LOW-LEVEL PIPELINE FOR EXTRACTING SCALAR DATA VALUES FROM WEATHER FORECAST SITES

Yaroslav Kravets, Iryna Liutenko
National Technical University «Kharkiv Polytechnic Institute», Kharkiv

The quality of weather forecasts is extremely important for the national economy and for ordinary citizens. Depending on the forecasts, management decisions are made, starting from "what to wear" and ending with "whether to start a military campaign". The authors proposed to use a multi-agent system [1-2] for processing/unification of weather data from WEB-sites. This article reveals a low-level approach to retrieving data from hierarchical formats such as HTML and Json. These formats were chosen because data on websites is most often presented in them.

When extracting weather data, simple data include date/time, humidity, temperature, wind direction and wind speed, pressure, precipitation, absolute/relative humidity, etc. In general, at a low level, the extraction of values for each of the weather parameters occurs as follows:

In the first step, data is obtained from a hierarchy element. The element can be different, depending on the format, for Json it is a Json Element, for HTML it is a HTML attribute or a HTML element. To select the desired hierarchy element for the HTML format, the XPATH language is used. For the JSON format, the JSON PATH query language is used, respectively. Regardless of the format of the processed file, the result of this step is the raw text data.

In the second step, post-processing of raw data from the previous step occurs using some regular expressions. This is necessary in order to extract valuable data from a raw text fragment. This step is optional, because in some cases the text itself contains only useful information and does not require such additional processing.

In the third, final step, the conversion of valuable data from the previous steps to the appropriate types is performed: for temperature, wind speed, humidity – the destination type is a floating-point number, for wind direction the azimuth angle is used, therefore, the conversion to floating point is also performed, for date/time – conversion to C# type DateTime. For some values, a simple conversion from string type to number occurs, but for wind direction and settlement dictionaries are used. It is needed, for example, to convert a verbal value such as "S.S." or "Kyiv" to a numeric or enum, respectively.

The proposed approach is a pipeline for extracting data from Json and HTML formats. In our study, it is used to obtain weather data. But it can also be used for mining other data that can be found on WEB pages. Also, due to the nature of the format, this approach will be relevant when working with XML files.

References:

1. Karina Melnyk, Yaroslav Kravets, Iryna Liutenko, Svitlana Yershova, Oksana Ivashchenko, Dmytro Yershov and Olena Odyntsova. Multi-Agent Approach for the Unification of Meteorological Data. *COLINS 2023*, URL: https://ceur-ws.org/Vol-3403/paper37.pdf

2. Liutenko I., Kravets Y. Principles of polite crawling for collecting weather data from websites. *Odesa 2024*, URL: https://repository.kpi.kharkov.ua/items/b6c46d23-48c1-400d-a6b9-87577232bcbd