Introduction

For quite some time, I've been intrigued by the algorithms capable of predicting weather forecasts and have been on the lookout for a suitable dataset to delve into this field. However, my search led me to a common hurdle—most resources were either paid or provided insufficient data to work with.

Finally, my persistence paid off when I stumbled upon a website that generously offered an extensive archive of weather data for various cities worldwide, all available in open access. That's the **weather diary for schoolchildren by Gismeteo**.

Gismeteo is one of the most popular CIS weather websites. The site was created on December 12, 2000, however weather forecasts have been published with forecasts from the 1990s on BBS and FIDONet. The site provides an API for professional meteorologists.

Frankly speaking, as a forecast, Gismeteo predictions very often are totally unreliable. But as far as they are one of the pioneers in weather forecasts I guess they at least have skills to reliably take readings from the weather stations and other sources.

Let's have fun scraping it …

Description

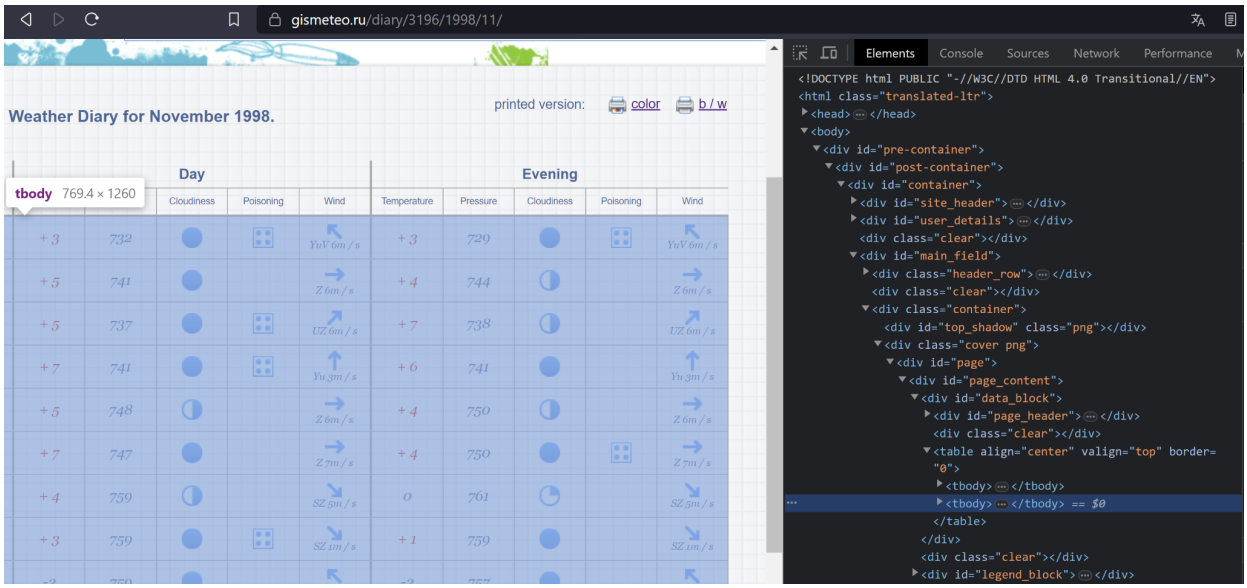
In the picture below there is an example of a weather diary for school children webpage (translated by browser + partially manually).

There is a block of selectors where users can select the country, city, month and year of the diary.

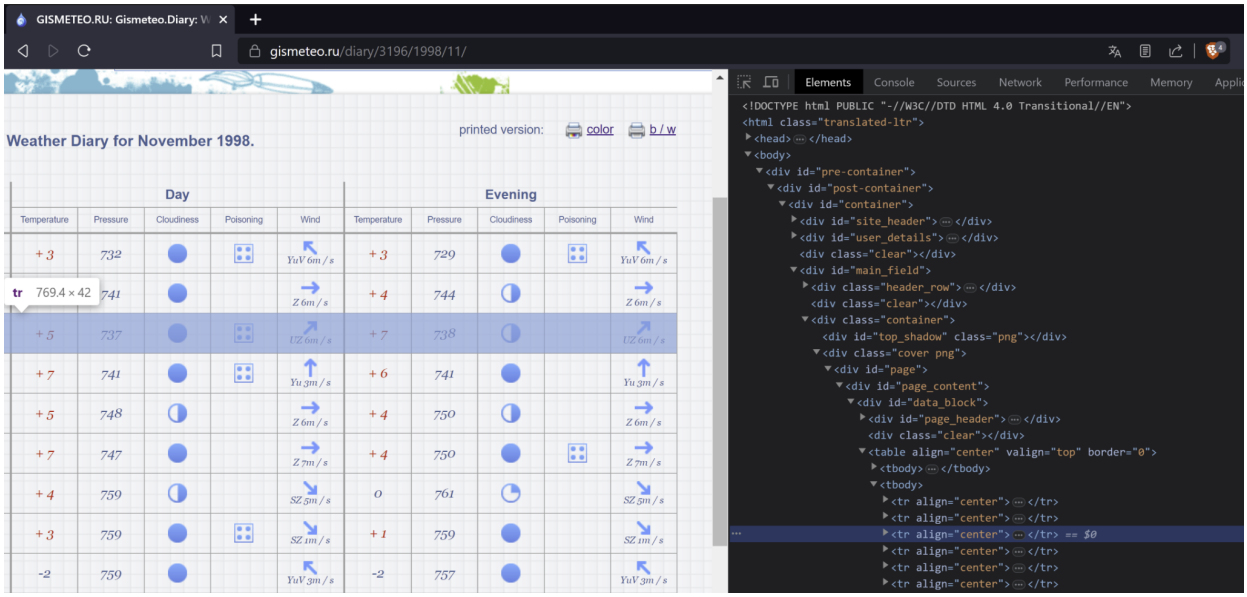
Number	Day					Evening				
	Temperature	Pressure	Cloudiness	Phenomena	Wind	Temperature	Pressure	Cloudiness	Phenomena	Wind
1	+6	746			S 3m/s	+5	741			S 3m/s
2	+4	743			W 7m/s	+1	743			W 7m/s
3	+11	740			NE 3m/s	+7	741			NE 3m/s
4	+10	738			E 5m/s	+7	738			E 5m/s
5	+14	735			SE 6m/s	+7	735			SE 6m/s
6	+9	741			NE 5m/s	0	744			NE 5m/s
7	+9	746			E 6m/s	+6	745			E 6m/s
8	+5	748			S 6m/s	+1	750			S 6m/s
9	+3	749			E 6m/s	+2	749			E 6m/s
10	+5	749			N 1m/s	+2	747			N 1m/s
11	-1	750			NW 8m/s	0	754			NW 8m/s
27	+19	743			SE 6m/s	+13	744			SE 6m/s
28	+15	746			E 6m/s	+8	748			E 6m/s
29	+15	753			E 3m/s	+9	754			E 3m/s
30	+16	755			E 4m/s	+8	756			E 4m/s
31	+17	756			NE 4m/s	+9	756			NE 4m/s

Next there is a table of observations for the selected time period. The data is organized in such a way that every row represents the date and includes the same 5 parameters (Temperature, Pressure, Cloudiness, Weather conditions, Wind) for both day and evening.

Inspection



The data of our interest is located in the **<tbody>** element. Every row of the table is defined by **<tr>** tag.



Every data cell is defined by **<td>** tag.

Day of month, Temperature and Pressure stored as text into **<font>** tag.

Cloudiness and Weather conditions are organized in the form of a gif icon with a unique filename for each factor.

Wind cell contains both icon and text inside for wind direction and wind speed.

```
<tr align="center">
  <td class="first">
    <font style="vertical-align: inherit;">
      <font style="vertical-align: inherit;">3</font>
    </font>
  </td>
  <td class="first_in_group positive">
    <font style="vertical-align: inherit;">
      <font style="vertical-align: inherit;">+ 5</font>
    </font>
  </td>
  <td>
    <font style="vertical-align: inherit;">
      <font style="vertical-align: inherit;">737</font>
    </font>
  </td>
  <td>
    <img src="//st6.gismeteo.ru/static/diary/img/dull.png" class="label_icon label_small screen_icon">
    <img src="//st7.gismeteo.ru/static/diary/img/dull-bw.gif" class="label_icon label_small print_icon">
  </td>
  <td>
    <img src="//st8.gismeteo.ru/static/diary/img/rain.png" class="label_icon label_small screen_icon">
    <img src="//st4.gismeteo.ru/static/diary/img/rain-bw.png" class="label_icon label_small print_icon">
  </td>
  <td>
    <span>
      <img src="//st5.gismeteo.ru/static/diary/img/w5.gif" class="screen_icon">
      <img src="//st6.gismeteo.ru/static/diary/img/w5-bw.gif" class="print_icon">
      <br>
      <font style="vertical-align: inherit;">
        <font style="vertical-align: inherit;">UZ 6m / s</font>
      </font>
    </span>
  </td>
```

The url for every single page (selected month/year) looks like: gismeteo.ru/diary/**{cityID}**/**{year}**/**{month}**

Scraper mechanics

Users define 3 parameters: **cityID**, **start date** and **finish date.**

Scraper creates a list of urls for every month between start and finish date: gismeteo.ru/diary/**{cityID}**/**{year}**/**{month}**.

Then scraper grabs the table cells line by line and appends it to the dataframe.

The first cell **Day** scraper converts into a **Date** adding {month} and {year} parameters.

Next two cells **Temperature** and **Pressure** parsed as text.

Cloudiness and **Weather conditions** (like snow, rain, storm, hail) are recognized by the uniqueness of the .gif icon name and stored in the dataframe as factors.

Last cell **Wind** when parsing is splitted into 2 cells: **Wind direction** and **Wind speed**.

When all the table is parsed scraper continues with the next url from the list and appends new data to the dataframe.

Scraper stops when all the links between start and finish dates are parsed and saves dataframe to .csv file.

Scrapers output

Date	D Temperature	D Pressure	D Cloudiness	D Weather Condition	D Wind Direction	D Wind Speed	N Temperature	N Pressure	N Cloudiness	N Weather Condition	N Wind Direction	N Wind Speed
1.4.1997	+15	753	Sun		З	8м/с	+8	753	Sun		З	8м/с
2.4.1997	+11	752	Dull		СЗ	5м/с	+7	752	Sun		СЗ	5м/с
3.4.1997	+13	743	Sun		ЮЗ	6м/с	+8	738	Mostly Clouds		ЮЗ	6м/с
4.4.1997	+4	743	Mostly Clouds		СЗ	9м/с	+1	745	Sun/Clouds		СЗ	9м/с
5.4.1997	+5	744	Mostly Clouds		СЗ	5м/с	0	741	Dull	Snow	СЗ	5м/с
6.4.1997	+2	748	Dull		С	7м/с	0	753	Sun		С	7м/с
7.4.1997	+1	760	Mostly Clouds		С	5м/с	-1	763	Sun		С	5м/с
8.4.1997	+5	763	Mostly Clouds		СЗ	6м/с	+2	762	Mostly Clouds		СЗ	6м/с
9.4.1997	+5	761	Dull		С	2м/с	+4	760	Dull		С	2м/с
10.4.1997	+9	752	Dull		СЗ	6м/с	+5	749	Dull	Rain	СЗ	6м/с
11.4.1997	+4	731	Mostly Clouds		СЗ	17м/с	+1	738	Dull		СЗ	17м/с
12.4.1997	0	747	Dull		С	7м/с	-1	751	Dull	Snow	С	7м/с
13.4.1997	+3	756	Mostly Clouds		СЗ	5м/с	0	756	Sun		СЗ	5м/с
14.4.1997	+4	745	Dull	Rain	ЮЗ	6м/с	+5	742	Dull	Rain	ЮЗ	6м/с
15.4.1997	+2	747	Mostly Clouds		СЗ	8м/с	-1	750	Sun/Clouds		СЗ	8м/с
16.4.1997	+1	750	Dull	Rain	С	4м/с	+1	750	Dull		С	4м/с
17.4.1997	+11	746	Mostly Clouds		С	8м/с	+5	746	Dull		С	8м/с
18.4.1997	+7	744	Dull		СЗ	5м/с	+3	743	Dull	Rain	СЗ	5м/с
19.4.1997	+5	747	Mostly Clouds		СЗ	6м/с	0	749	Sun		СЗ	6м/с
20.4.1997	+2	750	Mostly Clouds		СВ	3м/с	0	751	Mostly Clouds		СВ	3м/с
21.4.1997	+5	750	Mostly Clouds		С	2м/с	-1	750	Sun/Clouds		С	2м/с
22.4.1997	+7	747	Mostly Clouds		С	2м/с	+1	747	Sun		С	2м/с
23.4.1997	+7	752	Mostly Clouds		СЗ	8м/с	+2	754	Sun		СЗ	8м/с
24.4.1997	+14	752	Mostly Clouds		ЮЗ	9м/с	+9	750	Sun		ЮЗ	9м/с
25.4.1997	+9	749	Dull	Rain	З	4м/с	+5	751	Sun/Clouds		З	4м/с
26.4.1997	+11	753	Sun		СВ	3м/с	+5	754	Sun		СВ	3м/с
27.4.1997	+15	751	Sun/Clouds		ЮВ	6м/с	+10	752	Sun		ЮВ	6м/с
28.4.1997	+18	748	Mostly Clouds		ЮЗ	7м/с	+13	747	Mostly Clouds		ЮЗ	7м/с
29.4.1997	+16	745	Dull		ЮЗ	5м/с	+10	746	Mostly Clouds		ЮЗ	5м/с
30.4.1997	+17	750	Sun/Clouds		З	5м/с	+11	753	Sun/Clouds		З	5м/с
1.5.1997	+13	758	Mostly Clouds		СЗ	7м/с	+10	758	Dull		СЗ	7м/с
2.5.1997	+17	753	Mostly Clouds		З	9м/с	+14	750	Mostly Clouds		З	9м/с
3.5.1997	+16	746	Dull		СЗ	10м/с	+8	747	Mostly Clouds		СЗ	10м/с
4.5.1997	+16	745	Sun/Clouds		З	6м/с	+8	744	Sun		З	6м/с
5.5.1997	+24	735	Mostly Clouds		ЮЗ	5м/с	+16	734	Mostly Clouds		ЮЗ	5м/с
6.5.1997	+23	735	Mostly Clouds		ЮЗ	6м/с	+17	733	Sun/Clouds		ЮЗ	6м/с
7.5.1997	+12	744	Mostly Clouds		З	5м/с	+7	746	Sun/Clouds		З	5м/с
8.5.1997	+9	741	Dull	Rain	СВ	7м/с	+12	738	Dull		СВ	7м/с
9.5.1997	+13	744	Mostly Clouds		З	7м/с	+8	746	Sun		З	7м/с
10.5.1997	+17	752	Mostly Clouds		ЮЗ	5м/с	+10	753	Sun		ЮЗ	5м/с
11.5.1997	+22	756	Sun/Clouds		ЮЗ	4м/с	+15	756	Sun		ЮЗ	4м/с
12.5.1997	+24	756	Mostly Clouds		ЮЗ	3м/с	+17	756	Sun		ЮЗ	3м/с
13.5.1997	+19	757	Dull		С	4м/с	+14	758	Sun		С	4м/с
14.5.1997	+26	756	Sun/Clouds		ЮЗ	3м/с	+19	756	Sun/Clouds		ЮЗ	3м/с
17.5.1997	+21	754	Sun		В	4м/с	+14	755	Sun/Clouds		В	4м/с
18.5.1997	+23	749	Sun/Clouds		С	2м/с	+16	748	Sun		С	2м/с
19.5.1997	+20	743	Dull	Rain	С	2м/с	+15	744	Dull		С	2м/с
20.5.1997	+17	744	Mostly Clouds		С	2м/с	+12	744	Mostly Clouds		С	2м/с
21.5.1997	+15	743	Dull	Rain	З	4м/с	+13	744	Sun		З	4м/с
22.5.1997	+15	745	Dull		С	4м/с	+10	748	Dull		С	4м/с
23.5.1997	+11	753	Mostly Clouds		С	2м/с	+8	753	Dull		С	2м/с
24.5.1997	+12	756	Mostly Clouds		С	2м/с	+8	757	Mostly Clouds		С	2м/с
25.5.1997	+14	761	Mostly Clouds		СЗ	5м/с	+7	761	Sun		СЗ	5м/с
26.5.1997	+15	756	Mostly Clouds		С	2м/с	+9	753	Mostly Clouds		С	2м/с
27.5.1997	+11	749	Dull		С	2м/с	+8	748	Dull	Rain	С	2м/с
28.5.1997	+10	752	Dull	Storm	СЗ	2м/с	+9	754	Dull		СЗ	2м/с
29.5.1997	+14	753	Dull		З	5м/с	+8	752	Dull	Rain	З	5м/с
30.5.1997	+9	752	Dull		СЗ	6м/с	+7	754	Dull	Rain	СЗ	6м/с
31.5.1997	+13	753	Dull	Storm	З	4м/с	+9	753	Dull		З	4м/с
1.6.1997	+12	753	Dull	Rain	СЗ	3м/с	+10	752	Dull	Rain	СЗ	3м/с
2.6.1997	+15	750	Mostly Clouds		С	4м/с	+11	750	Mostly Clouds		С	4м/с
3.6.1997	+13	750	Dull		З	5м/с	+10	750	Mostly Clouds		З	5м/с
4.6.1997	+18	753	Mostly Clouds		З	3м/с	+13	754	Sun		З	3м/с
5.6.1997	+21	753	Mostly Clouds		No wind	No wind	+14	753	Sun/Clouds		No wind	No wind
6.6.1997	+22	753	Sun/Clouds		В	4м/с	+17	754	Sun		В	4м/с
7.6.1997	+22	755	Sun/Clouds		В	5м/с	+16	756	Sun/Clouds		В	5м/с
8.6.1997	+17	756	Dull	Storm	С	2м/с	+15	755	Sun/Clouds		С	2м/с
9.6.1997	+17	755	Dull	Storm	З	3м/с	+16	756	Mostly Clouds		З	3м/с
10.6.1997	+19	755	Mostly Clouds		СЗ	5м/с	+14	755	Sun		СЗ	5м/с
11.6.1997	+21	751	Sun/Clouds		С	4м/с	+16	751	Sun		С	4м/с
12.6.1997	+24	747	Sun/Clouds		С	2м/с	+17	747	Sun/Clouds		С	2м/с
13.6.1997	+25	744	Mostly Clouds		ЮЗ	3м/с	+18	745	Mostly Clouds		ЮЗ	3м/с
14.6.1997	+16	744	Dull	Storm	З	5м/с	+16	744	Mostly Clouds	Storm	З	5м/с
15.6.1997	+19	748	Dull	Storm	З	7м/с	+15	750	Mostly Clouds		З	7м/с

Scrapers performance

The best results in terms of program speed are observed in Scrapy-based solutions - this is due to the fact that the framework works asynchronously, simultaneously processing several links.

The second fastest solution showed BeautifulSoup - all links are processed sequentially, but very quickly individually - only one http request is required to analyze data for one month.

The worst solution for such a task is performed by Selenium. The low speed of work is due to the fact that the solution completely emulates a browser, which means it does a lot of unnecessary work: it loads additional css, js files; execute javascript code; sends additional http requests (for example, to the yadro.ru site, which is connected to gismeteo.ru)

Total running time for the whole dataset without page limits (using VPN):

Beautiful Soup: 1307 seconds

Scrapy: 33.5 seconds

Selenium: 6054 seconds

Data analysis

* for the whole period from 1997 to 2023