ИРКУТСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ О ТУРИСТИЧЕСКИХ МЕСТАХ БАЙКАЛЬСКОГО РЕГИОНА ИЗ ПОСТОВ В СОЦИАЛЬНЫХ СЕТЯХ

Магистрант Поддубный И.А. - poddubnyyiv@yandex.ru
Пестова Ю. В.
Николайчук О.А.

Иркутск 2023

Актуальность

- В Стратегии развития туризма в Российской Федерации на период до 2035 года отмечено, что для роста конкурентоспособности и раскрытия потенциала туристского продукта необходимо:
- Обеспечить повышение доступности актуальных отраслевых данных со стороны участников туристского рынка

Наименование	Тип объекта	Регион	Населенный
Рыбалка на Байкале	Водный спорт	Иркутская о	Иркутск г
Байкал в цифрах и фактах	Природные памятни	Иркутская о	Иркутск г
Ангарский городской музей	Художественные муз	Иркутская о	Ангарск г
Усть-Кутский музей истории школы №9	Краеведческие музе	Иркутская о	Усть-Кут г
Музей истории Байкальского государственного университета экономики и права	Музеи, галереи, выст	Иркутская о	Иркутск г

Достопримечательности, Ростуризм

«Здоровье» УСКС ОАО «АНХК»	(зима) база отдыха (лето) детский оздоровительный лагерь	Ангарское гор.пос., 18 км Савватеевского тракта
«Юбилейная» УСКС ОАО «АНХК»	(зима) база отдыха (лето) детский оздоровительный лагерь	Ангарское гор.пос., 1 км развилки автодорог Ангарск-Одинск-Савватеевка
ООО «Ангарская горка»	база отдыха	Ангарское гор.пос., 12 км Савватеевского тракта

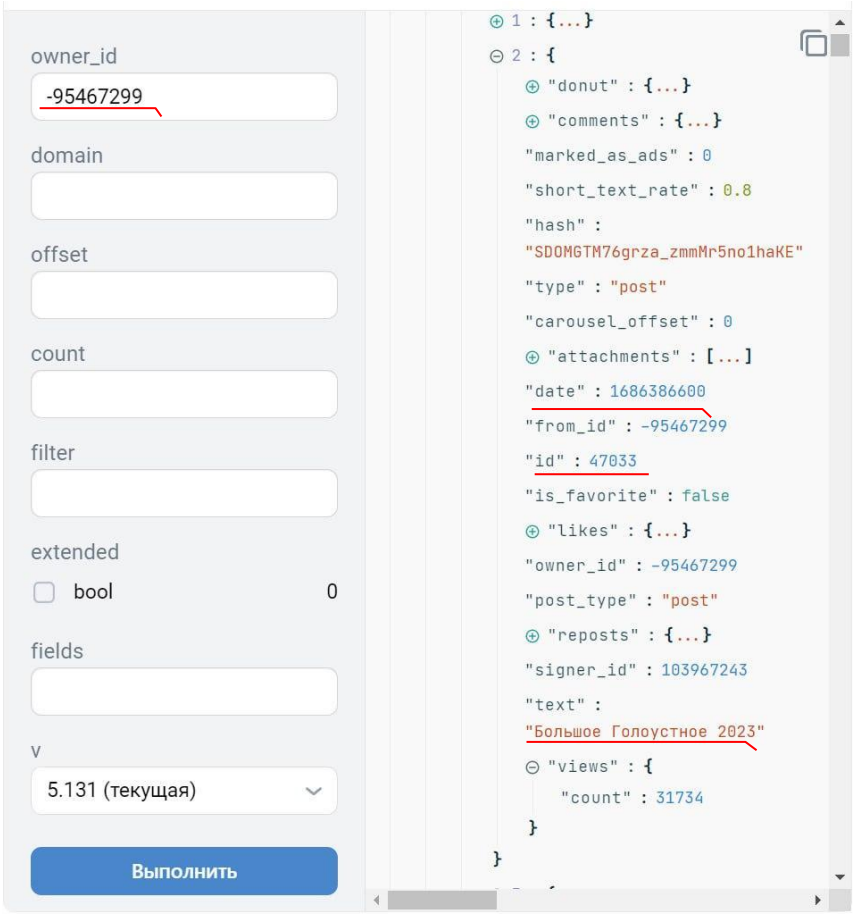
https://irkobl.ru/sites/tour/working/Общая%20информация.pdf

ЦЕЛЬ

Разработать метод и программную систему сбора информации из постов социальных сетей об объектах притяжения туристов (достопримечательностях) для последующего их мониторинга на территории Иркутской области, прилегающей к Байкалу

ИСТОЧНИКИ ДАННЫХ

- Источники данных – открытые тематические группы ВК:
  - Байкал
  - Мой Байкал | Экология
  - Байкал удивительный
  - Байкал для каждого
  - БАЙКАЛ
- ОК
- Все посты обезличены



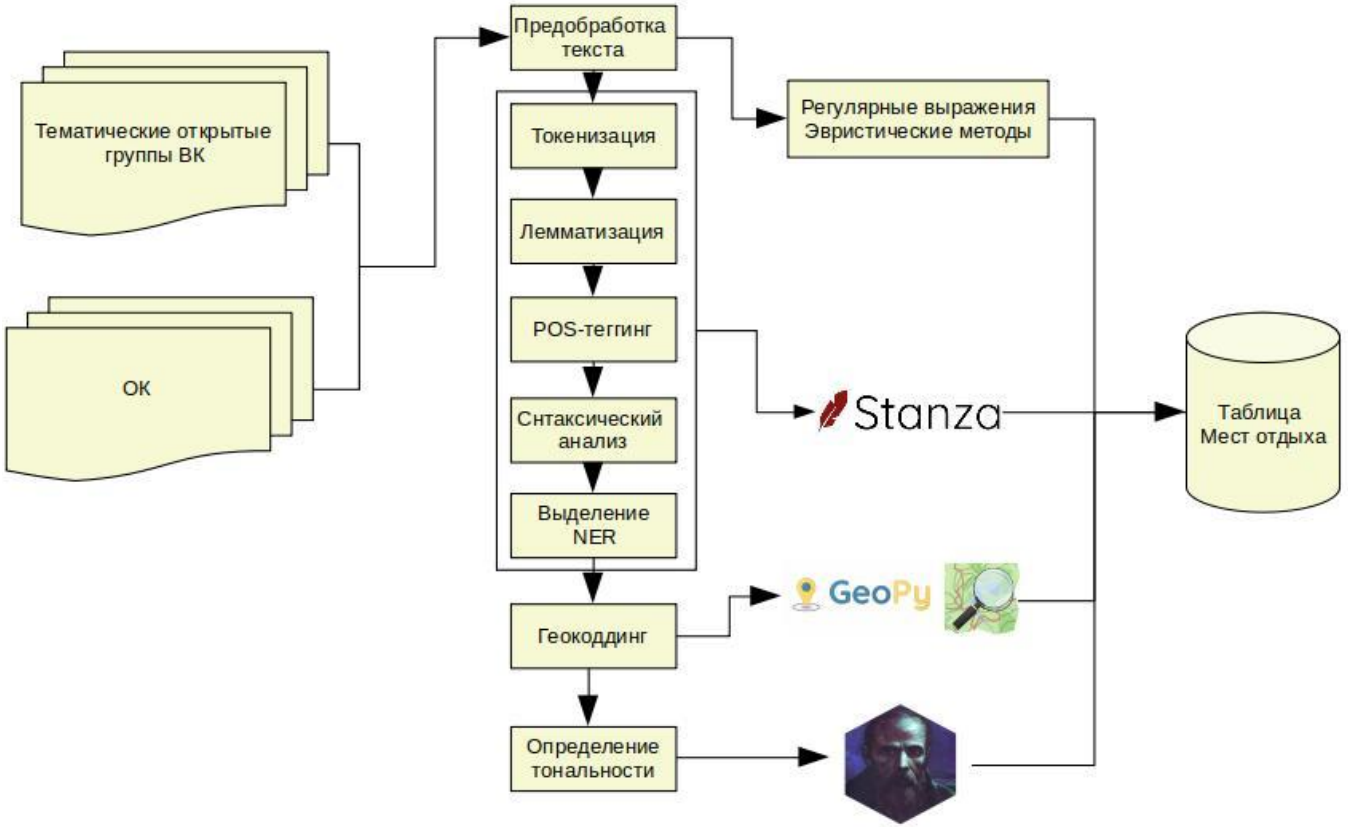
ОПИСАНИЕ ДАННЫХ

- ВК, ОК ~ 16000 постов
- В тексте встречаются:
  - Эмодзи и подобные им символы
  - Ссылки, адреса
  - Латиница, текст на английском
  - Лишние пробелы при использовании дефиса
- Местами отсутствует пунктуация

id	post_id	date	text
2061	2057	2022-03-31 00:43:07	«Экодесант» – это ...

Формат данных

ЭТАПЫ ОБРАБОТКИ ПОСТОВ



Stanza:

- Широкий набор решаемых задач
- **Поддержка русского языка**
- Работает из коробки

Dostoevsky:

- **Поддержка русского языка**
- Обучен на схожем с данными датасете
- Работает из корбки

geopy+Nominatim

- Бесплатен
- Есть возможность развернуть БД OSM на своем сервере

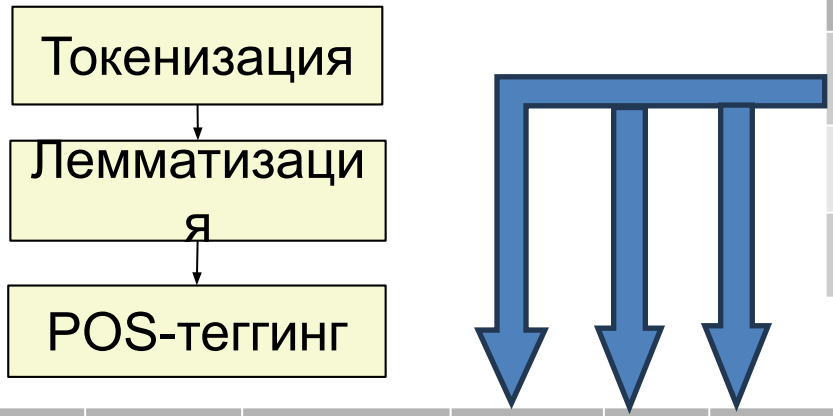
ЭТАП ПРЕДОБРАБОТКИ

- Удаляются эмодзи.
- Перевод строки заменяется на точки.
- Англоязычные фразы переводятся на русский язык.

id	post_id	date	text
11845	11831	2020-08-07 16:50:22	Недавно ездил на Байкал😁'\n' Сделал небольшой монтаж📹'\n' Оцените пожалуйста работу 🙏🙏🙏

id	post_id	date	text
11845	11831	2020-08-07 16:50:22	Недавно ездил на Байкал. Сделал небольшой монтаж. Оцените пожалуйста работу

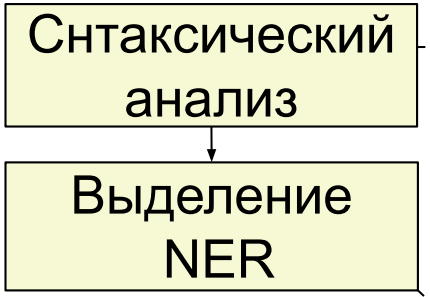
ЭТАП ОБРАБОТКИ ТЕКСТА



post_id	group_id	post_date	sentence	num_in_post
11845	12377	2020-08-07 16:50:22	Недавно ездил на Байкал.	0
11845	12377	2020-08-07 16:50:22	Сделал небольшой монтаж.	1
11845	12377	2020-08-07 16:50:22	Оцените пожалуйста работу .	2

post_id	group_id	post_date	num_in_post	num_in_sentence	normal_form	pos	word	ner
11845	12377	2020-08-07 16:50:22	0	1	недавно	ADV	Недавно	O
11845	12377	2020-08-07 16:50:22	0	2	ездить	VERB	ездил	O
11845	12377	2020-08-07 16:50:22	0	3	на	ADP	на	O
11845	12377	2020-08-07 16:50:22	0	4	Байкал	PROPN	Байкал	S-LOC
11845	12377	2020-08-07 16:50:22	0	5	.	PUNCT	.	O

ЭТАП ОБРАБОТКИ ТЕКСТА

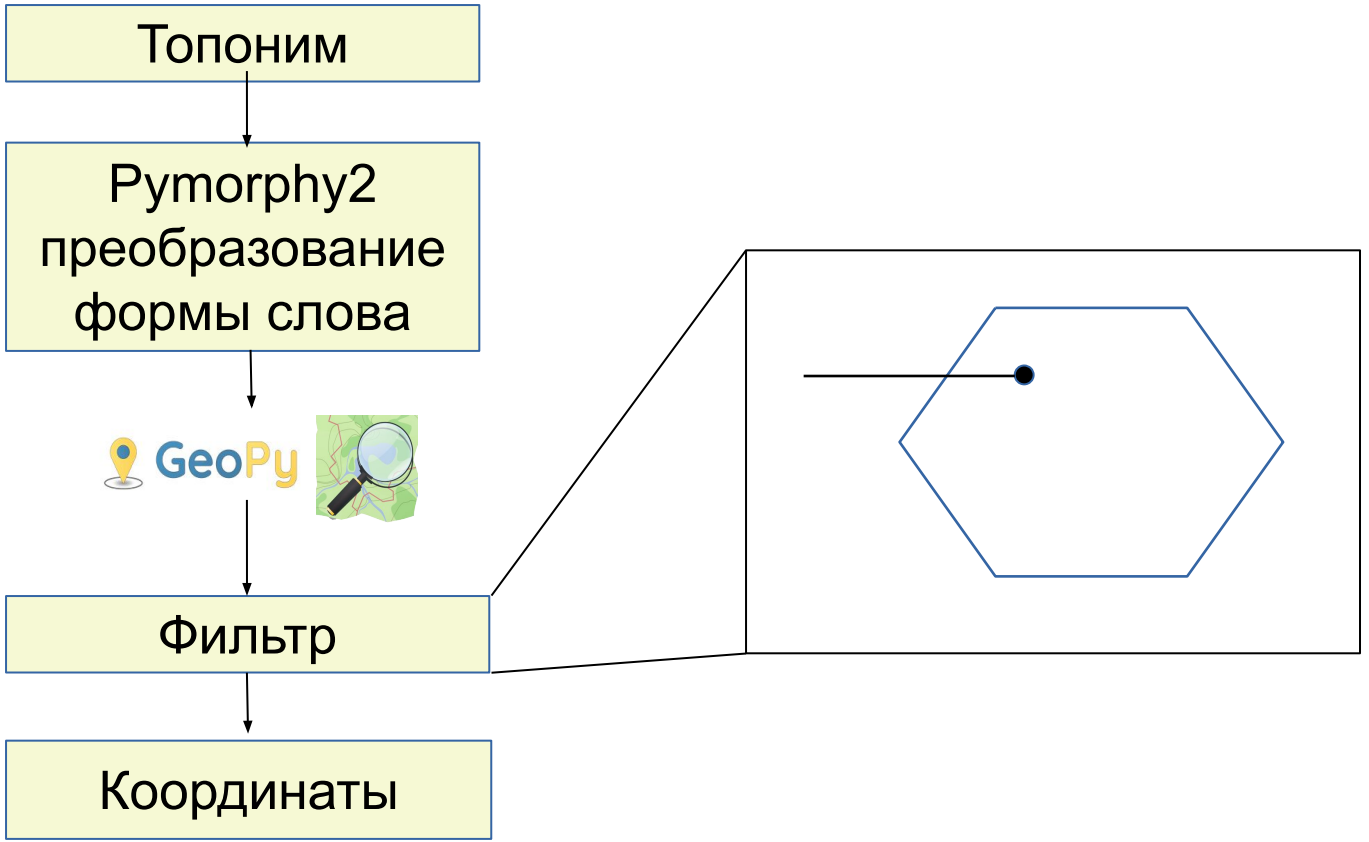


- Именованые сущности классифицируются;
- Среди них выбираются локации.

post_id	group_id	post_date	num_in_post	num_in_sentence	word	head_id	deprel
11845	12377	2020-08-07 16:50:22	0	1	Недавно	2	advmod
11845	12377	2020-08-07 16:50:22	0	2	ездил	0	root
11845	12377	2020-08-07 16:50:22	0	3	на	4	case
11845	12377	2020-08-07 16:50:22	0	4	Байкал	2	obl

post_id	group_id	post_date	num_in_post	num_in_sentence	word	loc_norm_form	norm_form
11845	12377	2020-08-07 16:50:22	0	4	Байкал	байкал	Байкал

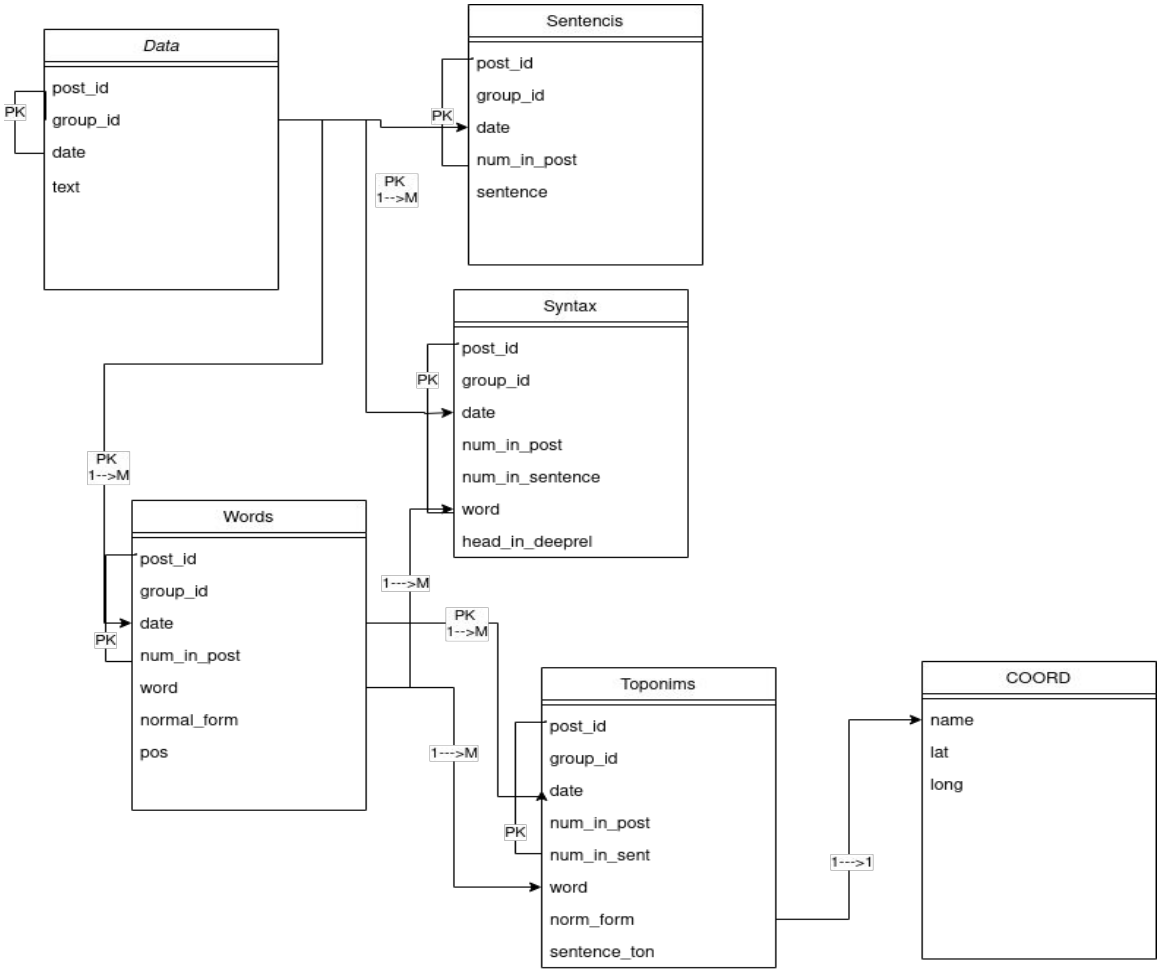
ЭТАП ГЕОКОДИНГА



ЭТАП ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ

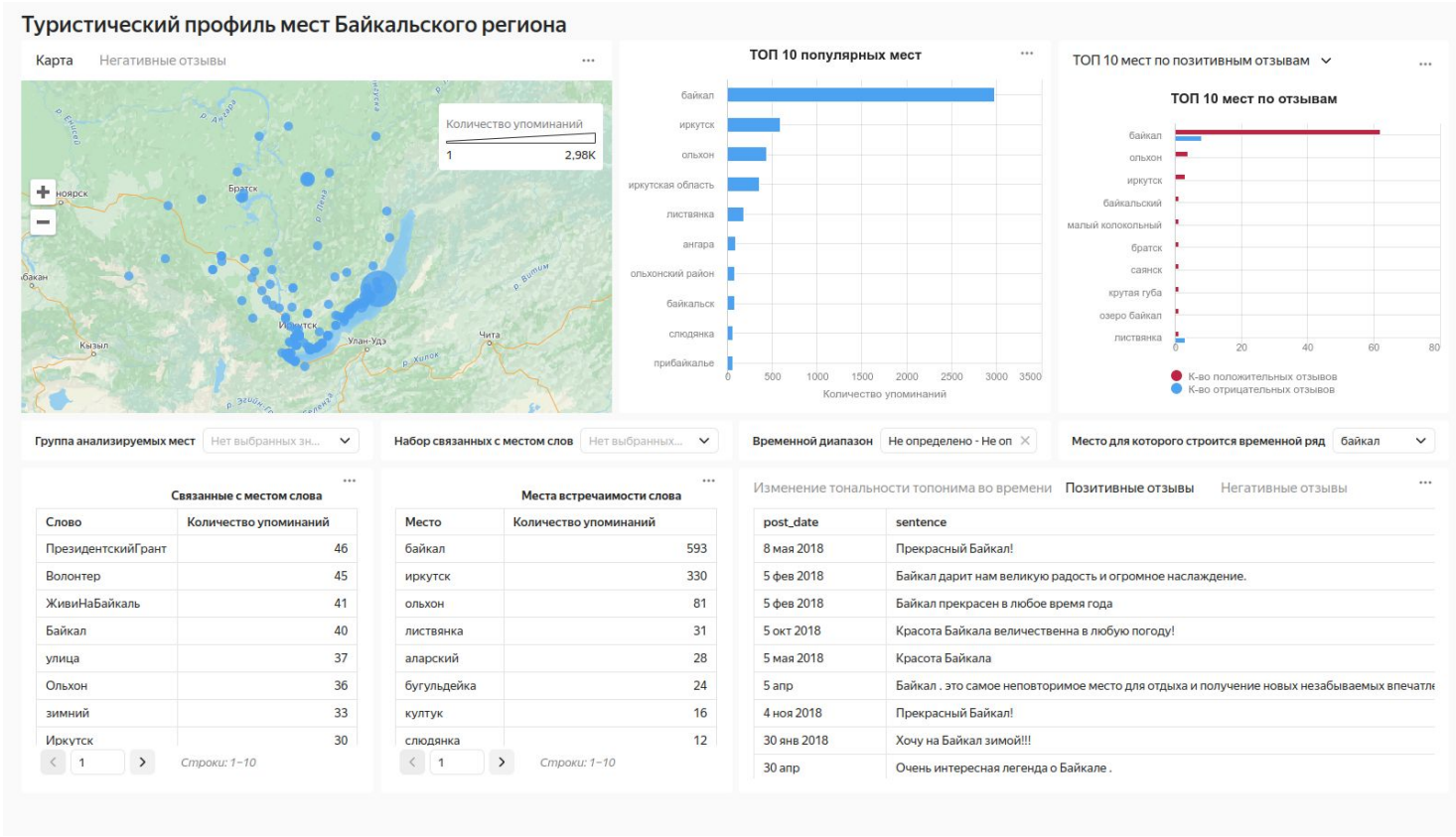
- Тональность топонима эквивалентна тональности предложения в котором он встечается.
- Тональность предложения определяется с помощью модуля Dostoevsky.
- При наличии нескольких топонимов в одном предложении, тональность всем топонимам присваевается одинаковая.

СХЕМА ДАННЫХ

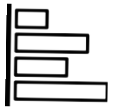


ВИЗУАЛИЗАЦИЯ

[Ссылка на дашборд]



ВИЗУАЛИЗАЦИЯ



Топы мест

1) по количеству упоминаний
2) по количеству положительных отзывов

Линейный график отображает суммарную месячную тональность выбранной локации в зависимости от времени публикации поста

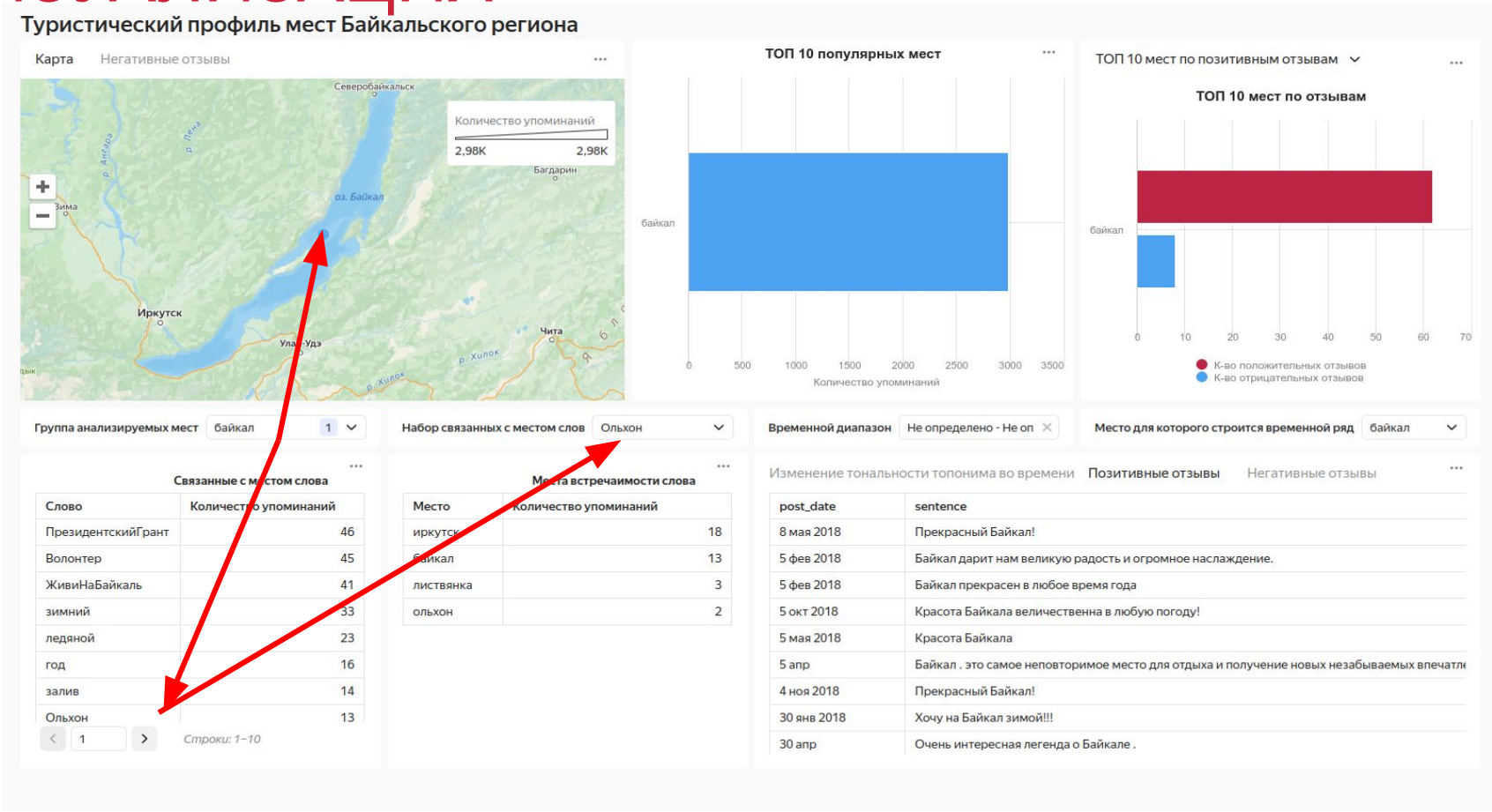
Карта с изображением выбранных локаций с учетом их посещаемости

Таблицы:

1) Места встречаемости – локации, с которыми связаны* выбранные слова
2) Связанные слова – слова связанные* с выбранными локациями

*связь – отношения между токенами, полученные в ходе синтаксического анализа с фильтрацией по части речи

ВИЗУАЛИЗАЦИЯ



РЕЗУЛЬТАТЫ АНАЛИЗА

- Обработано ~16000 постов;
- Из них выделено 243 локации Байкальского региона;
- Топ 3 - Байкал, Иркутск, Ольхон;
- Встречаются различные типы локаций: области, города, озера, реки, утесы, скальники, бухты и мысы
- Также упоминаются и объекты инфраструктуры(КБЖД, БАМ), различные базы отдыха, нац. парк
- Для Байкала позитивные оценки заметно превышают негативные, для Ольхона и Иркутска положительные равны, а негативных не обнаружено.

ЗАКЛЮЧЕНИЕ (первой части)

- В качестве источников данных выбраны группы ВК и ОК, из которых выполнен сбор данных посредством метода парсинга.
- Реализована метод анализа постов и выделения объектов притяжения туристов.
- Собранные данные обработаны, выполнена идентификация объектов притяжения туристов.
- Результаты анализа визуализированы посредством разработанных панелей визуализации с помощью возможностей BI-платформы Yandex DataLens.
- В дальнейшем планируется реализовать метод анализа постов и выделения проблем сферы туризма (проблемы размещения и питания, оказания услуг и т.п.).

НОВЫЕ ЗАДАЧИ

- Выделение и классификация проблем, связанных с туристическими объектами, из постов
- Выделение достоинств туристических объектов из постов

ИСПОЛЬЗОВАНИЕ БЯМ

- Предпосылки:
- Высокая популярность методов
- Большой объем знаний моделей, универсальность
- ПРОМТ (нет необходимости в датасете)

- Недостатки
- Необходимо большое к-во ресурсов для запуска и огромное для дообучения
- Сложность в оценке результатов
- Частичная случайность ответов, галлюцинации

Эксперементы с промтом

Инференсы CPU:

- Optimum
- Llama.cpp

Квантизация!

Hugging Face

Промт:

- zero shot
- few shot

Рекомедации по составлению промта

- Примеров много не бывает, бывает мало токенов)
- На каждый термин нерйонка имеет свое определение, иногда даже несколько. (Лучше показать на примерах)
- Если не обозначить пример, в итоге будет каша
- Задачу для модели лучше вставлять в конце (начало помнит хуже)
- Советуют в начале обзывать модель в соответствии с задачей

Итог экспериментов

Промт: zero shot + few shot
saiga2 7b
Экспертная оценка: 00000000110100110100l 33%
Экспертная оценка: 100001001000001 27%
Для 7b модели 1й промт показал себя на 6% лучше
Общее качество очень мало.
Модель слишком маленькая для данных промтов
saiga2 70b
Экспертная оценка: 001011101101011001011 57%
Экспертная оценка: 101111111001101001111 71%
Для 70b модели 2 промт оказался лучше на 14%

ЗАКЛЮЧЕНИЕ (второй части)

- 7b мадель слишком маленькая для задачи.
- Комбинация saiga2 70b + промт 2 показали лучший результат, потенциально достаточный для введения в общую систему
- Комбинация saiga2 70b + промт 2 показывает лучшую способность не выделять проблемы, если их нет
- Few-shot лучше подходит для контроля формата
- Zero-shot — для смысла

Спасибо за внимание!

Работа выполняется при поддержке проекта Российского научного фонда №23-28-00844 «Мониторинг сферы регионального туризма на основе анализа данных из открытых источников».

Комментарии:

- Возможно есть готовые модули и подходы для решения проблем предобработки
- Найти инструменты для верной расстановки точек
- Добавить словарь топонимов и за счет него исправлять опечатки
- При анализе тональности анализировать деревья deprel для разбиения сложноподчиненных и т.п. предложений с двумя основами
- Рассказывать как определяются связанные с топонимом слова