

# WEBからの図書情報の収集による図書館システムの試作

---

東京情報大学 総合情報学科 情報システム学系

知能情報システム研究室 永井ゼミ

# 研究背景

---

- データの継続的な開発に伴い、個人のデータに対する需要も日々高まり、個人や中小企業が必要とする小規模なデータをWebクローラーを通じて取得することができます。Webクローラープログラムは、URLアドレスを介してWebページ情報を読み取り、Webページのテキスト情報を解析およびフィルタリングして、目標データを形成します。 Pythonで記述されたWebクローラーは、 マルチスレッド (multithreading) を使用してプログラムの効率を向上させ、ロックして取得したデータの正確性を確保し、データベースを使用してクロールされた大規模データを保存します。

# 研究目的

---

- このシステムがクローラープログラムは、 、主に複数のWebサイトの書籍情報をユーザーに提供し、最終的にWebページの形式でユーザーに提示されます。まず、データベースからウェブページのアドレスと対応するRE規則を読み取り、クローラープログラムがウェブページのRE規則を組み合わせて、データを、接続されているデータベースにデータを保存してから、HTMLテキストにデータを書き込むことで、情報集約プロセス全体が完了します。
- 例： Amazon, bookoffなど値段を比較
- 各地図書館書籍情報集約、資料整合

# 進捗と予定

---

- 今まで
- データベース：Book表、Table\_re表、Table\_web表
- データベースの接続： テスト完了
- Webクローラー：ウェブとアドレスをキャプチャー、コーディング問題また存在する。
- これから
- 9月前：ウェブが正確な分析して、書籍タグ部分を取得し、データベースに添加する
- 10月前：book表に内容を読み込みして、htmlを生成する。
- 11月前：全部機能をテストする。
- 12月前：卒業論文を書く。

# プログラム一覧

---

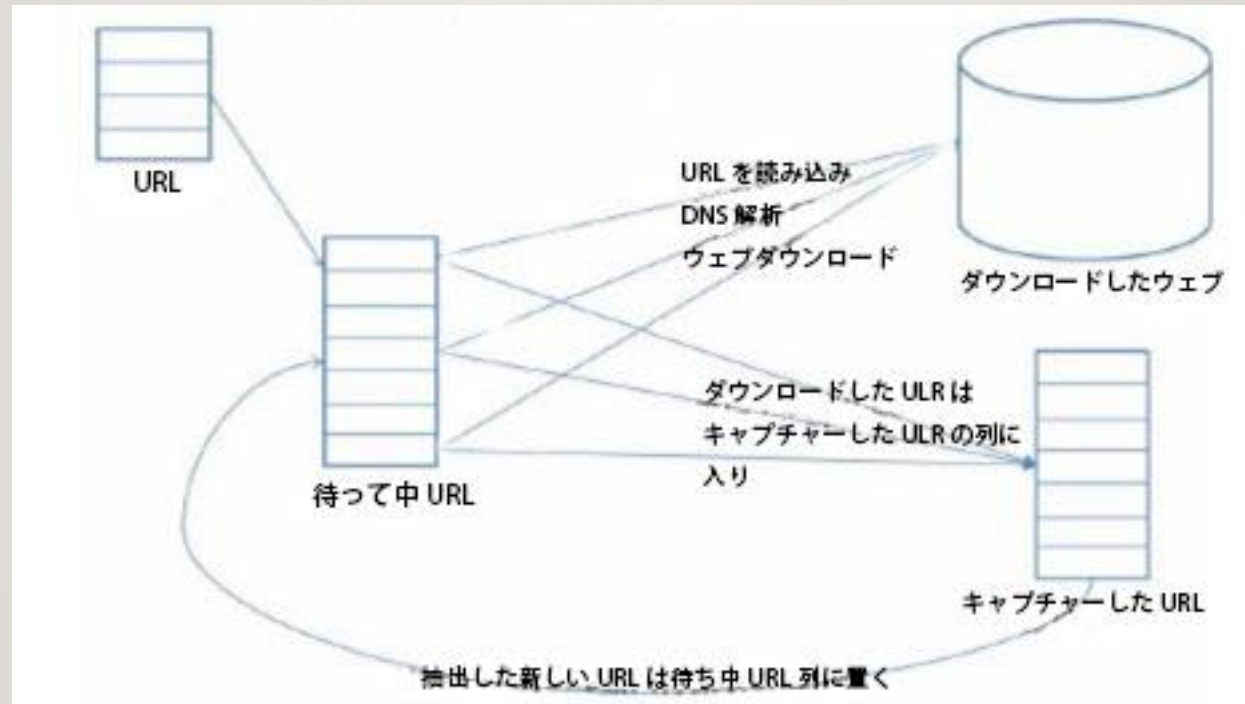
- Python3.6
- Visual Studio Code
- MySQL
- MySQL connector Python
- MySQL Workbench

# 設計概要

---

- この設計は、書籍情報の集約を実現するために、複数のWebサイトの書籍情報はクローラーを通してキャプチャーするプログラム。
- プログラムは最初に各URLのアドレスと対応するWebサイトがデータベースをキャプチャーする必要があるページ数を読み取りプログラムのデータを取得してから、Pythonプログラムをキャプチャーして目的の書籍の目的情報にキャプチャーし、キャプチャーされたデータを自分のデータベースに移動し、それを取り出してHTML文書に書き込み、Webページの形式でユーザーに表示します。

# ウェブクロールフレーム



# 機能設計…データベースの接続

---

- 一、はpythonとデータベースの間に正しい接続します。その機能はデータベースの操作に対する機能の基礎です。

# 機能設計…データベースの読み込み

---

- データベースはすべてのURLアドレスとそのウェブのhtml文書の正規表現式 (RE) を読み込み、
- この機能を実行できない場合は、各正規表現をプログラムに手動で書き込む必要があります。各正規表現をプログラムに手動で書き込む必要があります。この場合、データを抽出できるのは1つのWebサイトだけであり、データの変更も非常に不便です。
- 今の解決策：read関数を使うで内容を読み込み、コーディング問題 UTF8、GBK と GB2312など、解決策は組み込み関数decodeとencodeで運用

# 機能設計…データベースの入力

---

- 図書の各方面の情報を抽出し、データベースの図書表に入力します

# 機能設計…ウェブページの読み込み

---

- ウェブ情報を読み込み、一つのウェブHTML文書を開けてから、その後処理します

# 機能設計…ウェブページの分析

---

- ウェブページの分析、HTMLファイルの読み込みの後 発見した 一つの図書販売のウェブにたくさんな本がある。まず、毎本のタグ部分を取得し、次、その部分の本の他の属性を抽出します。

# 機能設計… HTMLのファイルの正確な生成

---

- HTMLのファイルの正確な生成 ユーザに向けるUIの設定を提供します。全部キャプチャーした図書情報はこのウェブに示します 今まで、この部分を考えているからである。

# データベース

---

- 表が3つあります
- 1 **web spider** を使うしたい**URL**アドレスを保存します。
- 2 正規表現式と他の情報を保存する表。
- 3 抽出した図書情報を保存する表

# TABLE\_WEB表

- 
- Field | Type | Null | Key | Default | Extra
  - +-----+-----+-----+-----+-----+-----
  - | web\_id | int(10) | No | ☒ | NULL | web番号
  - | web\_name | varchar(100) | No | | NULL | web名前
  - | web\_url | varchar(255) | No | | NULL | url
  - | web\_num | int(10) | No | | NULL | webページ

# TABLE\_RE表

- 
- Field | Type | Null | Key | Default | Extra
  - +-----+-----+-----+-----+-----+-----
  - | re\_id | int(10) | No | ☒ | NULL | 正則番号
  - | web\_id | int(10) | No | | NULL | web番号
  - | row\_name | varchar(100) | No | | NULL | RE名
  - | re | varchar(255) | No | | NULL | 正則表現式

# BOOK表

- 
- Field | Type | Null | Key | Default | Extra
  - +-----+-----+-----+-----+-----+-----
  - | insert\_time | varchar | No | | | 時間
  - | web\_name | varchar(100) | No | ○ | NULL | web名前
  - | title | varchar(255) | No | | NULL | 本名
  - | book\_url | varchar(255) | No | | NULL | 本のurl
  - | book\_id | int(10) | No | ○ | NULL | 本の番号
  - | author | varchar(255) | Yes | | NULL | 作者
  - .....

---

ご清聴ありがとうございます