

Команда 37.

Прогнозирование тем и рейтинга статей с `habr.ru`

Состав команды:

- Караулов Виталий (EDA, ML-Baseline, Service)
- Мясникова Ангелина (EDA, ML-Baseline, Service)
- Писцов Георгий (EDA, ML-Baseline, Service)
- Сопельник Дмитрий (Парсинг, EDA, ML-Baseline, Service)

Куратор проекта:

Чуприн Александр

Цель проекта

- Разработка сервиса с внедрением обученных ML/DL-моделей для предсказания тематики и рейтинга статей на основе данных, собранных с habr.ru

Задачи:

- Парсинг статей с habr.ru
- Предобработка данных и EDA
- Построение бейзлайна ML-модели
- Создание MVP-сервиса на основе Streamlit и FastAPI ([мы находимся здесь](#))
- Улучшение бейзлайна
- Доработка ML-решения и внедрение DL
- Доработка сервиса

Собранный датасет, его обработка

	Title	Author	Publication_date	Hubs	Tags	Content	Comments	Views	URL	Reading_time	Images_links	Individ/Company	Rating	Positive/Negative	Bookmarks_cnt
0	Лечение приступов лени	complex	2009-08-03 14:34:35+00:00	GTD	лень, учись работать, самомотивация, мотивация	Пора лишать девственности свой блогик.\nТак как это происходит сегодня, в по...	67	6800	https://habr.com/ru/articles/66091/	2.0		individual	4	positive	25.0
1	Как я работал по два часа в день	Konovalov	2009-07-30 13:50:03+00:00	GTD	тайм-менеджмент, timemanagement, работа, эффективность, организация дел	Когда я только перешёл от офисной работы к домашней, первое время был на сед...	99	21000	https://habr.com/ru/articles/65783/	3.0		individual	193	positive	114.0

Размер исходного датасета

```
1 df.shape
```

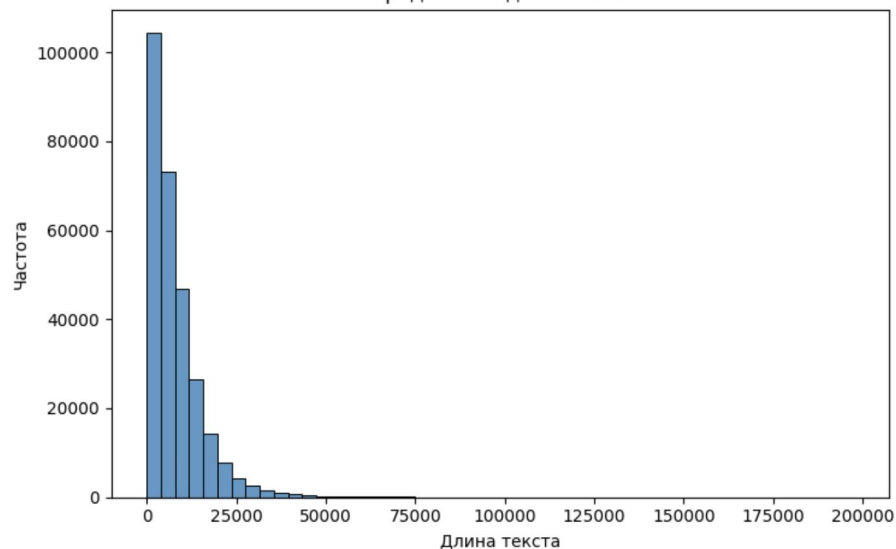
(285499, 15)

Текстовые колонки (Title, Author, Hubs, Tags, Content) были преобразованы так:

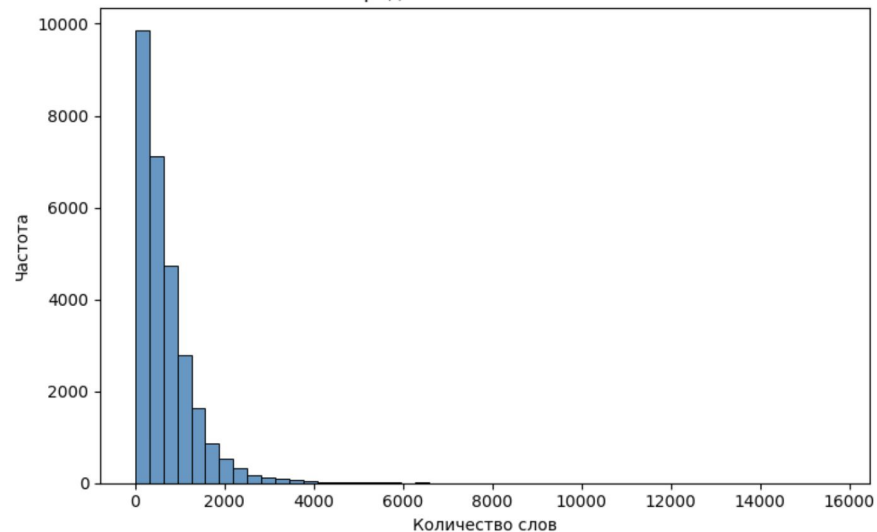
- Удалены лишние символы: html-тэги, специальные символы, цифры, лишние пробелы
- Удалены стоп-слова
- Слова приведены к нижнему регистру
- Проведена лемматизация и токенизация

EDA. Тексты статей

Распределение длины текстов



Распределение количества слов



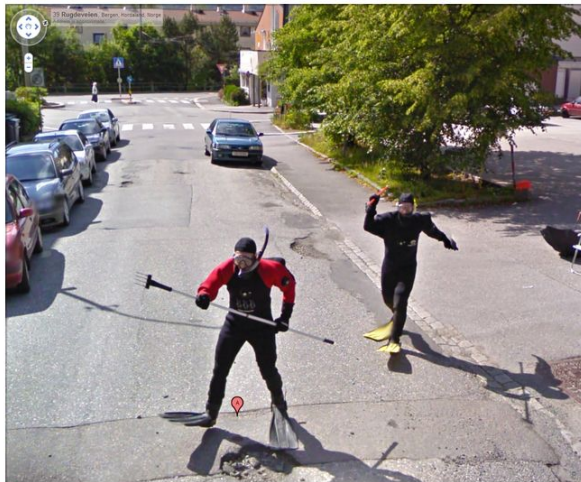
EDA. Тексты статей

 Kuper 11 фев 2010 в 00:23

Агрессивные норвежцы на Google Street View

🕒 1 мин 👁 545

Геоинформационные сервисы*




Я не верю!!

Теги: Google Street View, юмор

👍 +57 🗨 10 ➡ 24

Пример статьи с околонулевой длиной текста

 fur_habr 20 мая 2019 в 12:14

Помощь и просьба о ней. Статья про информационную безопасность для рядовых пользователей

🕒 118 мин 👁 103K

Информационная безопасность*

Из песочницы

Я предлагаю вам некоторые шаги по повышению безопасности и приватности в интернет сети (и не только) для рядовых пользователей. Обоснование почему это необходимо — в начале статьи. Для тех, кто всё знает и недоумевает, почему этот текст находится здесь — просьба прочитать пункт «Для тех, кто уже всё знает». Три месяца назад я написала этот текст, но в связи с моей необразованностью и нескончаемым потоком новостей о новых угрозах безопасности, мне надоело переделывать, так что пусть в этом тексте остаётся всё как было).

Вступление и обоснование

Общие рекомендации

Пароли и учётные записи

Голосовые помощники

СИМ карты

--Дополнительная СИМ карта

Проверить все устройства антивирусами

Сохранить все данные на внешние носители/сделать бэкап

Проверить обновления устройств

Удалить всё лишнее из всех соц. сетей и сайтов

--Для Google

--Для Яндекс

--Для Bing и MAIL.RU

--Для Вконтакте

--Для Одноклассники

Удалите все мессенджеры от глобальных компаний

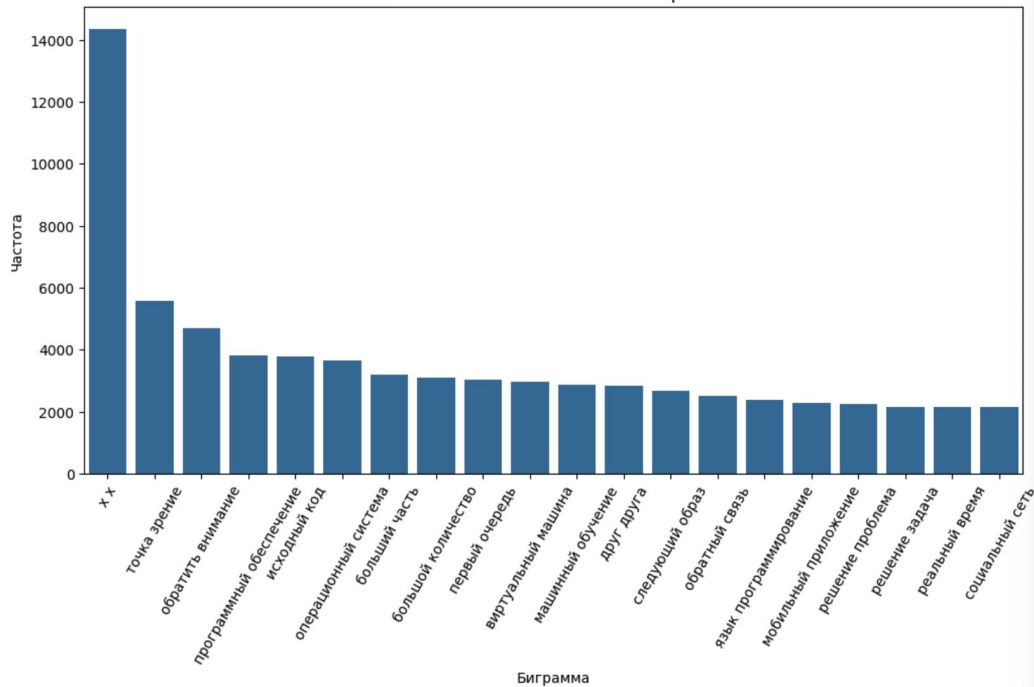
Не используйте системы оплаты типа vk.pay, Яндекс.Деньги и Master Card

Удалить всё с облаков

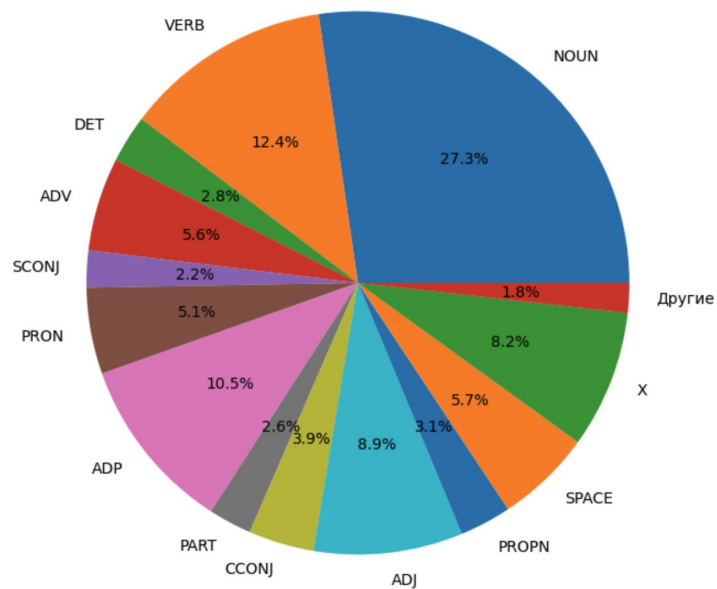
Самая длинная статья в датасете
(197359 символов)

EDA. Тексты статей

Топ-20 наиболее частотных биграмм



Распределение частей речи



EDA. Облако слов текстов статей

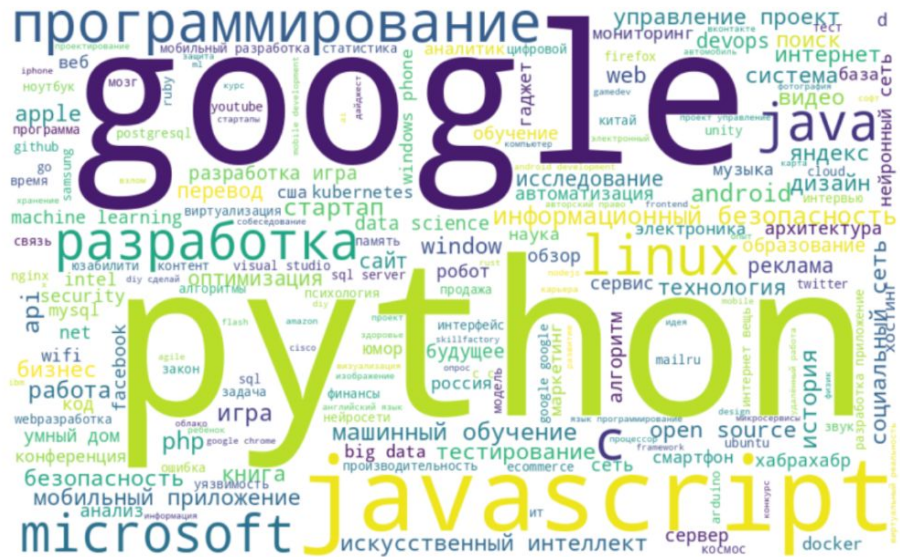


EDA. Облака слов заголовков и тегов

Облако слов для заголовков



Облако слов для тегов



EDA. Итоговый вид датасета и распределение классов

	author	publication_date	hubs	comments	views	url	reading_time	individ/company	bookmarks_cnt	text_length	tags_tokens	title_tokens	rating_new	text_tokens	text_pos_tags
0	complex	2009-08-03 14:34:35+00:00	GTD	67	6800	https://habr.com/ru/articles/66091/	2.0	individual	25.0	2027	['лень' 'учись' 'работать' 'самомотивация' 'мотивация']	['лечение' 'приступ' 'лень']	4.0	['лишать' 'девственность' 'блужик' 'происходить' 'сегодня' 'понедельник' 'д...	[NOUN, VERB, NOUN, DET, NOUN, ADV, SCONJ, PRON, VERB, ADV, ADP, NOUN, ADP, N...
1	popotam2	2009-07-15 20:24:31+00:00	GTD	13	3100	https://habr.com/ru/articles/64586/	1.0	individual	6.0	424	['развитие' 'работоспособность' 'организация' 'дело' 'дело']	['организация' 'рабочий' 'время' 'помощь' 'цвет']	1.0	['предлагать' 'вариант' 'сделать' 'организованный' 'успеть' 'сделать' 'кале...	[VERB, ADV, NUM, NOUN, VERB, PRON, ADV, ADJ, CCONJ, PRON, VERB, PRON, VERB, ...

Размер итогового датасета

```
df.shape
```

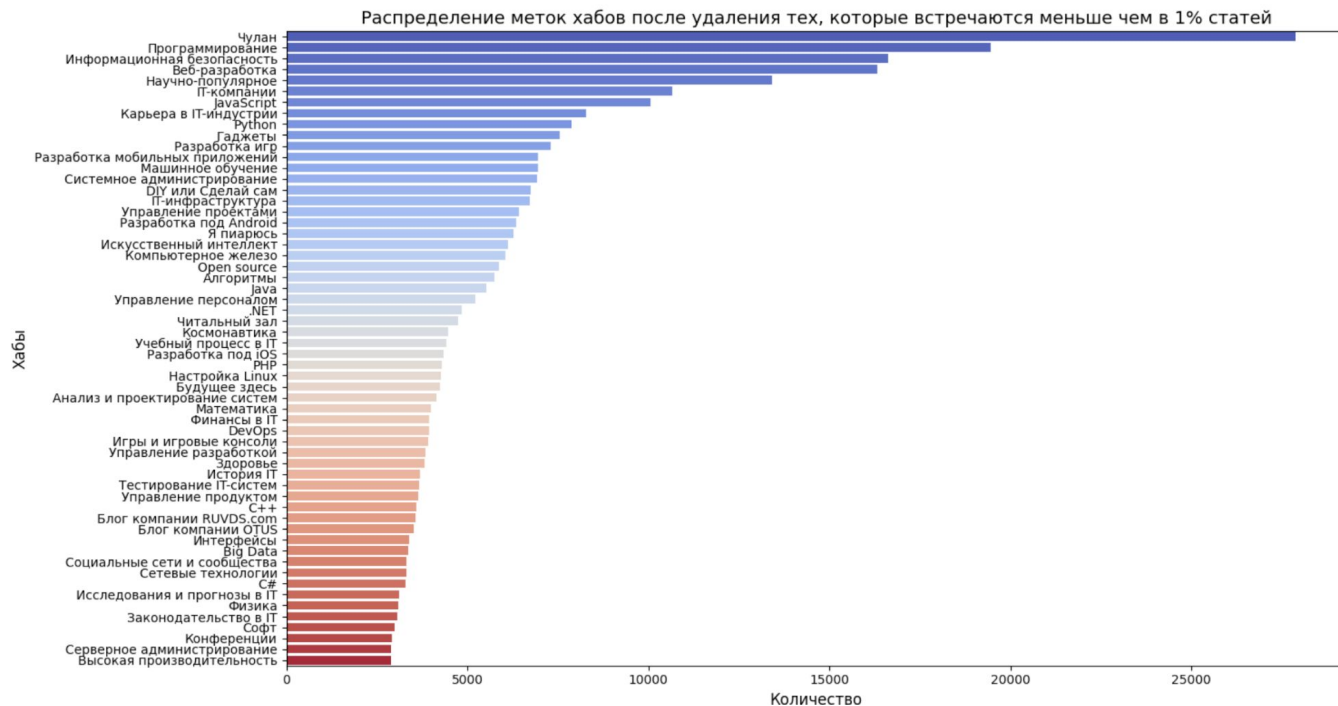
```
(282640, 15)
```

Распределение классов в хабах и рейтинге

hubs			
Чулан	27888		
Программирование	19469		
Информационная безопасность	16627	1	110927
Веб-разработка	16320		
Научно-популярное	13431	0	75106
...			
Блог компании «Медиа Грус»	1	2	57297
Блог компании Fixico	1		
Блог компании iRuPay	1	-1	7468
Блог компании Чистилка	1		
Блог компании Indexisto	1	-2	3578

Name: count, Length: 2255, dtype: int64

ML. Распределение классов в хабах (темах)



После удаления редких классов всего осталось 58 хабов (тем)

ML. Метрики лучших моделей

Хабы - TfidfVectorizer + OneVsRestClassifier(LinearSVC):

micro - {'Precision': 0.3828, 'Recall': 0.7205, 'F1-Score': 0.4999, 'Hamming Loss': 0.0167}

macro - {'Precision': 0.377, 'Recall': 0.7084, 'F1-Score': 0.4766, 'Hamming Loss': 0.0167}

weighted - {'Precision': 0.4405, 'Recall': 0.7205, 'F1-Score': 0.5361, 'Hamming Loss': 0.0167}

Рейтинг - TfidfVectorizer + Logistic Regression:

micro - {'Precision': 0.496, 'Recall': 0.496, 'F1-Score': 0.496, 'Hamming Loss': 0.504}

macro - {'Precision': 0.2889, 'Recall': 0.4111, 'F1-Score': 0.288, 'Hamming Loss': 0.504}

weighted - {'Precision': 0.5604, 'Recall': 0.496, 'F1-Score': 0.5173, 'Hamming Loss': 0.504}

MVP-сервис. Демонстрация работы

