報道各位

2020年9月7日
株式会社インプレスR&D
https://nextpublishing.jp/

Webスクレイピングで楽々データ収集！

『スクレイピング・ハッキング・ラボ　Pythonで自動化する未来型生活』発行

技術の泉シリーズ、9月の新刊

インプレスグループで電子出版事業を手がける株式会社インプレス R&D は、『スクレイピング・ハッキング・ラボ Pythonで自動化する未来型生活』(著者:齊藤 貴義)を発行いたしました。

最新の知見を発信する『技術の泉シリーズ』は、「技術書典」や「技術書同人誌博覧会」をはじめとした各種即売会や、勉強会・LT 会などで頒布された技術同人誌を底本とした商業書籍を刊行し、技術同人誌の普及と発展に貢献することを目指します。

『スクレイピング・ハッキング・ラボ　Pythonで自動化する未来型生活』
https://nextpublishing.jp/isbn/9784844378860

著者:齊藤 貴義
小売希望価格:電子書籍版　1800円(税別)／印刷書籍版　2200円(税別)
電子書籍版フォーマット:EPUB3／Kindle Format8
印刷書籍版仕様:B5判／モノクロ／本文268ページ
ISBN:978-4-8443-7886-0
発行:インプレスR&D

＜＜発行主旨・内容紹介＞＞

本書ではPythonを使ったWebスクレイピングテクニックについて解説します。いろいろなことをPythonで自動化していきましょう。日本の主要なサービスを題材に、スクレイピングでデータを取得する方法と、そのデータを元に分析や可視化していく手法を紹介していきます。

スクレイピング環境の構築、スクレイピングを行うにあたって便利なライブラリの選定、ターゲットとなるWebサービスの選定、データ分析の手法など、初心者にもわかりやすく解説しています。

(本書は、次世代出版メソッド「NextPublishing」を使用し、出版されています。)

スクレイピングの基礎と Python の導入を解説

あるいは、Webサービスの RSS（Rich Site Summary）フィードを、RSS リーダーで読み取って新着情報を得る手段もあります。現在、サービスを提供している代表的な Web ベースの RSS リーダーには、Feedly[3]や Inoreader[4]などがあります。しかし、SNS の増加や Google Reader のサービス終了などで、RSS の役目は歴史的役割を終えたという認識も一部で広まっています。あらゆる情報を RSS フィードで生成・取得する文化は、徐々に衰退しつつあります。大手 Web サービスでも、最近は RSS フィードを生成していないサービスが増えています。また、RSS フィードの内容は提供側の意図で決められているため、必ずしも意図どおりの情報を得られるとは限りません。

それ以外では、最近ではスマホアプリのプッシュ通知や、ブラウザーの Web Push などで更新情報を得る手段もあります。近年では、多くのサービスがプッシュ通知に対応しています。しかし、この手段も利用者が意図どおりの結果を得られるとは限りません。また、Web Push は Chrome など主要なブラウザーに限られます。加えて、通知するタイミングや詳細な購読管理ができない、Web Push 自体がブラウザーにとってノイズになっている[5]、などの問題点があります。

1.3　スクレイピングの利点

スクレイピングは、自分で記述したプログラムを実行して、Web サービスから必要な情報を抽出します。スクレイピングでは、基本的に Web ページとして提供されているデータを自由に取得して、加工・整理することができます。また、実行したいタイミングで動かすことができます。スクレイピングは、RSS フィードやプッシュ通知などで見られた、サービス提供側の制約をあまり受けません[6]。

また、近年はスクレイピング関連のライブラリーが整備されてきており、それらのライブラリーを組み込めば、簡単なプログラムでスクレイピングを実行できる環境が実現できます。プログラミングについて専門知識を持たない人でも、本書に書かれているようなスクレイピングを実行して、自由に情報を得ることができます。スクレイピングは、Web の自由や民主化を補完して拡張するものといえます。

1.4　スクレイピングの問題点

スクレイピングを実行する際には、法律に注意しなければなりません。スクレイピングやクローリングの手法は、Google を始め多くの企業や公共機関で採用されています。ただし、スクレイピングの実装内容や対象サービスによっては、著作権法や不正アクセス禁止法の違反、あるいは偽計業務妨害に問われる可能性があります。

岡崎市立中央図書館事件（Librahack 事件）

2010 年 3 月に、愛知県岡崎市の岡崎市立図書館で、蔵書検索システムに障害が発生しました。そして、岡崎市立図書館の蔵書検索システムからクローラーで自動的に情報を収集していた男性が、愛知県警によって偽計業務妨害容疑で逮捕されました。この事件は「岡崎市立中央図書館事件」、または男性が Librahack というサイト[7]を開設したことから、「Librahack 事件」と呼ばれています。

男性は勾留と取り調べの後に起訴猶予処分となりました。ですが、クローラーは一定間隔空けてからアクセスするように実装されており、これが偽計業務妨害に該当するかは、多くの専門家が疑問視しています。ただ、近年の WizardBible 事件や Coinhive 事件、アラートループ事件などを見ても、警察は IT に関して無知であることが多いようです。そのため、実際には法律に抵触しなくても、家宅捜索や起訴をされるリスクはあります。私達は慎重に行動しなければなりません[8]。

本書では Web サイトに設置された robots.txt を解析して、それをスクイピングに組み込んでいく手法も紹介しています。

1.5　スクレイピングに Python を使用する

スクレイピングは様々なプログラミング言語で実装できますが、本書では Python3 を使ったスクレイピング技法について解説していきます。

Python を使うメリット

Python は、世界中で多数の開発者が使用している言語です。1991 年に登場したこの言語は、オープンソースのスクリプト言語として開発され、現在では Google、Dropbox、Instagram など、著名なサービスでも使われています。

シンプルでわかりやすい文法と豊富なライブラリーが Python の特徴で、初心者にも学びやすいコンピューター言語です。数多くのライブラリーを pip コマンドで容易に組み込むことができ、機械学習や深層学習などの人工知能研究でも、Python が多数使われています。

3.Feedly https://feedly.com/
4.Inoreader https://www.inoreader.com/
5.mizchi 氏「自社の懸念どおり、ブラウザーのプッシュ通知は悪質に使われはじめている。実装側はクリックまで購読確認を待つべき。」https://mizchi.hatenablog.com/entry/2017/12/05/223520
6.RSS から情報を抽出することも、広義のスクレイピングに該当しますが、本書では主に、Web ページを自由に操作するスクレイピングのプログラムを扱います。
7.Librahack http://librahack.jp/
8.この問題を考えるにあたって、詳細は自著「Web セキュリティーのミライ　不正指令電磁的記録問題に寄せて」を参照ください。

スクレイピングのテクニックや考慮すべき点を解説

スマートフォンのユーザーエージェントでアクセスする

モバイル限定の Web サイト、またはモバイルで表示が変わる Web サイトからスクレイピングしたい場合は、たとえば iPhone7 Plus の場合は、以下のようにユーザーエージェントを指定すると良いでしょう。

リスト6.38:
```
user_agent = "Mozilla/5.0 (iPhone; CPU iPhone OS 12_4_1 like Mac OS X)
AppleWebKit/605.1.15 (KHTML, like Gecko) Version/12.1.2 Mobile/15E148
Safari/604.1"
```

Android（Huawei P20 Pro）の場合は、一例として以下のように記述します。

リスト6.38:
```
user_agent = "Mozilla/5.0 (Linux; Android 8.1.0; HW-01K) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/77.0.3865.73 Mobile Safari/537.36"
```

6.4　リファラを設定する

リファラはアクセス元の URL を指します。Web サービスには、リファラによって挙動を変えたりアクセスを制限しているページが存在します。ユーザーエージェント情報と同じように、リファラもヘッダー情報として定義できます。

リスト6.38:
```
refer = "https://www.yahoo.co.jp/"
user_agent = "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/77.0.3864.0 Safari/537.36"
headers = {
    "refer" : refer,
    "User-Agent": user_agent
}
```

6.5　文字列内の特殊文字をエスケープする

Web の世界では、どのような文字が使われるかわかりません。中には、HTML として保存すると、潜在的に危険性のある文字列も存在します。文字列中の &、<、>、"、' を HTML セーフなシーケンスに変換するには、以下のようにすると良いでしょう。

リスト6.38:
```
import html
data = html.escape("""  Hello "World" """)
print(data)

# Hello &amp;quot;World&amp;quot;
```

エスケープされた文字列をアンエスケープもできます。

リスト6.38:
```
import html
data = html.unescape("""  Hello &amp;quot;World&amp;quot; """)
print(data)

# Hello "World"
```

6.6　HTTP ステータスコード

接続先の Web サイトに正常接続できたのかを確認するために、HTTP ステータスコードを取得するならば、requests ライブラリーでは、status_code を使ってステータスコードを取得できます。今回は、アニメ「SteinsGate」の公式サイトの HTTP ステータスコードを取得してみます。

http://steinsgate.tv

図6.3:

各種フレームワークを使った実際のスクレイピングの活用法を紹介

図 10.3:

わずか4行でfor文を実行することなく、上場企業の平均年収のデータを取得することができました。

Pandasでは、この結果を簡単にCSVで保存することもできます。

リスト10.20:

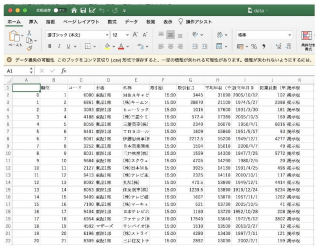
```
import pandas as pd

url = "https://info.finance.yahoo.co.jp/ranking/?kd=45"

df = pd.read_html(url)
df[0].to_csv("data.csv", encoding="SHIFT-JIS")
```

data.csvを開くと、以下のような内容になっており、出力に成功しています。

図 10.4:



今まで、スクレイピングの結果のファイル出力はCSVを中心に解説してきましたが、PandasのDataFeame型ならば、Excel形式（xlsxファイル）にも簡単に出力できます。まず、openpyxlライブラリーをインストールします。

```
$ pip3 install openpyxl
```

次に、先ほどのソースコードを以下のように修正します。修正した箇所は、2行目のopenpyxlのインポートと、最終行のファイル出力をto_excel()メソッドに変更したのみです。

リスト10.20:

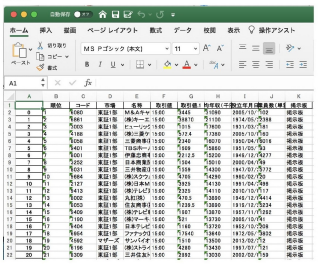
```
import pandas as pd
import openpyxl

url = "https://info.finance.yahoo.co.jp/ranking/?kd=45"

df = pd.read_html(url)
df[0].to_excel("data.xlsx", encoding="SHIFT-JIS")
```

これで、data.xlsxというExcel形式のファイルが出力されました。このファイルを開いてみます。以下のようなExcelデータであることを確認できます。

図 10.5:



CSVと比べたときのExcel形式のメリットとして、シートを分けてデータを管理できる点があります。今回は、上場企業の平均年収ランキングと、時価総額ランキングの結果を別々のシートに保

＜＜目次＞＞

＜＜著者紹介＞＞

齊藤 貴義

1979 年 9 月 11 日、福島県相馬市生まれ。Web エンジニア・インフラエンジニア。ニックネームはサイバーメガネ。高校時代にパソコン通信でネットワークの面白さに触れてコンピュータにのめり込む。Web ベンチャーでモバイルサービスや SNS をメインとしたシステムを多数開発。2019 年 2 月に、IPUSIRON（@ipusiron）と共にサークル「ミライ・ハッ

キング・ラボ」を結成して同人誌の販売を開始。執筆活動・受託開発・コンサルティングなどを行っている。趣味はブログと写真。

＜＜販売ストア＞＞

電子書籍:

Amazon Kindle ストア、楽天 kobo イーブックストア、Apple Books、紀伊國屋書店 Kinoppy、Google Play Store、honto 電子書籍ストア、Sony Reader Store、BookLive!、BOOK☆WALKER

印刷書籍:

Amazon.co.jp、三省堂書店オンデマンド、honto ネットストア、楽天ブックス

※　各ストアでの販売は準備が整いしだい開始されます。

※　全国の一般書店からもご注文いただけます。

【インプレス R&D】 https://nextpublishing.jp/

株式会社インプレスR&D（本社:東京都千代田区、代表取締役社長:井芹昌信）は、デジタルファーストの次世代型電子出版プラットフォーム「NextPublishing」を運営する企業です。また自らも、NextPublishing を使った「インターネット白書」の出版など IT 関連メディア事業を展開しています。

※NextPublishing は、インプレス R&D が開発した電子出版プラットフォーム（またはメソッド）の名称です。電子書籍と印刷書籍の同時制作、プリント・オンデマンド（POD）による品切れ解消などの伝統的出版の課題を解決しています。これにより、伝統的出版では経済的に困難な多品種少部数の出版を可能にし、優秀な個人や組織が持つ多様な知の流通を目指しています。

【インプレスグループ】 https://www.impressholdings.com/

株式会社インプレスホールディングス（本社:東京都千代田区、代表取締役:松本大輔、証券コード:東証 1 部 9479）を持株会社とするメディアグループ。「IT」「音楽」「デザイン」「山岳・自然」「モバイルサービス」「学術・理工学」「旅・鉄道」を主要テーマに専門性の高いメディア&サービスおよびソリューション事業を展開しています。さらに、コンテンツビジネスのプラットフォーム開発・運営も手がけています。

【お問い合わせ先】

株式会社インプレス R&D NextPublishing センター
TEL 03-6837-4820
電子メール: np-info@impress.co.jp