

## Написание парсера значений дневника погоды сервиса Gismeteo

*Вихляев Дмитрий Романович*

*Приамурский государственный университет имени Шолом-Алейхема*

*Студент*

### Аннотация

В данной статье приведён метод парсинга числовых, строковых и графических значений предоставляемых в дневнике погоды системы Gismeteo. Пример реализации программы написан на языке python. В результате работы будет описан процесс написания парсера, а также представлены инструменты для эффективной работы с извлечёнными данными.

**Ключевые слова:** парсинг, python, Cookie, сессия.

## Writing a parser for the weather diary values of the Gismeteo service

*Vikhlyayev Dmitry Romanovich*

*Sholom-Aleichem Priamursky State University*

*Student*

### Abstract

This article provides a method for parsing numeric, string and graphic values provided in the weather diary of the Gismeteo system. An example program implementation is written in Python. As a result of the work, the process of writing a parser will be described, as well as tools for effectively working with extracted data will be presented.

**Keywords:** parsing, python, Cookie, session.

## 1 Введение

### 1.1 Актуальность

Сбор метеорологических данных через парсинг интернет-ресурсов становится все более актуальным в современных научных исследованиях и прикладных задачах. Это обусловлено высокой доступностью и широким охватом данных, которые предоставляют специализированные сайты, веб-сервисы и API. В условиях цифровизации многие метеорологические агентства и частные компании публикуют информацию о погодных условиях, климатических изменениях и прогностических моделях в режиме реального времени, что делает такие данные востребованными для анализа.

Особенности сбора метеоданных через парсинг включают в себя ряд ключевых аспектов. Во-первых, этот метод позволяет автоматизировать процесс получения информации, минимизируя ручной труд и исключая человеческие ошибки. Во-вторых, парсинг дает возможность извлекать данные с различных источников одновременно, что повышает точность и

репрезентативность собранной информации. В-третьих, современные инструменты парсинга поддерживают интеграцию с языками программирования, такими как Python, что упрощает последующую обработку, хранение и визуализацию данных.

### **1.2 Обзор исследований**

И.В.Зайцева, С.Н.Фадеев, К.Р.Гончик, Т.И.Малыгина применили метод Монте-Карло для исследования климатических показателей на основании статистических данных Гисметео [1]. Е.А.Коркина, И.С.Бакланова рассчитали климатический показатель, для электронной карты проанализировав архивные климатические данные с портала GISmeteo для г. Нижневартовска [2]. А.С.Кучерина написала обзор методов и систем обработки данных с веб-страниц [3]. Р.Ю.Демина, Д.Э.Шукралиева исследовали методы защиты web-ресурсов от вредоносных парсеров [4]. В.Ю.Чистяков в своей статье обработал результаты парсинга [5].

### **1.3 Цель исследования**

Цель исследования – описать метод реализации парсинга простых и сложных типов данных на примере дневника погоды Gismeteo.

## **2 Материалы и методы**

Для работы используется язык программирования python и загрузочные модули. Извлечение данных из изображений осуществляется, через определения имени файла в ссылке.

## **3 Результаты и обсуждения**

Сервис Gismeteo предоставляет на странице дневника погоды в архивные погодные данные, в свободном доступе [7]. Данные представлены в виде HTML таблицы с двумя уровнями заголовков. Колонки разделяют показатели на дневные и вечерние значения. Среди показателей присутствуют температура воздуха, давление, облачность, наличие осадков и данные ветра. При этом значения температуры и давления представлены в числовом виде. Информация о наличие облачности и осадков указана виде условных обозначений растровым изображением. Данные о ветре приведены одновременно как строка и изображение, для направления ветра и числовое значение для скорости. Данные можно получить только, как совокупность за конкретный месяц года (рис.1).

Дневник погоды в Биробиджане за Ноябрь 2023 г. печатная версия:  цветная  Ч/Б

| Число | День        |          |                                                                                     |         |          | Вечер       |          |                                                                                       |         |          |
|-------|-------------|----------|-------------------------------------------------------------------------------------|---------|----------|-------------|----------|---------------------------------------------------------------------------------------|---------|----------|
|       | Температура | Давление | Облачность                                                                          | Явления | Ветер    | Температура | Давление | Облачность                                                                            | Явления | Ветер    |
| 1     | +4          | 755      |    |         | СЗ 3 м/с | -4          | 756      |    |         | СЗ 3 м/с |
| 2     | -2          | 750      |    | *       | В 4 м/с  | -3          | 745      |    | *       | В 4 м/с  |
| 3     | -6          | 756      |    |         | З 5 м/с  | -13         | 761      |    |         | З 5 м/с  |
| 4     | -5          | 764      |    |         | ЮЗ 4 м/с | -13         | 765      |    |         | ЮЗ 4 м/с |
| 5     | -1          | 765      |    |         | Ш        | -6          | 765      |    |         | Ш        |
| 6     | -4          | 752      |    |         | СВ 3 м/с | -2          | 749      |    |         | СВ 3 м/с |
| 7     | -4          | 752      |    | *       | СВ 1 м/с | -5          | 758      |    | *       | СВ 1 м/с |
| 8     | -3          | 762      |    |         | В 3 м/с  | -5          | 762      |    |         | В 3 м/с  |
| 15    | -1          | 760      |    |         | Ю 1 м/с  | -14         | 761      |    |         | Ю 1 м/с  |
| 16    | -4          | 758      |    |         | СВ 2 м/с | -4          | 758      |    | *       | СВ 2 м/с |
| 17    | 0           | 751      |    | *       | С 4 м/с  | +1          | 749      |    | *       | С 4 м/с  |
| 18    | +5          | 740      |    |         | СЗ 2 м/с | 0           | 742      |    |         | СЗ 2 м/с |
| 19    | -1          | 743      |    |         | СЗ 2 м/с | -5          | 746      |    |         | СЗ 2 м/с |
| 20    | +1          | 750      |    |         | ЮЗ 2 м/с | -3          | 749      |    |         | ЮЗ 2 м/с |
| 21    | +3          | 748      |    |         | ЮЗ 2 м/с | -6          | 752      |    |         | ЮЗ 2 м/с |
| 22    | -2          | 756      |    |         | СЗ 2 м/с | -6          | 754      |    |         | СЗ 2 м/с |
| 23    | -6          | 749      |   |         | С 4 м/с  | -11         | 752      |   |         | С 4 м/с  |
| 24    | -14         | 755      |  |         | З 5 м/с  | -18         | 757      |  |         | З 5 м/с  |
| 25    | -13         | 759      |  |         | З 6 м/с  | -17         | 761      |  |         | З 6 м/с  |
| 26    | -11         | 757      |  |         | З 6 м/с  | -15         | 758      |  |         | З 6 м/с  |
| 27    | -14         | 753      |  |         | ЮВ 1 м/с | -22         | 752      |  |         | ЮВ 1 м/с |
| 28    | -12         | 750      |  |         | ЮВ 1 м/с | -19         | 752      |  |         | ЮВ 1 м/с |
| 29    | -14         | 753      |  |         | Ю 2 м/с  | -22         | 755      |  |         | Ю 2 м/с  |
| 30    | -13         | 754      |  |         | ЮЗ 3 м/с | -23         | 755      |  |         | ЮЗ 3 м/с |

Рис. 1. Таблица архивных показателей дневника погоды в городе Биробиджане за ноябрь 2023 года

Если брать во внимание только показатели со значениями в виде условных обозначений, то имеется 7 уникальных изображений. Четыре обозначения для описания облачности и три для описания вида осадков. При этом показатель облачности всегда имеет значение, а показатель осадков допускает пропуск, если осадки отсутствовали. Стоит отметить, что каждое изображение имеет своё уникальное имя (рис.2).

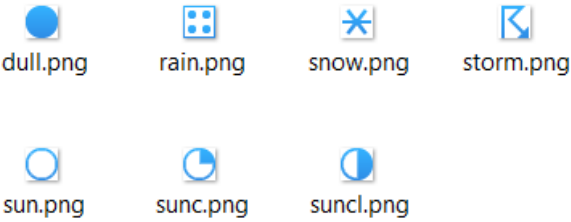


Рис. 2. Условные обозначения показателей в дневнике погоды

При обращении к дневнику погоды, требуется указать три параметра. Идентификатор города, является числом и уникальным значением города в системе Gismeteo. Значение можно получить из url-адреса на странице прогноза для данного города или после выборки в дневнике погоды. Город Биробиджан имеет идентификатор 4860. Остальные два параметра, год и месяц соответственно, указанные в числовом виде. Сервис требует наличие заголовочного значения User-Agent, а также идентификатора сессии записанного в Cookie. Модуль requests в языке python обладает методом Session, с помощью которого можно автоматически обновлять значение Cookie в заголовке при обращении к ресурсу (рис.3).

```
url = "https://www.gismeteo.ru/diary/{}/{}/{}/".format(city_id,year,month)

headers = {
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36"
}

session = requests.Session()
session.headers.update(headers)

try:
    response = session.get(url)
    response.raise_for_status() # Проверяет, был ли ответ успешным (код 2xx)
    return response.text
except requests.exceptions.RequestException as e:
    print(f"Ошибка запроса: {e}")
    return None
```

Рис. 3. Метод отправки GET-запроса на ресурс дневника погоды

На рисунке 4 представлено содержание первой половины строки HTML-таблицы. Чтобы получить необходимые значения не используя, методика распознавания изображений, достаточно вытащить название изображения из адреса пути указанного в теги img. Причём достаточно получить только первую ссылку.

```
<tr align="center"> == $0
  <td class="first">2</td>
  <td class="first_in_group">-2</td>
  <td>750</td>
  <td>
    
    
  </td>
  <td>
    
    
  </td>
  <td>
    <span>
      
      
      <br>
      "B 4м/с"
    </span>
  </td>
```

Рис. 4. Содержание строки HTML-таблицы

С помощью инструмента BeautifulSoup, доступного как загрузочного модуля языка python, можно без труда распарсить html страницу. Метод find находит первый встретившийся элемент. Метод find\_all находит все

элементы в файле. Произведя обход по каждой строке и ячейке нужно проверить наличие тега `img`. Если элемент найден, извлекается значение из атрибута `src`, а затем вытаскивается имя изображения без расширения. Полученная строка полностью заменяет содержимое ячейки (рис.5).

```
# Парсинг через BeautifulSoup
soup = BeautifulSoup(html_doc, "html.parser")

# Извлекаем таблицу
table = soup.find('table')

# Теперь работаем с таблицей
for row in table.find_all('tr'):
    # Находим все <td> в строке
    td_tags = row.find_all('td')

    # Если в строке есть изображение
    for td_tag in td_tags:
        img_tag = td_tag.find('img', src=True)
        if img_tag:
            # Извлекаем ссылку из атрибута src
            img_url = img_tag.get('src')

            # Извлекаем имя файла без расширения
            filename = os.path.splitext(os.path.basename(img_url))[0]

            # Заменяем тег <img> на имя файла
            img_tag.replace_with(filename)

# Извлекаем обновленную таблицу
return str(soup.find('table'))
```

Рис. 5. Извлечение имени условных обозначений

После, того как HTML-таблица имеет только числовые и строковые значения, можно воспользоваться методом `read_html` модуля `pandas`. Такой метод быстро распределит значения в табличную структуру данных `DataFrame`. Но если бы внутри ячейки содержались более сложные данные как вложенные HTML-теги, как было изначально для показателей с уникальными значениями, данные для этих полей заполнились пропусками.

Стоит отметить, что в данном случае метод распределил заголовки в конце таблицы как значения. Для удобного обращения последние две строки вручную установлены как заголовки, а структура изменена на `MultiIndex`, позволяющий иметь иерархическую структуру заголовков или индексов (рис.6).

```
dfs = pd.read_html(StringIO(html_doc), index_col = 0);
df = dfs[0]

# Извлечение последних двух строк как заголовков
header_row_1 = df.iloc[-2]
header_row_2 = df.iloc[-1]

# Создание MultiIndex из двух строк
multi_index = pd.MultiIndex.from_arrays([header_row_1, header_row_2])

# Обновление DataFrame с MultiIndex
df.columns = multi_index
df = df.iloc[:-2] # Удаляем строки, которые использовали как заголовки
```

Рис. 6. Преобразование HTML-таблицы в pandas DataFrame MultiIndex

Извлечённые данные имеют нормализованный вид и хранятся в структуре данных позволяющих эффективно проводить анализ и продолжать обработку данных (рис.7).

| Число | День        |          |            |         |            | Вечер       |          |            |         |            |
|-------|-------------|----------|------------|---------|------------|-------------|----------|------------|---------|------------|
| Число | Температура | Давление | Облачность | Явления | Ветер      | Температура | Давление | Облачность | Явления | Ветер      |
| 0     |             |          |            |         |            |             |          |            |         |            |
| 1     | 4           | 755      | sunc       |         | w7 C3 3м/с | -4          | 756      | sun        |         | w7 C3 3м/с |
| 2     | -2          | 750      | dull       | snow    | w2 B 4м/с  | -3          | 745      | dull       | snow    | w2 B 4м/с  |
| 3     | -6          | 756      | sun        |         | w6 3 5м/с  | -13         | 761      | sun        |         | w6 3 5м/с  |
| 4     | -5          | 764      | sun        |         | w5 Ю3 4м/с | -13         | 765      | sun        |         | w5 Ю3 4м/с |
| 5     | -1          | 765      | sunc       |         | Ш          | -6          | 765      | dull       |         | Ш          |
| 6     | -4          | 752      | dull       |         | w1 CB 3м/с | -2          | 749      | dull       |         | w1 CB 3м/с |
| 7     | -4          | 752      | dull       | snow    | w1 CB 1м/с | -5          | 758      | dull       | snow    | w1 CB 1м/с |
| 8     | -3          | 762      | dull       |         | w2 B 3м/с  | -5          | 762      | dull       |         | w2 B 3м/с  |
| 15    | -1          | 760      | sun        |         | w4 Ю 1м/с  | -14         | 761      | sun        |         | w4 Ю 1м/с  |
| 16    | -4          | 758      | dull       |         | w1 CB 2м/с | -4          | 758      | dull       | snow    | w1 CB 2м/с |
| 17    | 0           | 751      | dull       | snow    | w0 C 4м/с  | 1           | 749      | dull       | snow    | w0 C 4м/с  |
| 18    | 5           | 740      | sunc       |         | w7 C3 2м/с | 0           | 742      | sun        |         | w7 C3 2м/с |
| 19    | -1          | 743      | sunc       |         | w7 C3 2м/с | -5          | 746      | sun        |         | w7 C3 2м/с |
| 20    | 1           | 750      | sunc       |         | w5 Ю3 2м/с | -3          | 749      | sun        |         | w5 Ю3 2м/с |
| 21    | 3           | 748      | sun        |         | w5 Ю3 2м/с | -6          | 752      | sun        |         | w5 Ю3 2м/с |
| 22    | -2          | 756      | sunc       |         | w7 C3 2м/с | -6          | 754      | sun        |         | w7 C3 2м/с |
| 23    | -6          | 749      | sunc       |         | w0 C 4м/с  | -11         | 752      | sun        |         | w0 C 4м/с  |
| 24    | -14         | 755      | sun        |         | w6 3 5м/с  | -18         | 757      | sun        |         | w6 3 5м/с  |
| 25    | -13         | 759      | sun        |         | w6 3 6м/с  | -17         | 761      | sun        |         | w6 3 6м/с  |
| 26    | -11         | 757      | sun        |         | w6 3 6м/с  | -15         | 758      | sun        |         | w6 3 6м/с  |
| 27    | -14         | 753      | sunc       |         | w3 ЮВ 1м/с | -22         | 752      | sunc       |         | w3 ЮВ 1м/с |
| 28    | -12         | 750      | sunc       |         | w3 ЮВ 1м/с | -19         | 752      | sun        |         | w3 ЮВ 1м/с |
| 29    | -14         | 753      | sun        |         | w4 Ю 2м/с  | -22         | 755      | sunc       |         | w4 Ю 2м/с  |
| 30    | -13         | 754      | sunc       |         | w5 Ю3 3м/с | -23         | 755      | sun        |         | w5 Ю3 3м/с |

Рис. 7. Извлечённые данные в табличном представлении

#### 4 Выводы

В результате работы рассмотрен пример парсинга HTML-таблицы и загрузки данных в структуру pandas DataFrame. Описан способ создания и поддержания сессии. Представлен вариант извлечения данных изображения по имени файла. Извлечённые данные приведены к виду, способствующему дальнейшей обработке.

#### Библиографический список

1. Зайцева И.В., Фадеев С.Н., Гончик К.Р., Малыгина Т.И. Применение метода Монте-Карло для исследования климатических показателей // Информационные технологии и системы: управление, экономика, транспорт, право. 2022. № S1. С. 44-46.
2. Коркина Е.А., Бакланова И.С. Расчет климатического показателя для электронной карты почвенно-экологической оценки земель таежной зоны западной Сибири (в пределах нижневартковского района) // Бюллетень науки и практики. 2017. № 5 (18). С. 86-94.
3. Кучерина А.С. Обзор методов и систем обработки данных с веб-страниц // В сборнике: Проблемы управления в социально-экономических и технических системах. Материалы XIX Международной научно-практической конференции. Саратов, 2023. С. 390-394.

4. Демина Р.Ю., Шукралиева Д.Э. Методы защиты web-ресурсов от вредоносных парсеров // В сборнике: Проблемы повышения эффективности научной работы в оборонно-промышленном комплексе России. Материалы VI Всероссийской научно-практической конференции. Составитель С.Н. Бориско. Астрахань, 2023. С. 96-100.
5. Чистяков В.Ю. Обработка результата парсинга // RSDN Magazine. 2014. № 1. С. 47-54.
6. Жукова Ж.С., Ерофеева В.В., Кулагин Д.А., Шварцман К.С., Яблочников С.Л. Разработка комплекса программ парсеров для сбора климатических данных // Вестник Кыргызско-Российского Славянского университета. 2024. Т. 24. № 4. С. 18-25.
7. Дневник погоды в населённом пункте Биробиджане на сервисе Gismeteo URL: <https://www.gismeteo.ru/weather-birobidzhan-4860/archive/> (дата обращения: 02.06.2025).