世界のけんティーと学ぶ
ウキウキ
スクレイピング講座 Ver1.1

~ワイ将、pythonと共に~

(2020/03/02)

1.1. 導入

- 環境はUbuntu14.04(virtualbox)
- 使うのはpython3.4.3

- プリミティブな開発を行いたいので、極力IDEは使用しない(というか入れ方がわからない)。

- 必要な物をpip3コマンドでインストールする。

1.2. pip3とは…

- The python package indexに公開されているpythonパッケージのインストールなどを行うユーティリティ
- おそらくpython3用なので、pip3だと思われる。

- これをターミナルで打ち込めば色々パッケージをインストールできる。

1.3. pip3を用いてインストールするもの

- RequestsとBeautifulSoup4をインストールする

- **Requests**

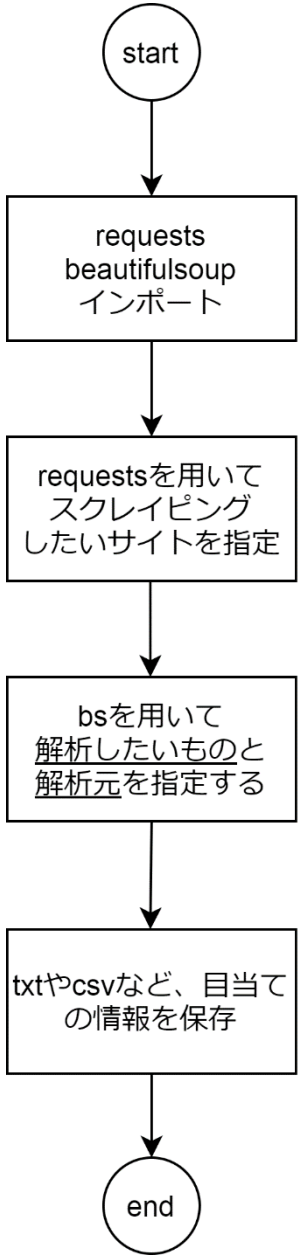
→URLを指定して、コマンドにより指定した形でデータを取得(おそらく)

- **BeautifulSoup4**

→htmlなどからデータを抽出するために使用(スクレイピング本体)

1.4. フローチャート

基本的にはmain関数の中身をこのようにする。



*bs=beautiful soup

1.5. 準備

- pip3をインストール

```
-> sudo apt-get install python3-pip
```

- Requestsをインストール

```
-> pip3 install requests
```

- BeautifufSoupをインストール

```
-> sudo pip3 install beautifulsoup4
```

*beautifulsoupのインストールはsudoコマンドが無いとキレられる^^;

1.6.1 簡単なスクレイピング

- title.py(yahooニュースのタイトルを取得)

```
import requests
from bs4 Import BeautifulSoup

r = requests.get('https://news.yahoo.co.jp/')

#第一引数:解析したいもの,第二引数:何を元にするか
soup = BeautifulSoup(r.content,"htmp.parser")

#ニュース一覧のテキストのみを表示
print(soup.find("ul","newsfeed_list").text)
```

1.6.2. title.py 結果

```
[~/python]--> python3 title.py
藤田ニコル、無観客のTGCで“投げキス” ミニスカートでランウェイを歩き「ファンサ神すぎ」ENCOUNT浜田ブリトニー 1歳娘が溶連菌性感染症と診断「心配です」デイリースポーツ清原和博、銀座で大荒れして警察沙汰に【画像あり】SmartFLASH汚部屋一直線！衣類で部屋が散らかる人の特徴5つサンキュ！付き合っていないのにキスをする意味・男子の心理・女子のホンネ12選CanCam.jp錦戸亮 仕事減で吐露した焦り…知人らに“オファー要求”も女性自身3:12＜動画＞Matt、TGC初出演で美しい歌声披露 MC田中みな実「この世のものとは思えない」と絶賛MANTANWEB小倉優子 第3子妊娠発表までの不安を吐露「胎盤剥離が見られたり…心配ばかりでした」スポニチアネックストリンドル玲奈、芸名は母の旧姓から「肉丸玲奈」も考えていたSmartFLASH’19ベストセラーバイクはZ900RS/Ninja400/レブル250WEBヤングマシンこれはヤバい…!?和牛・水田、インスタで“ファンのマナー違反に苦言か？”と思いきや…スポニチアネックス薬物を断つのは本当に無理なのか…治療的司法のススメTOKYO MX＜麒麟がくる＞織田信長役・染谷将太が語る登場シーン 「本番直前に奇跡的に太陽が…」MANTANWEB【2020年3月★蟹座】今月の運勢・無料占いハルメクWEB高知東生さん「これだけのカメラは久しぶりで…ドキッっとする」逮捕から4年 初の公の場中日スポーツ【プチプラパンツコーデ】春まで着られる組み合わせを大公開！集英社ハピプラニュース＜白石麻衣＞「乃木坂46」としてTGCラストパフォーマンス 無観客開催の大トリ務める毎日キレイ辻希美、三男とのチュウ顔2SHOT &“つけ心地が良すぎる”自作マスク姿公開「自慢じゃないが…」E-TALENTBANK[自転車世界選手権]ホッ！落車の梶原悠未は軽傷「骨折しなかったのが何より」笑顔の金メダリスト中日スポーツ岡村隆史から吉本に誘われた中居正広「ギャラ8：2なら行く」SmartFLASH3歳から乗れる電動バイクで、キッズにバイクの楽しさを体験させるGQ JAPAN“3児の母グラドル”熊田曜子が明かすグラビア界の現状ザテレビジョン＜NGT48荻野由佳＞オーディション勝ち抜きTGCに ミニドレスでランウエー歩く 蜷川実花ステージにAKBグループ15人毎日キレイ大桃美代子、震災後の『めちゃイケ』収録を回顧「いずれ笑いが必要」SmartFLASH森麻季アナ“女子アナの実態”に言及、田中みな実が「凄い」理由明かすスポニチアネックスジジ＆ベラの母、ヨランダ・ハディッドがランウェイに復帰！ 美しきモデル一家の共演にラブコール続出【SPURセレブ通信】集英社ハピプラニュース「喜寿になりました！」 加藤綾菜、加藤茶の77歳バースデーを祝福“ラブラブ同居”10年目ねとらぼ舛添要一氏、「サンデー・ジャポン」に落胆 「影響力を考えると少しがっかり」ENCOUNTしっぽピーン！ 嬉しい時にする「うれしっぽ」ねこのきもち WEB MAGAZINE【正規ディーラーでも勧められることあり！】クルマのカーボナイザー洗浄・スラッジナイザー洗浄はやるべき？WEB CARTOP2:13母親が新型コロナに感染した女性が語る"医療現場の実態" 発熱・倦怠感あってもすぐに検査受けられず…北海道ニュースUHB＜翔＞インスタで人気の美少年、13歳になり大人っぽく TGCランウエーに毎日キレイ「めちゃめちゃ美しいです」大島優子、中山美穂&木村多江との豪華3ショット公開ENCOUNT理不尽すぎる？「完璧妻」の夫が不倫に走る理由All About名古屋市、50、80代女性2人の感染確認 既に入院の70代女性と同居 新型コロナ毎日新聞香取慎吾、加藤浩次と「ななにー」ショット 稲垣＆草ナギとの貴重4ショットも「加藤さん自撮り練習してね」の声スポーツ報知GU賢者の「GU MEN」活用術 | Martmagacol有吉、テレビ番組の“お約束行動”に不満「そのくだりが余計なんだよ！」スポニチアネックス父を無視・あからさまに嫌な顔をする娘に、母は「父娘が遭遇しないよう配慮」…思春期の娘への過剰な気遣いが逆効果となって…読売新聞（ヨミドクター）占い師ラヴィー・ヒトミの“心を調律する”アドバイス【第3回】「出世するにはどうしたらいいか？」ＢＥＳＴ ＴＩ ＭＥＳ空気を読めない人と読み過ぎる人、どちらが悪い？All Aboutともさかりえ 中3息子の「ふつうの毎日を最後までちゃんとやりたかった」に胸痛デイリースポーツ
[~/python]--> 
```

1.7.1. もし実行できなかったら…

- RequestsとBeautifilsoup4が異なるディレクトリに保存されている場合、どちらかが読み込まれない可能性がある。

- その場合、環境変数PYTHONPATHを設定すれば解消できる

*ちゃんと実行できてたら1.7節は飛ばしてクレメンスm(_ _)m

1.7.2. PYTHONPATH設定

- まずrequestsとbeautifulsoupがどのディレクトリにあるか調べる
  →pip3で再度インストールすると場所を教えてくれる

- 場所がわかったら~/.bashrcにPYTHONPATHを設定(記述)し、reboot
- あちきの場合は

.bashrc

```
export PYTHONPATH="¥usr¥local¥lib¥python3.4¥dist-packages:¥usr¥lib¥python3¥dist¥packages"
```

Requestsの場所

複数指定の場合、「:」記述

bsの場所

まとめと展望(Ver1.1)

- まとめ

→スクレイピングで使用するライブラリの準備

→スクレイピングの流れの理解(データ化→データ抜き出し)

- 展望

→htmlのおおまかな構造の理解

→requestsとBeautifulSoup4の関数の使い方の確認