

TypeScript を用いたスクレイピングとデータ整形について

5435 藤本 祥太 (指導教員：才田 聡子)

Scraping and Data Formatting Using TypeScript

概要：ここに概要を書く。

キーワード スクレイピング, TypeScript

1. はじめに

電離圏研究のために実施されている短波ドップラー (HFD) 観測のデータはオープンデータとして公開されており、自由な使用が認められている。現状、データ活用を促進するために数件の web アプリケーションが開発されているが、多くが開発運用の困難性やユーザーエクスペリエンス等の観点で問題を抱えている。この課題点を解消するため既存のアプリケーションの課題点から必要要件を洗い出し、ページレンダリング手法を考慮した web アプリケーションの新規設計を行う。本研究ではその中の web スクレイピングを用いてデータを取得する処理、および、取得したデータを整形して扱いやすくする処理に焦点を当て研究を行う。本研究では Next.js と呼ばれるフレームワークを用いて実装を行う。JavaScript の UI ライブラリの一つである React.js がベースとなっており、サーバ機能も有しているため効率的な web 開発が可能である。また、使用する言語を統一して管理コストの低減を図るため、web スクレイピングは TypeScript を用いて行う。

2. Web スクレイピングについて

web スクレイピングとは、web サイトから情報を抽出するコンピュータソフトウェア技術のことを指す。HTML フォーマットからテキストを抽出してスプレッドシートや json ファイル等の構造化データへの変換を行うことができ、web 上にあるデータを取得して扱うことを目的とする。web スクレイピングの手法としては、正規表現や DOM 解析、HTML パーサ等を用いたものがある。本研究では、ISR を用いた web アプリケーションを設計するため、スクレイピングの処理に速度を求める必要がない。そのため、本研究では軽量でサーバへの負担が小さい cheerio を HTML パーサとして用いる。

3. 使用ライブラリ

本研究で使用した主な TypeScript ライブラリについて以下に示す。

3.1 Superaagent

HTML リクエストを送信することのできるモジュール。本研究では HFD 観測データが公開されている web ページから HTML レスポンスとしてテキストデータを取得する際に用いる。²⁾

3.2 Cheerio

HTML パーサに jQuery のサブセットを実装したもの。jQuery の関数を用いてスクレイピングの処理を行うことができる。本研究では HTML レスポンスからテキストデータだけを抜き出す際に用いる。³⁾

4. 実装

4.1 目標

本開発では、Next.js を用いて開発運用及びユーザーエクスペリエンスの観点で従来のものよりも優れた Web アプリケーションを作成することを目標とする。その中でも、本研究ではスクレイピングを用いてデータを取得し、それをフロントエンドで扱いやすいように整形する処理を行うことまでを目標とする。

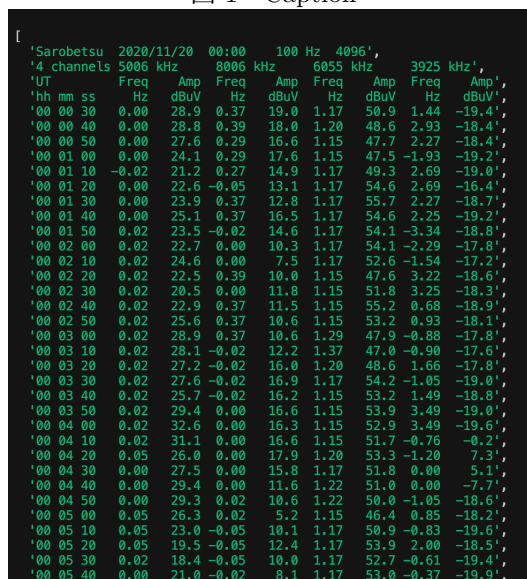
4.2 スクレイピング処理

Superaagent を用いて、HF ドップラーのデータが公開されている web サイト⁴⁾ に HTML リクエストを送る。返ってきた HTML レスポンスから cheerio を用いて body を抜き出し、テキストデータに変換する。

4.3 データ整形処理

前項で取得したテキストデータはそのままでは扱いにくいいため、扱いやすい形に整形する。例として、2020/11/20Sarobetsu というデータを示す。

図 1 Caption



	2020/11/20	00:00	100 Hz	4096'
'Sarobetsu	2020/11/20	00:00	100 Hz	4096'
'4 channels	5006 kHz	8006 kHz	6055 kHz	3925 kHz'
'UT	Freq	Amp	Freq	Amp
'hh mm ss	Hz	dBuV	Hz	dBuV
'00 00 30	0.00	28.9	0.37	19.0
'00 00 40	0.00	28.8	0.39	18.0
'00 00 50	0.00	27.6	0.29	16.6
'00 01 00	0.00	24.1	0.29	17.6
'00 01 10	-0.02	21.2	0.27	14.9
'00 01 20	0.00	22.6	-0.05	13.1
'00 01 30	0.00	23.9	0.37	12.8
'00 01 40	0.00	25.1	0.37	16.5
'00 01 50	0.02	23.5	-0.02	14.6
'00 02 00	0.02	22.7	0.00	10.3
'00 02 10	0.02	24.6	0.00	7.5
'00 02 20	0.02	22.5	0.39	10.0
'00 02 30	0.02	20.5	0.00	11.8
'00 02 40	0.02	22.9	0.37	11.5
'00 02 50	0.02	25.6	0.37	10.6
'00 03 00	0.02	28.9	0.37	10.6
'00 03 10	0.02	28.1	-0.02	12.2
'00 03 20	0.02	27.2	-0.02	16.0
'00 03 30	0.02	27.6	-0.02	16.9
'00 03 40	0.02	25.7	-0.02	16.2
'00 03 50	0.02	29.4	0.00	16.6
'00 04 00	0.02	32.6	0.00	16.3
'00 04 10	0.02	31.1	0.00	16.6
'00 04 20	0.05	26.0	0.00	17.9
'00 04 30	0.00	27.5	0.00	15.8
'00 04 40	0.00	29.4	0.00	11.6
'00 04 50	0.00	29.3	0.02	10.6
'00 05 00	0.05	26.3	0.02	5.2
'00 05 10	0.05	23.0	-0.05	10.1
'00 05 20	0.05	19.5	-0.05	12.4
'00 05 30	0.02	18.4	-0.05	10.0
'00 05 40	0.00	21.0	-0.02	8.1

4.4 完成品

完成したサイトがこちらになります。

参考文献

- 1) 中嶋 柊, HF ドップラー観測データの利活用を目的とする web アプリケーションのフロントエンド設計, 令和4年度 制御情報システム創造演習 報告書
- 2) Superagent-npm <https://www.npmjs.com/package/superagent> 2023/01/25
- 3) cheeriojs/cheerio: Fast, flexible, and lean implementation of core jQuery designed specifically for the server. <https://github.com/cheeriojs/cheerio> 2023/01/25
- 4) HF Doppler Sounding Experiment in Japan - HFDOPE <http://gwave.cei.uec.ac.jp/~hfd/pre.html> 2024/