進捗報告

清水　豪士

Department of Information Systems Engineering,
Graduate School of Engineering
u155016@st.pu-toyama.ac.jp

12:10-12:35 Friday, December 10, 2021,
Toyama Prefectural University.

近況

前回のおさらい

- エクセルファイルから授業計画のデータを取得できるようにした.
- 先生はバラバラだが，授業内容が重複しているものは 1 つにした.
- 看護学部だけのものや，大学院の授業は除外した.

前回の課題

- Selenium だけでなく，bs4 に対応できるようにする.
- 通販サイトなどの EC サイトを適切に処理する.
- 授業計画のいろんな書き方に対応する.

進捗

bs4 での対応

- Selenium と同様な処理を bs4 でできるようにした.
- これで，Selenium でエラーが起きても対応可能になった.

Selenium と bs4

- Selenium では，ページ数でスクレイピングができ，bs4 ではページ数ではなく検索結果の上位何件を自分で指定して取得できる.
- どちらでも取得できる数の制限をかけることで同様の結果となる.

進捗

EC サイトについて

- 一応 Amazon や楽天などの有名どころの EC サイトは除去したが，根本的な解決はできていない．
- また，調べる授業計画の内容によっては，検索結果の 8 割が EC サイトとかになってしまう．

対策の考え

- どの EC サイトにも「購入」ボタンがあると思うので，それらを発見したら，その URL をスクレイピングしないみたいな感じにする．
- 他の意見があれば提案してくれると幸いです．

まとめ

まとめ

- Selenium だけでなく bs4 でも同様の処理を行えるようにした.
- エグゼクティブサマリーを作成した.

課題

- 授業計画のいろんな書き方に対応する.
- EC サイトの選別
- プログラムの理解 + 論文を読む