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： 電離圏研究のために実施されている 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，この観測データの利活用のためにいくつかの web アプリケーションが開発されているが，その多くで課題点を抱えてしまっている．そのため，主な利用者である研究者たちの要望も取り入れつつ，よりよい web アプリケーションの提案，開発する．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扱いやすくする処理に焦点を当てる．

キーワード　スクレイピング，TypeScript

1. はじめに

地震等に伴う電離圏の変動を観測するため，HF 帯電波を用いた HFD 観測という手法が取られている．この HFD(短波ドップラー) 観測のデータは電離圏研究のために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．[1)]

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．しかし，設計時点で複数の言語が使用されている，ページの読み込みに時間がかかるなど，その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．また，web アプリケーションの主な使用者となる電離圏の研究者から，いくつかの要望が出されている．それらをいくつか抜粋して，以下表 1 に示す．

表 1　web アプリケーションに対する要望の抜粋

- 周波数固定で全ての局を見せる
- 局固定で全ての周波数を見せる
- 局，周波数を選んでカスタムプロットしたい
- ダイナミックスペクトルを見たい
- PDF などの形に出力できると良い
- パンとズームができると良い

既存の web アプリケーションにおける課題点の解消，及び研究者からの要望を反映した web アプリケーションを提供するため，web アプリケーションの新規設計を行う．既存の web アプリケーションにおける課題点を解消するため，ページレンダリング手法を考慮して開発を進めていく．本研究では Next.js と呼ばれる JavaScript の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，サーバーサイドレンダリング (SSR) や静的サイト生成 (SSG) といった機能を持つ web サイトの作成を得意とする．そのため，今回開発する web アプリケーションの要件に合致していると判断した．[2)]

本研究ではこの web アプリケーションにおける web スクレイピングを用いてオープンデータが公開されている web サイトから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フロントエンドで扱いやすく整形する処理に焦点を当てる．また，使用する言語を web アプリケーション開発全体で統一して管理コストの低減を図るため，web スクレイピングに関する処理は 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 TypeScritp ライブラリについて以下に示す．

3.1 Superagent

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．[4)]

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．[5)]

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

Superagent 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 [7] 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ため，扱いやすい形に整形する．例として，2020/11/20Sarobetsu のデータを示す．

図 1　Caption

```
[
  'Sarobetsu  2020/11/20  00:00    100 Hz  4096',
  '4 channels 5006 kHz     8006 kHz     6055 kHz     3925 kHz',
  'UT         Freq    Amp  Freq    Amp  Freq    Amp  Freq    Amp',
  'hh mm ss     Hz   dBuV    Hz   dBuV    Hz   dBuV    Hz   dBuV',
  '00 00 30   0.00   28.9  0.37   19.0  1.17   50.9  1.44  -19.4',
  '00 00 40   0.00   28.8  0.39   18.0  1.20   48.6  2.93  -18.4',
  '00 00 50   0.00   27.6  0.29   16.6  1.15   47.7  2.27  -18.4',
  '00 01 00   0.00   24.1  0.29   17.6  1.15   47.5 -1.93  -19.2',
  '00 01 10  -0.02   21.2  0.27   14.9  1.17   49.3  2.69  -19.0',
  '00 01 20   0.00   22.6 -0.05   13.1  1.17   54.6  2.69  -16.4',
  '00 01 30   0.00   23.9  0.37   12.8  1.17   55.7  2.27  -18.7',
  '00 01 40   0.00   25.1  0.37   16.5  1.17   54.6  2.25  -19.2',
  '00 01 50   0.02   23.5 -0.02   14.6  1.17   54.1 -3.34  -18.8',
  '00 02 00   0.02   22.7  0.00   10.3  1.17   54.1 -2.29  -17.8',
  '00 02 10   0.02   24.6  0.00    7.5  1.17   52.6 -1.54  -17.2',
  '00 02 20   0.02   22.5  0.39   10.0  1.15   47.6  3.22  -18.6',
  '00 02 30   0.02   20.5  0.00   11.8  1.15   51.8  3.25  -18.3',
  '00 02 40   0.02   22.9  0.37   11.5  1.15   55.2  0.68  -18.9',
  '00 02 50   0.02   25.6  0.37   10.6  1.15   53.2  0.93  -18.1',
  '00 03 00   0.02   28.9  0.37   10.6  1.29   47.9 -0.88  -17.8',
  '00 03 10   0.02   28.1 -0.02   12.2  1.37   47.0 -0.90  -17.6',
  '00 03 20   0.02   27.2 -0.02   16.0  1.20   48.6  1.66  -17.8',
  '00 03 30   0.02   27.6 -0.02   16.9  1.17   54.2 -1.05  -19.0',
  '00 03 40   0.02   25.7 -0.02   16.2  1.15   53.2  1.49  -18.8',
  '00 03 50   0.02   29.4  0.00   16.6  1.15   53.9  3.49  -19.0',
  '00 04 00   0.02   32.6  0.00   16.3  1.15   52.9  3.49  -19.6',
  '00 04 10   0.02   31.1  0.00   16.6  1.15   51.7 -0.76   -0.2',
  '00 04 20   0.05   26.0  0.00   17.9  1.20   53.3 -1.20    7.3',
  '00 04 30   0.00   27.5  0.00   15.8  1.17   51.8  0.00    5.1',
  '00 04 40   0.00   29.4  0.00   11.6  1.22   51.0  0.00   -7.7',
  '00 04 50   0.00   29.3  0.02   10.6  1.22   50.0 -1.05  -18.6',
  '00 05 00   0.05   26.3  0.02    5.2  1.15   46.4  0.85  -18.2',
  '00 05 10   0.05   23.0 -0.05   10.1  1.17   50.9 -0.83  -19.6',
  '00 05 20   0.05   19.5 -0.05   12.4  1.17   53.9  2.00  -18.5',
  '00 05 30   0.02   18.4 -0.05   10.0  1.17   52.7 -0.61  -19.4',
  '00 05 40   0.00   21.0 -0.02    8.1  1.17   53.0 -0.37  -19.9',
```

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) 吉川晃平，HF ドップラーにより観測された地震に伴う電離圏変動の中性待機波導数値シミュレーションによる定量的評価，電気学会論文誌 A Vol.136 No.5 pp.259 264
2) Next.js by Vercel - The React Framework `https://nextjs.org/`　2024/1/13
3) 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
4) Superagent-npm `https://www.npmjs.com/package/superagent` 2023/01/25
5) cheeriojs/cheerio: Fast, flexible, and lean implementation of core jQuery designed specifically for the server. `https://github.com/cheeriojs/cheerio` 2023/01/25
6) HF Doppler Sounding Experiment in Japan - HFDOPE `http://gwave.cei.uec.ac.jp/~hfd/pre.html`　2024/13