



ЛОКАЦИОННАЯ СОСТАВЛЯЮЩАЯ И ЛЕКСИЧЕСКАЯ НАПОЛНЕННОСТЬ ТЕЛЕГРАМ-КАНАЛА О ПУТЕШЕСТВИЯХ

ПУТЕШЕСТВИЯ ПО РОССИИ • GO RUSSIA

[HTTPS://T.ME/TRAVEL_RUSSIA1](https://t.me/travel_russia1)

186 930 ПОДПИСЧИКОВ

→ ГОРА «ЧЁРТОВ ПАЛЕЦ», АЛТАЙ

Компьютерная лингвистика — 2025

ЗАДАЧИ ПРОЕКТА

- собрать корпус текстов телеграм-канала «Путешествия по России • Go Russia»;
- структурировать данные телеграм-канала, работа с Pandas;
- определить самые частые локации по количеству упоминаний на канале;
- оценить взаимосвязь количества реакций с популярностью контента;
- сравнить лексическое наполнение статей в зависимости от:
 - разных регионов;
 - разных периодов публикаций;
- биграммы как основа рекламных текстов.



→ Гора Паасо, Республика Карелия



ЗАДАЧА 1



Собрать корпус текстов телеграм-канала
«Путешествия по России • Go Russia»

- экспорт HTML-файлов;
- извлечение данных с помощью библиотеки BeautifulSoup и модуля re;
- теги `<body>`, `</body>`.

→ вид с острова Атласова, Курилы



ЗАДАЧА 2

Структурировать данные телеграм-канала, работа с Pandas

df. head(250)

	Заголовок	Текст	Дата публикации	Эмодзи	Локация на карте	Название региона
0	Работа на Юге. ТОП-15 Вакансий	Строительная сфера на юге развивается семимиль...	23.07.2023 10:13:35 UTC+04:00	{'👍': 23, '❤️': 1, '😬': 1, '😬': 1}		
1	Куда дети перевозят своих родителей? Где прове...	Климат морской полезен для всех. Очень часто м...	24.07.2023 11:45:50 UTC+04:00	{'👍': 23, '👍': 3}		
2	Головой на море, "одним местом" на диване.	Если вы в эти дни не на морене расстраивайтесь...	25.07.2023 12:48:20 UTC+04:00	{'👍': 11, '❤️': 2, '😬': 1}		
3	Лучший город России. Рейтинг	Ты все ещё ищешь лучший город России? Блогер-у...	26.07.2023 14:52:33 UTC+04:00	{'👍': 53, '👍': 13, '❤️': 5}		
4	Жить у моря полезно.	Я тут наткнулся на статью про воду там ученые ...	27.07.2023 10:55:21 UTC+04:00	{'👍': 17, '❤️': 2, '😬': 1}		
...
245	Уксинская озовая гряда – чудо природы, оставше...	Это одно из самых красивых мест Карелии ! Ледн...	26.09.2023 17:36:07 UTC+04:00	{'❤️': 268, '👍': 182, '😬': 38, '😬': 9, '😬': 3, ...}	https://yandex.ru/maps? whatshere%5Bpoint%5D=31...	#республикакарелия
246	Необычные Сырные скалы в Карачаево-Черкесии.	В Карачаево-Черкесии есть одно необычное место...	26.09.2023 19:18:01 UTC+04:00	{'😬': 260, '👍': 170, '❤️': 42, '❤️': 11, '😬': ...}	https://yandex.ru/maps/org/ syrynaya_peshchera/1...	#карачаевочеркесия
247	Усадьба Нероново – погибающее имперское наследие.	Усадьба Нероново относится к числу наиболее ин...	26.09.2023 22:07:06 UTC+04:00	{'❤️': 249, '😬': 204, '👍': 91, '😬': 30, '😬': 10...	https://yandex.ru/maps/org/ usadba_neronovo/202...	#костромскаяобласть
248	Особняк Василия Каншина, содержателя питейных ...	Василий Семёнович Каншин был богатейшим челове...	27.09.2023 09:07:01 UTC+04:00	{'👍': 320, '❤️': 52, '❤️': 32, '😬': 23, '😬': 2}	https://yandex.ru/maps? whatshere%5Bpoint%5D=30...	#санктпетербург
249	Горнолыжный курорт «Матлас	в Дагестане. На горнолыжном курорте Матлас раз...	27.09.2023 12:14:01 UTC+04:00	{'👍': 238, '😬': 78, '😬': 41, '❤️': 31, '😬': 9, ...}	https://yandex.ru/maps/org/ gornolyzhny_kurort_...	#республикадагестан

250 rows × 6 columns

Структура датафрейма:

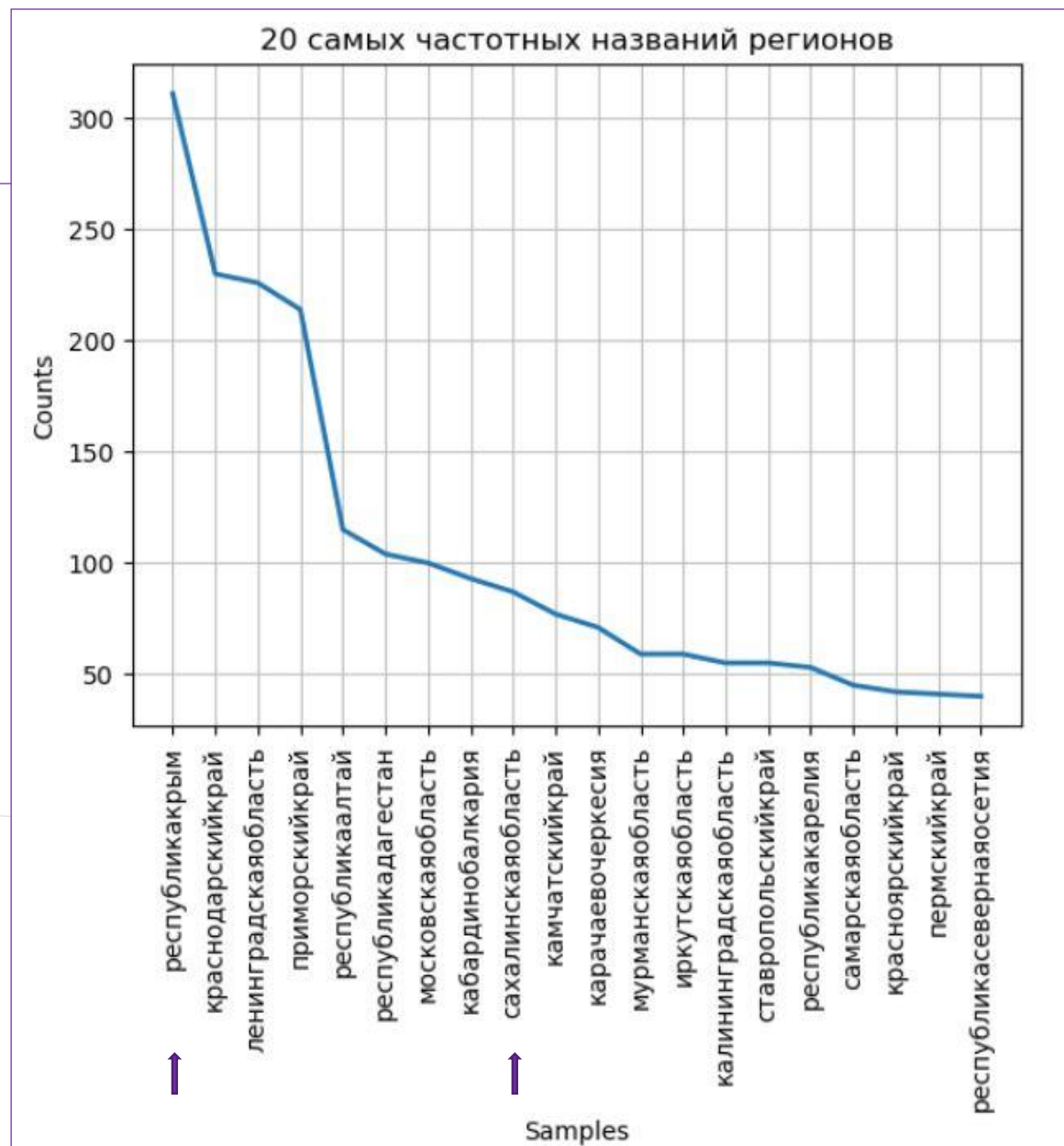
- заголовок поста;
- текст поста;
- дата публикации;
- реакции;
- локация на карте;
- название региона.



ЗАДАЧА 3

Определить самые частые локации по количеству упоминаний на канале

- работа с датафреймом: названия регионов, общий список;
- предобработка;
- частотное распределение FreqDist

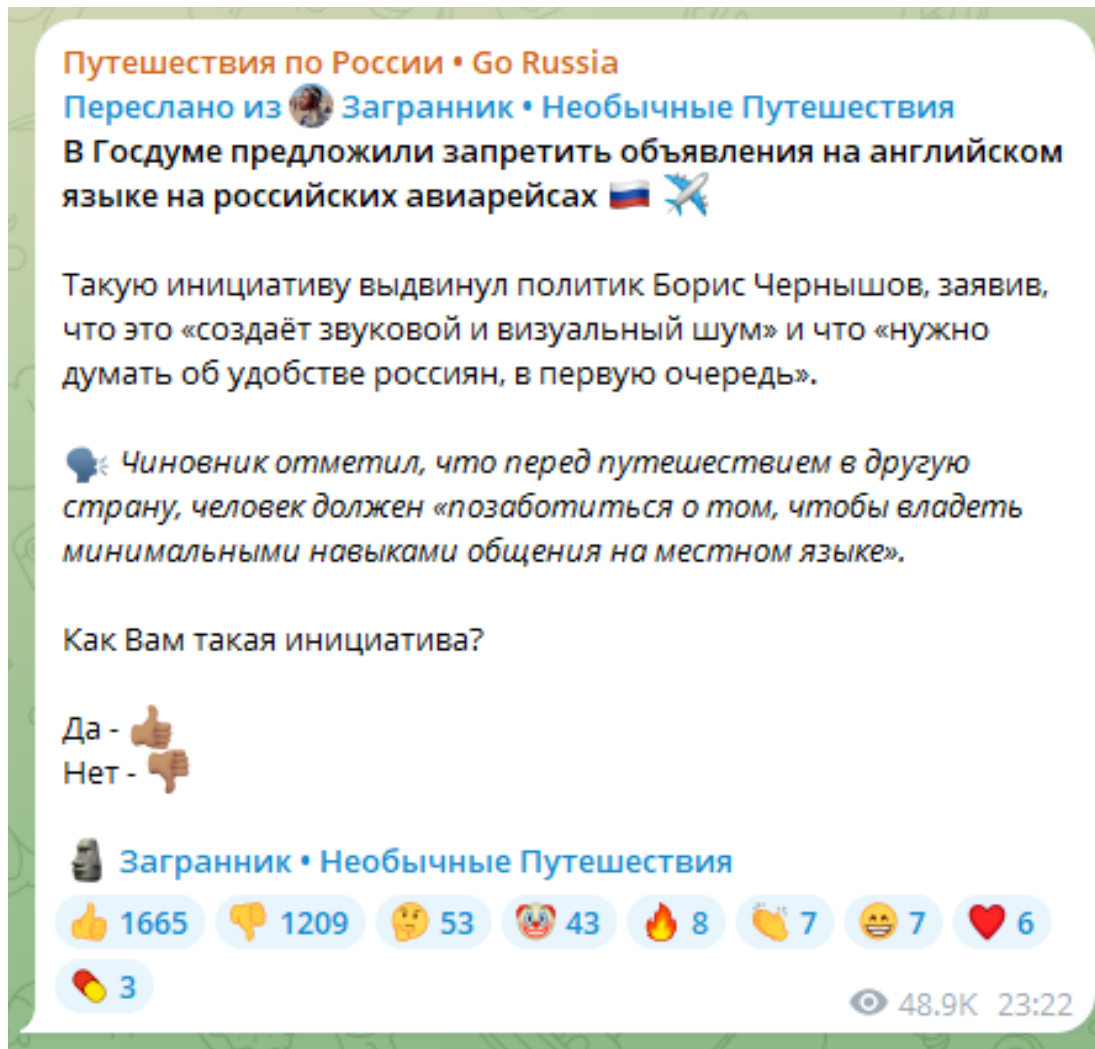


ЗАДАЧА 4

- общее количество реакций — 2999;
- сторонний пост из другого телеграм-канала;
- прямое обращение к пользователю → побуждение к действию.

Вывод: количество реакций не всегда показатель заинтересованности и тематических предпочтений читателей

Оценить взаимосвязь количества реакций с популярностью контента

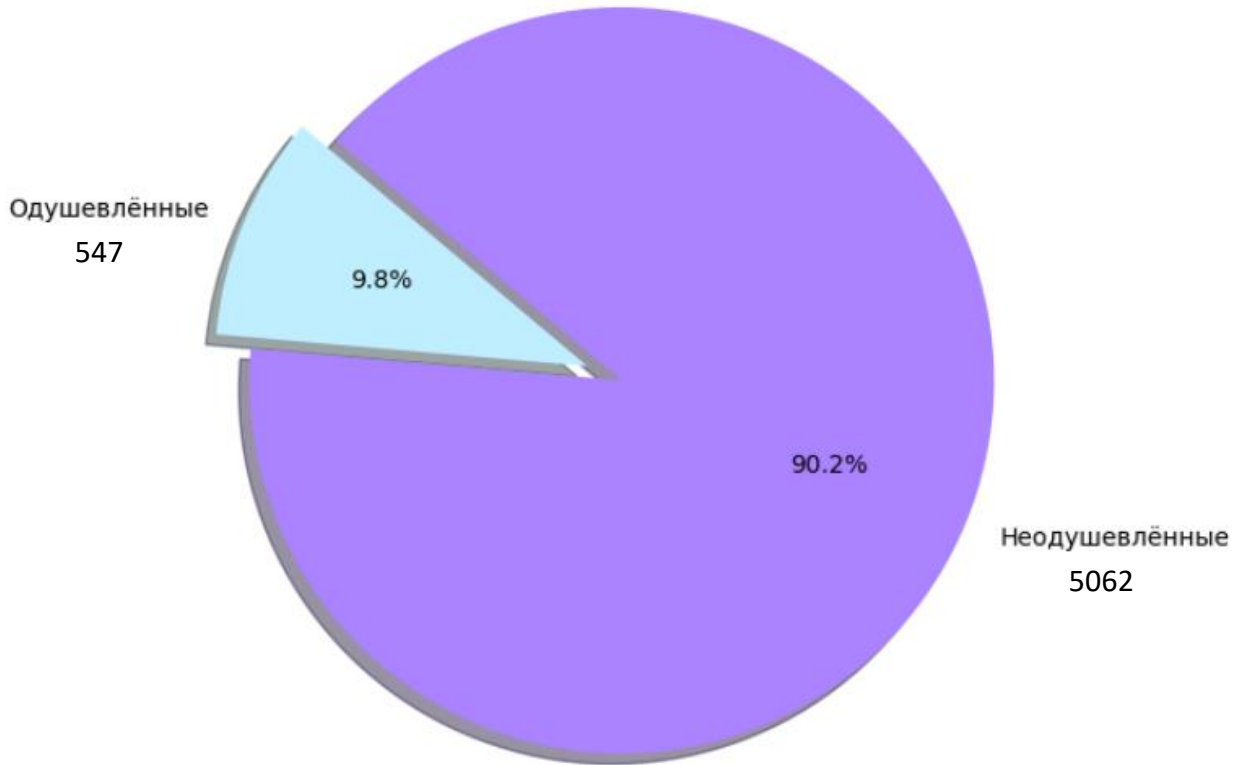


ЗАДАЧА 5

Сравнить лексическое наполнение статей в зависимости от разных регионов

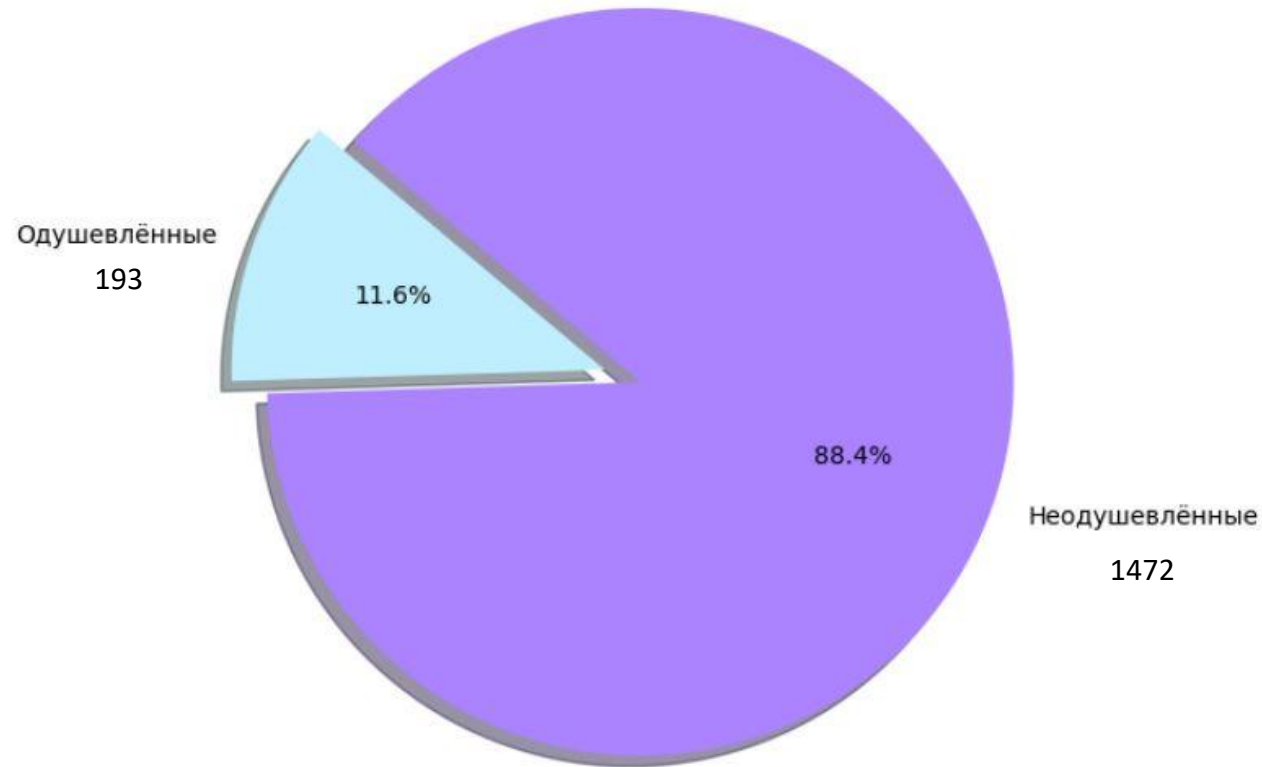
Республика Крым

Состав существительных по одушевленности



Сахалинская область

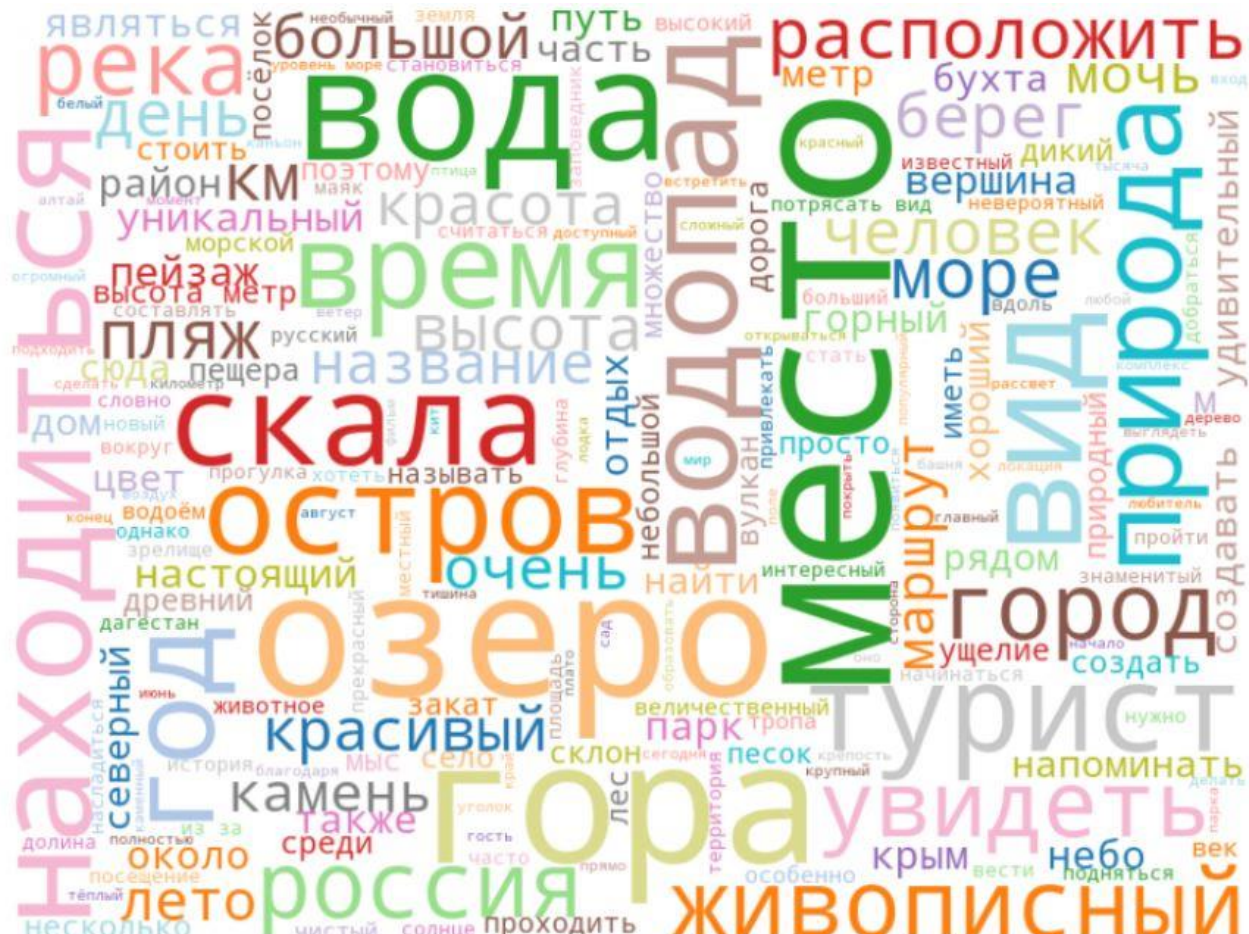
Состав существительных по одушевленности



- количественное соотношение одушевлённых/неодушевлённых существительных в текстах

Сравнить лексическое наполнение статей в зависимости от периодов публикаций

Лето — 2024 • количество лемм — 131876

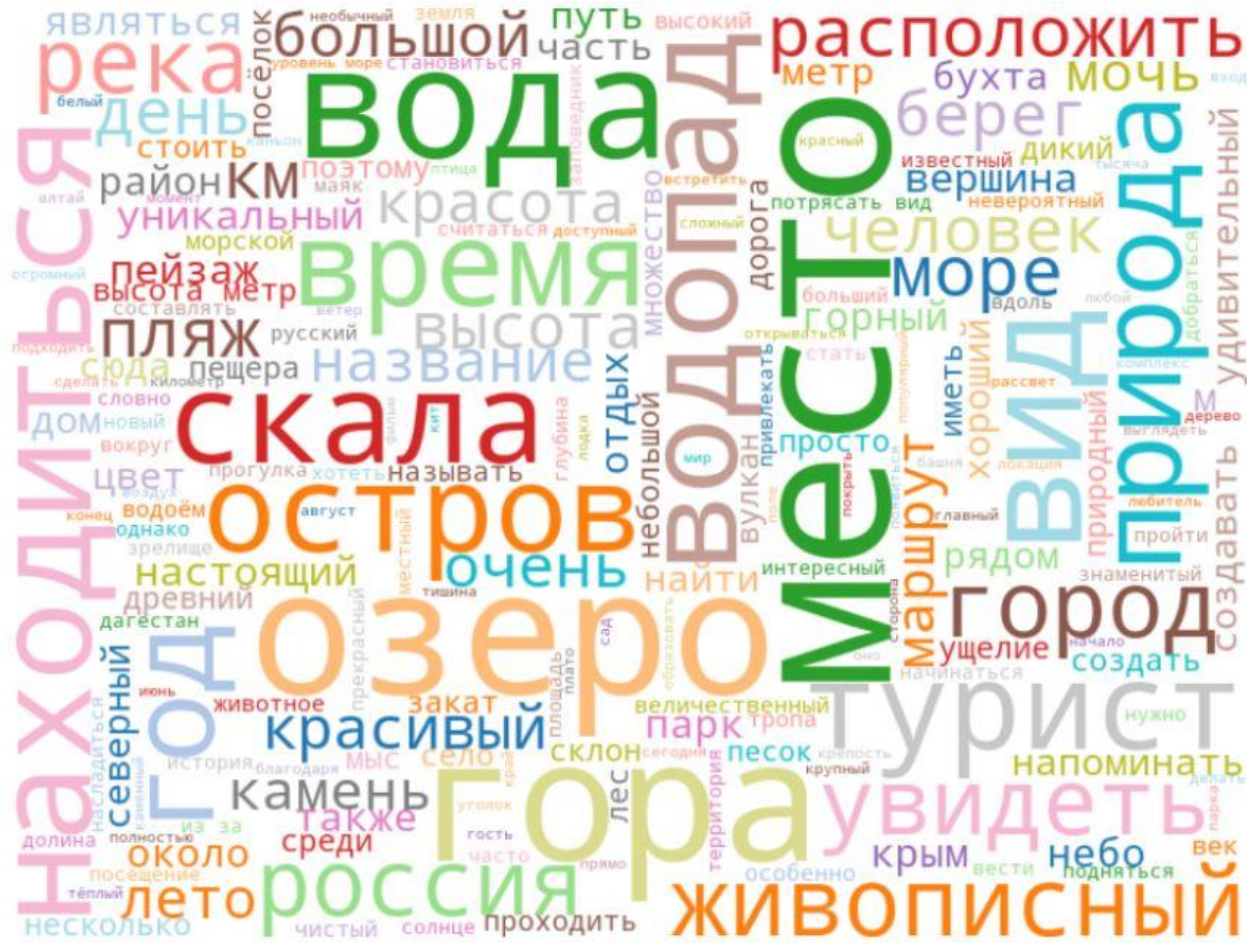


Сравнить лексическое наполнение статей в зависимости от периодов публикаций

- количество лемм — 118881





- количество лемм — 131876



ЗАДАЧА 6

Биграммы как основа рекламных текстов



Погрузитесь в удивительный мир, где сливаются магия северного сияния и потрясающие пейзажи! 🌲🌌

Приглашаем вас в наш уникальный путешествие в национальный парк, где вас ждут захватывающие моменты и незабываемые впечатления. Откройте для себя диковинную дикую природу, наполненную яркими красками и удивительными звуками.

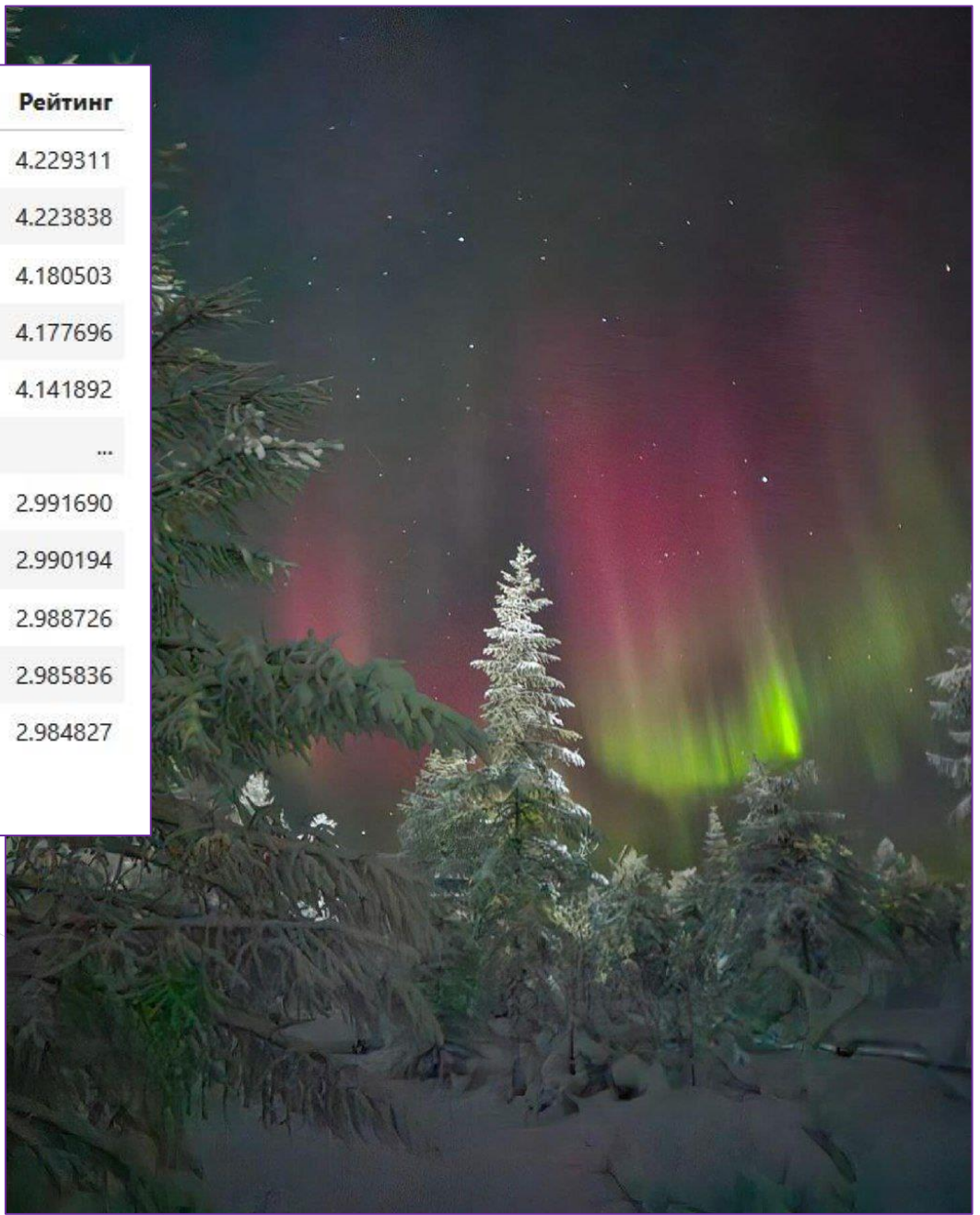
Насладитесь красивыми видами, которые будут радовать ваш взгляд, и ощутите всю силу природной красоты, окружающей вас. Этот парк — настоящий рай для фотографов и любителей активного отдыха!

Не упустите шанс стать частью этого волшебства. Забронируйте свой тур уже сегодня и откройте для себя северное сияние в полной красе! 🌌💚

18:48

	Биграммы	Рейтинг
39	северный сияние	4.229311
40	национальный парк	4.223838
41	красивый вид	4.180503
42	дикий природа	4.177696
43	природный красота	4.141892
...
139	хвойный лес	2.991690
140	баренцев море	2.990194
141	туристический инфраструктура	2.988726
142	балтийский море	2.985836
143	неповторимый атмосфера	2.984827

105 rows × 2 columns





СПАСИБО ЗА ВНИМАНИЕ!

→ ПОСЁЛОК СИМЕИЗ, РЕСПУБЛИКА КРЫМ