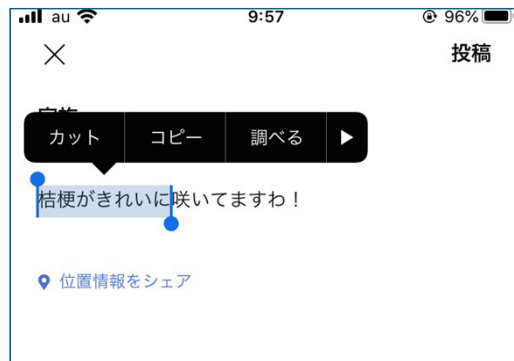


YouTubeの動画情報
を分析してみよう！

Webのデータ収集

- 以下のような時に、「範囲選択 → コピー → 貼り付け」をしていたら日が暮れてしまう
 - 気象庁Webサイトにある天気情報を10年分取得したい
 - Amazonの商品で、名前、価格、評価点数を100製品分取りたい
 - Xのポストを10万件取得したい

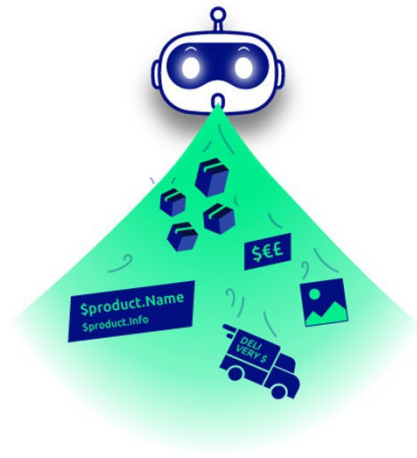
そのために
使われる技術が
Webスクレイピング



Webスクレイピング

Web + Scraping （こする・削る・かき出す）

- Web の情報をかき出して保存していく技術
- 手動コピペより効率よく行うためのテクニック
- 類義語としてWebクロールがある



スクレピングとクローリング

どちらも似た単語であるが、目的が異なる

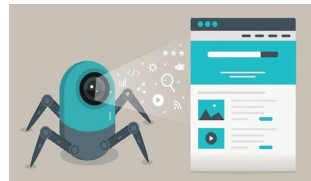
Webスクレイピング

- 特定のページや目的がある
 - 価格やコメントなど特定の情報の抽出



Webクローリング

- 広範なWebページを巡回
 - 検索エンジンが活用できるインデックスを作成
 - サイト内ページの関係性をマッピング



1. データ収集の仕組み

- Web上のデータを自動的に取る方法は、大きく分けて2種類

1. HTMLの解析

- HTMLのソースコードからデータを抜き出す

2. APIを利用する（今回はこっち）

- 定められたURLにアクセスしてデータを受け取る

プログラムによるデータ収集の可否

● Webサイトの利用規約

- Amazon:

<https://www.amazon.co.jp/gp/help/customer/display.html?nodeId=GLSBYFE9MGKKQXXM>

利用許可およびサイトへのアクセス

本規約およびサービス規約の遵守を条件とし、アマゾンまたはコンテンツ提供者は、アマゾンサービスを限定的、非独占的、非商業的および個人的に利用する権利をお客様に許諾します（譲渡およびサブライセンス不可）。この利用許可には、アマゾンサービスまたはそのコンテンツの転売および商業目的での利用、製品リスト、解説、価格などの収集と利用、アマゾンサービスまたはそのコンテンツの二次的利用、第三者のために行うアカウント情報のダウンロードとコピーやその他の利用、データマイニング、ロボットなどのデータ収集・抽出ツールの使用は、一切含まれません。本規約またはその他の規約

● robots.txt

- クローラーに対するアクセス許可管理のファイル
- 通常はルート直下にある
`example.com/robots.txt`

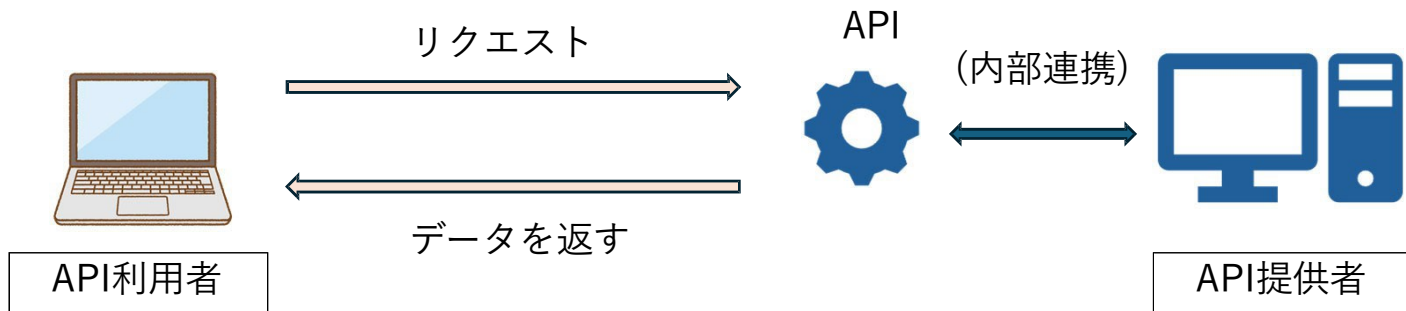
```
User-Agent: *  
Disallow: /images/  
Disallow: /backup/  
Disallow: /cgi-bin/  
Disallow: /shops/  
Disallow: /shop/ascosing/  
Disallow: /shop/egs-g/  
Disallow: /shop/tripleo-shop/  
Disallow: /shop/foody/  
Disallow: /shop/kira-con02/  
Disallow: /*/lightbox_*.html  
Disallow: /aboutus/am/maint/
```

<https://www.rakuten.co.jp/robots.txt>

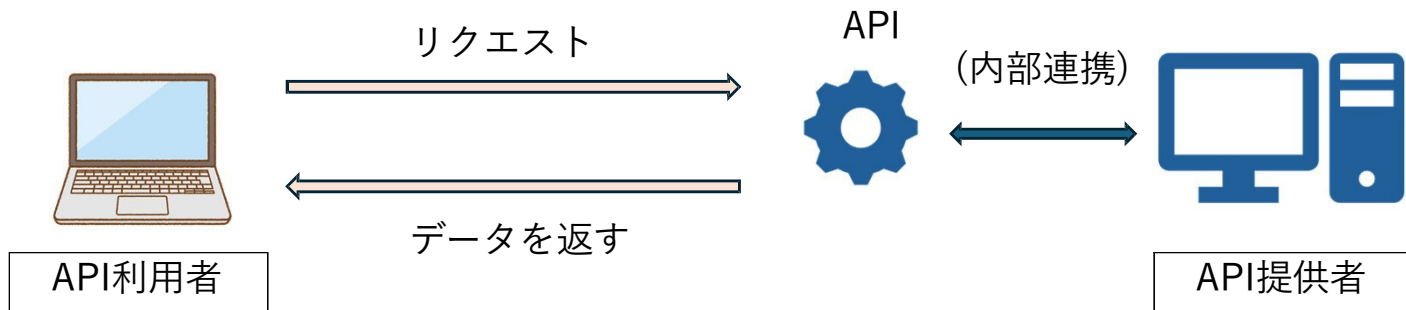
2. APIの利用

Application Programming Interface

- ソフトウェアやプログラム、Webサービスの間をつなぐインターフェース
- インターフェースとは何らかの境界面・接点を意味し、異なる2つをつなぐ
- データを管理する側がAPIを提供



APIを設置するメリット



データ利用者

- データ取得PGを自作しなくて済む
- JSONなど扱いやすい形でデータを大規模に取得できる
- 利用規約に基づいている

データ提供者

- アクセス管理・制限ができる
- サーバー負荷の軽減
- 収益化の機会となる

今回利用するAPI

- Youtube Data APIを利用
- みなさんの好きなユーチューバーや動画のデータを取得して分析を行う

例) 特定のチャンネル, ジャンル, 動画リスト等

今回行うこと



1. データを取得するユーチューバーや動画を決める
2. ユーチューバーのチャンネルID、動画の動画IDを取得する
3. プログラムにチャンネルIDや動画IDを設定する
4. プログラムを実行する
5. Webからさまざまなデータが取得できること、取得したデータを使ってさまざまな分析ができることを知る
6. 分析結果から知見を読みとる（考察する） ← 2年生の認知情報科学実験で実施

注意：この2つは別物

- **結果**：出てきた数値，得られたデータ（客観的，事実ベース）
- **考察**：数値やデータ，その分析結果から言えそうなこと（主観的，考えられること）

例) このアルゴリズムでは，精度が10%向上した。
したがって，この結果からこのアルゴリズムは有効であるといえる。

サンプルプログラムの説明

1.スクレイピング(チャンネル：タイトルと日付と視聴回数).py

1-1.データ分析：棒グラフ(日毎の視聴回数).py

1-2.データ分析：棒グラフ(月毎の視聴回数).py

1-3.データ分析：棒グラフ(年毎の視聴回数).py

1.** ~ 4.**のプログラムを実行すると、その結果はoutput.txt というファイルに出力される。

2.スクレイピング(チャンネル：視聴回数と高評価).py

2-1.データ分析：散布図(視聴回数と高評価).py

2-2.データ分析：評価率トップ10.py

1-1.* ~ 1-3.* のプログラムは、1.* の結果を利用する

2-1.* ~ 2-2.* のプログラムは、2.* の結果を利用する

3.スクレイピング(チャンネル：視聴回数とコメント数).py

3-1.データ分析：散布図(視聴回数とコメント数).py

3-2.データ分析：コメント率トップ10.py

3-1.* ~ 3-2.* のプログラムは、3.* の結果を利用する

4.スクレイピング(動画のコメント一覧).py

4-1.データ分析：コメントのワードクラウド.py

4-2.データ分析：コメントのネガポジ分析.py

4-1.* ~ 4-2.* のプログラムは、4.* の結果を利用する

※ 人気のユーチューバーや動画では実行に少し時間がかかるかもしれません

チャンネルID、動画IDを設定する

次のプログラムにチャンネルIDを設定してください

1.スクレイピング(チャンネル：タイトルと日付と視聴回数).py

```
67  #タイトルと視聴回数
68  API_KEY = "" # APIキー
69  channel_id = "" # チャンネルID
```

2.スクレイピング(チャンネル：視聴回数と高評価).py

```
62  # APIキーとチャンネルID
63  API_KEY = "" # APIキー
64  channel_id = "" # チャンネルID
```

3.スクレイピング(チャンネル：視聴回数とコメント数).py

```
62  # APIキーとチャンネルID
63  API_KEY = "" # APIキー
64  channel_id = "" # チャンネルID
```

次のプログラムに動画IDを設定してください

4.スクレイピング(動画のコメント一覧).py

```
42  # APIキーと動画ID
43  API_KEY = "" # APIキー
44  video_id = "" # 動画ID
```

チャンネルIDと動画IDの探し方は先輩(4年生)に聞いてください！

3. YouTubeの世界を覗いてみよう

みなさんの関心があるユーチューバーや動画は？

YouTuber

東海オンエア
フィッシュャーズ
レイクレ
ヒカキン
はじめしゃちょー
よにのチャンネル
カジサック
ばんばんざい
平成フラミンゴ

人気アニメ

ONEPIECE
僕のヒーローアカデミア
SPY×FAMILY
呪術廻戦
チェンソーマン
鬼滅の刃
名探偵コナン
ハイキュー!!
かぐや様は告らせたい
ブルーロック

人気ゲーム

- ・ HikakinGames
- ・ Minecraft
サワヤンゲームズ
- ・ マリオカート
ロードシルク
- ・ スマッシュブラザーズ
V.I.P
- ・ プロ野球スピリッツA

スポーツ

OパリーグTV (野球)
DAZN Japan (サッカー、野球など)
B.LEAGUE (バスケ)
TBSスポーツ (スポーツ全般)
JAPAN RUGBY TV (ラグビー)

楽しめましたか？

千葉工業大学 情報変革科学部 認知情報科学科

情報科学と認知科学の両面からコンピュータの仕組み（ソフトウェア／AI／メディア／Web）と人間の仕組み（知能／感性／心理）、またそれらの関わり（エンジニアリング／プロジェクト／学習・教育／デザイン）について網羅的に学修します

