

# Образ Сахалина в японских СМИ

**Студенты:** Якушева Елена, Цыбикова Екатерина,  
Лаптева Ксения, Ников Максим, Шестопалов Тимофей

**Тьюторы:** Ушак Ксения, Вещикова Оксана



## Исследовательский вопрос

Как сегодня представлен Сахалин в Интернет-пространстве Японии, в том числе в туристических блогах?

## Актуальность

Туризм - приоритетное направление развития Сахалинской области

## Источники

Более 500 статей

- news.yahoo.co.jp
- 4travel.jp

The screenshot shows a Jupyter Notebook interface with the following details:

- Header:** CO, Диалог с Азией - Туризм (Япония), ☆, Chat icon, Settings icon, Share icon, User icon (Поделиться), K.
- Toolbar:** Файл, Изменить, Вид, Вставка, Среда выполнения, ▾
- Search Bar:** Команды, + Код, + Текст, ▶ Выполнить все ▾
- File Status:** ✓ ОЗУ [green bar], Диск [yellow bar] ▾
- Left Sidebar:** Icons for Cell, Cell Editor, Diff, Kernel, and File.
- Code Cell:** Contains Python code for generating a word cloud from travel blog data.

```
# Построение облака слов по всему корпусу текстов (Заголовки + Тексты)

# --- Функция для создания облака слов из DataFrame ---
def create_wordcloud_from_df(df, columns, title="", max_words=30):

    # Объединяем все токены из всех колонок
    all_words = []
    for col in columns:
        for tokens in df[col]:
            all_words.extend(tokens)

    # Генерируем облако слов
    wc = WordCloud(
        font_path='/usr/share/fonts/opentype/ipafont-gothic/ipagp.ttf',
        width=800,
        height=400,
        background_color='white',
        max_words=max_words
    ).generate(" ".join(all_words))

    # Отображаем
    plt.figure(figsize=(12,6))
    plt.imshow(wc, interpolation='bilinear')
    plt.axis("off")
    plt.title(title, fontsize=16)
    plt.show()

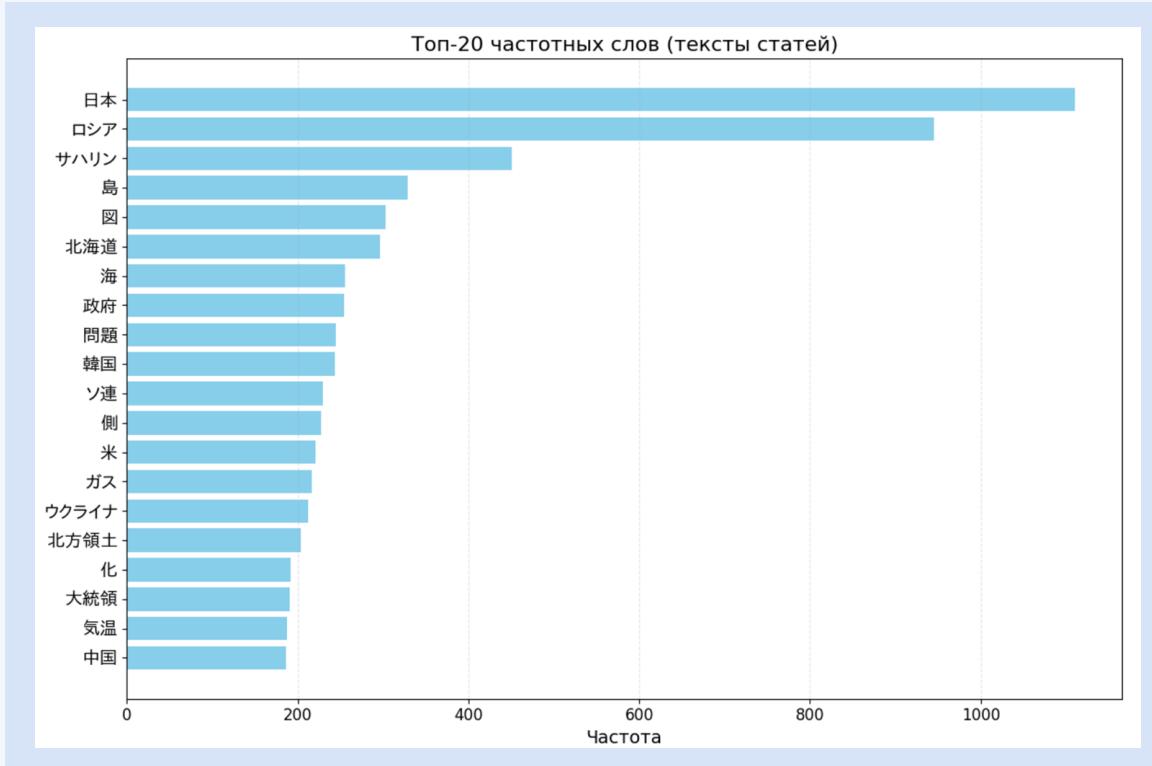
# --- Создаём общее облако слов для всех текстов ---
create_wordcloud_from_df(df_karafuto, columns=['tokens_header_clean', 'tokens_body_clean'])
```
- Bottom Navigation:** Переменные, Терминал, 00:04, Python 3

# Методы исследования данных

<b>Парсинг</b>	Автоматизированный сбор данных с веб-сайтов. Программа открывает страницы, анализирует содержимое и извлекает нужную информацию (заголовки, даты, тексты, ссылки), экономя время на ручном сборе данных.
<b>Токенизация</b>	Разбиение непрерывного текста на отдельные смысловые единицы (токены): слова, знаки препинания, числа.
<b>Частотный анализ</b>	Используется для выявления определения наиболее употребительных слов; ключевых тем и понятий.
<b>Анализ совместной встречаемости (co-occurrence analysis)</b>	Метод помогает понять какие слова в тексте часто встречаются вместе, т.к. это явный признак того, что они связаны между собой.
<b>LDA</b>	Метод тематического моделирования, который помогает автоматически выявлять скрытые темы в большом наборе текстов.

# Анализ по поиску “サハリン” (Сахалин)

## Частотный анализ



## LDA : Темы

Международные отношения

Проблема Северных  
территорий

Энергетика и  
международная торговля

Коренные народы

Погода

# Анализ по поиску “樺太” (Карафуто)

## LDA : Темы

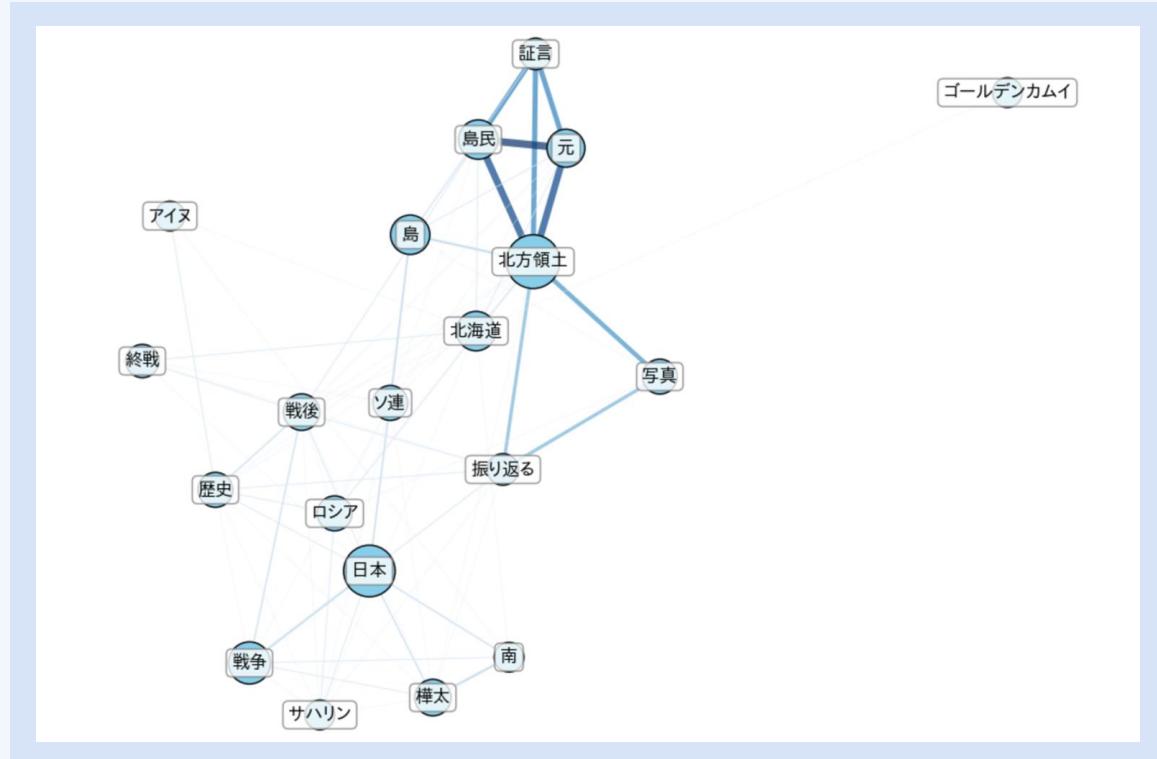
История Японии 20 века

Древние коренные народы

Манга и аниме

Погода в  
историческом контексте

## Анализ совместной встречаемости



# Анализ туристических заметок

## Облако слов



## Выводы

## Главный туристический интерес – японское прошлое Сахалина

## Целевая аудитория – японцы возрастом 60+

## Сообщение не только паромом из Хоккайдо, но и в рамках круизов

## Японцы активно пользуются электронной визой при въезде в Россию