовые структуры	31.7%	6.2%	1.1%	0.7%	0.2%	0.3%	0.2%	46.7%	10.7%	2.0%	0.2%
Наука и техника	3.2%	4.8%	2.5%	3.5%	0.2%	1.2%	1.1%	1.7%	81.1%	0.5%	0.1%
Бывший СССР	11.1%	7.3%	2.9%	1.2%	0.3%	0.9%	0.8%	0.7%	0.5%	74.1%	0.2%
Дом	8.7%	1.2%	7.8%	0.5%	0.4%	1.9%	1.5%	0.3%	0.3%	0.4%	76.9%
	Россия	Мир	Экономика	Интернет и СМИ	Спорт	Культура	Из жизни	Силовые структуры	Наука и техника	Бывший СССР	Дом

Predicted

Случайный лес - удалили 2 смешанных класса

	precision	recall	f1-score	support
Экономика	0.56	0.94	0.70	15906
Интернет и СМИ	0.75	0.28	0.41	8933
Спорт	0.91	0.97	0.94	12883
Культура	0.72	0.81	0.76	10759
Из жизни	0.00	0.00	0.00	5521
Силовые структуры	0.50	0.00	0.00	3919
Наука и техника	0.56	0.74	0.63	10627
Бывший СССР	0.77	0.80	0.78	10680
Дом	0.90	0.53	0.67	4347
accuracy			0.69	83575
macro avg	0.63	0.56	0.54	83575
weighted avg	0.66	0.69	0.64	83575

Confusion Matrix

	Экономика	Интернет и СМИ	Спорт	Культура	Из жизни	Силовые структуры	Наука и техника	Бывший СССР	Дом
Экономика	94.2%	0.3%	0.5%	0.3%	0.0%	0.0%	1.6%	2.4%	0.9%
Интернет и СМИ	35.7%	27.8%	3.7%	11.0%	0.0%	0.0%	14.6%	7.1%	0.1%
Спорт	0.9%	0.1%	97.4%	0.4%	0.0%	0.0%	0.7%	0.4%	0.0%
Культура	8.2%	0.8%	1.1%	80.7%	0.0%	0.0%	5.5%	3.4%	0.2%
Из жизни	17.2%	4.6%	6.8%	28.5%	0.0%	0.0%	36.9%	5.7%	0.3%
Силовые структуры	50.4%	0.7%	2.3%	4.4%	0.0%	0.0%	27.8%	13.6%	0.7%
Наука и техника	18.2%	3.2%	0.9%	2.2%	0.0%	0.0%	73.7%	1.7%	0.2%
Бывший СССР	11.3%	0.5%	1.0%	1.4%	0.0%	0.0%	5.8%	79.8%	0.2%
Дом	33.4%	0.5%	1.3%	3.7%	0.0%	0.0%	5.5%	2.2%	53.4%

Actual

Predicted

Логистическая регрессия - удалили 2 смешанных класса

	precision	recall	f1-score	support
Экономика	0.88	0.92	0.90	15906
Интернет и СМИ	0.80	0.77	0.78	8933
Спорт	0.97	0.97	0.97	12883
Культура	0.89	0.90	0.89	10759
Из жизни	0.75	0.73	0.74	5521
Силовые структуры	0.81	0.75	0.78	3919
Наука и техника	0.86	0.86	0.86	10627
Бывший СССР	0.88	0.89	0.88	10680
Дом	0.88	0.83	0.86	4347
accuracy			0.87	83575
macro avg	0.86	0.85	0.85	83575
weighted avg	0.87	0.87	0.87	83575

Confusion Matrix

	Экономика	Интернет и СМИ	Спорт	Культура	Из жизни	Силовые структуры	Наука и техника	Бывший СССР	Дом
Экономика	91.9%	1.9%	0.3%	0.3%	0.6%	0.5%	1.1%	1.8%	1.6%
Интернет и СМИ	5.7%	77.0%	1.2%	3.5%	3.6%	0.9%	4.8%	3.1%	0.3%
Спорт	0.4%	0.7%	96.8%	0.3%	1.0%	0.2%	0.2%	0.4%	0.1%
Культура	0.7%	2.5%	0.3%	89.5%	3.3%	0.4%	1.5%	1.2%	0.6%
Из жизни	2.2%	6.4%	1.9%	7.6%	72.9%	0.9%	3.5%	3.9%	0.7%
Силовые структуры	3.1%	2.0%	0.1%	0.9%	0.9%	75.2%	12.2%	4.8%	0.7%
Наука и техника	2.6%	3.6%	0.2%	1.7%	2.0%	2.3%	86.5%	1.0%	0.2%
Бывший СССР	3.7%	1.9%	0.3%	1.0%	1.7%	1.4%	0.7%	89.0%	0.3%
Дом	9.5%	0.9%	0.6%	2.1%	1.3%	1.0%	0.5%	1.0%	83.1%

Actual

Predicted

Поиск похожей новости

Что мы имеем в виду

Результатом алгоритма по поиску похожей новости должна являться новость из нашей базы данных, векторное представление (эмбединг) которой находится ближе всех к эмбедингу новости, введенной рользователем.

Описание алгоритма поиска похожей новости

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Проблемы

Долгая обработка признаков

Частые вылеты из-за ООМ

Долгое обучение моделей

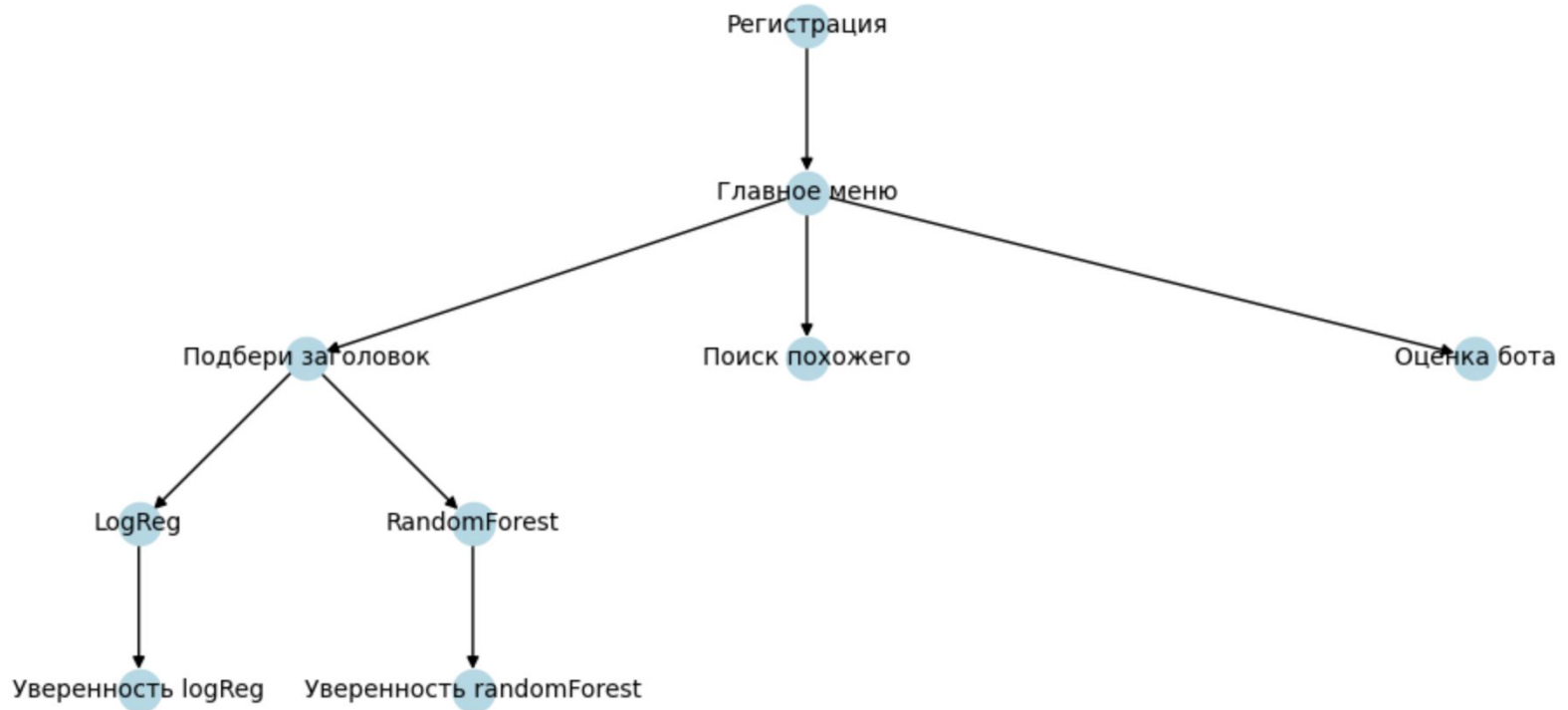
Асинхронный бот в Telegram

Функционал бота

Для удобного доступа пользователей к результатам нашей работы мы написали асинхронного бота со следующим функционалом:

- Регистрация
- Панель главного меню
- Оценка бота: сохранение в файл и проверка на то, что пользователь уже ставил оценки
- Поиск похожей новости
- Подбор названия новости через логистическую регрессию
- Показ уверенности бота в своем ответе для логистической регрессии
- Механизм отлова неправильного ввода и вывода подсказок
- Механизм кнопки назад: иерархическая структура интерфейса бота

Иерархическая структура интерфейса бота



Интеграция с моделями

Бесшовная Интеграция: Асинхронный бот в Telegram успешно интегрирован с разработанной моделью логистической регрессии. Это позволяет боту анализировать новостные данные в реальном времени и предоставлять актуальную информацию пользователям.

Доступ к нашей базе данных: Пользователи могут запрашивать поиск похожих новостей через бота. Бот обрабатывает запросы, и возвращает самую похожую новость из нашей базы данных

Попробуйте сами!

[@news_hse_project_bot](#)



Заключение

Итоги проекта

Нам удалось добиться хорошей точности моделей машинного обучения:

- Достигнута высокая точность в классификации и анализе новостей благодаря моделям логистической регрессии и случайного леса.

Нам удалось успешно внедрить асинхронного бота в Telegram:

- Создан и успешно интегрирован асинхронный бот в Telegram, обеспечивающий легкий и удобный доступ пользователей к полученным в данной работе моделям.

Планы на будущее

Используя более продвинутые методы МО и глубокого машинного обучения сделать классификатор “кликбейтных” новостей.

Превратить бота в новостной агрегатор, фильтрующий или переименовывающий “кликбейтные” новости.