

Deep Web Crawler

(ระบบรวบรวมข้อมูลจาก Deep Web)



Deep Web คืออะไร

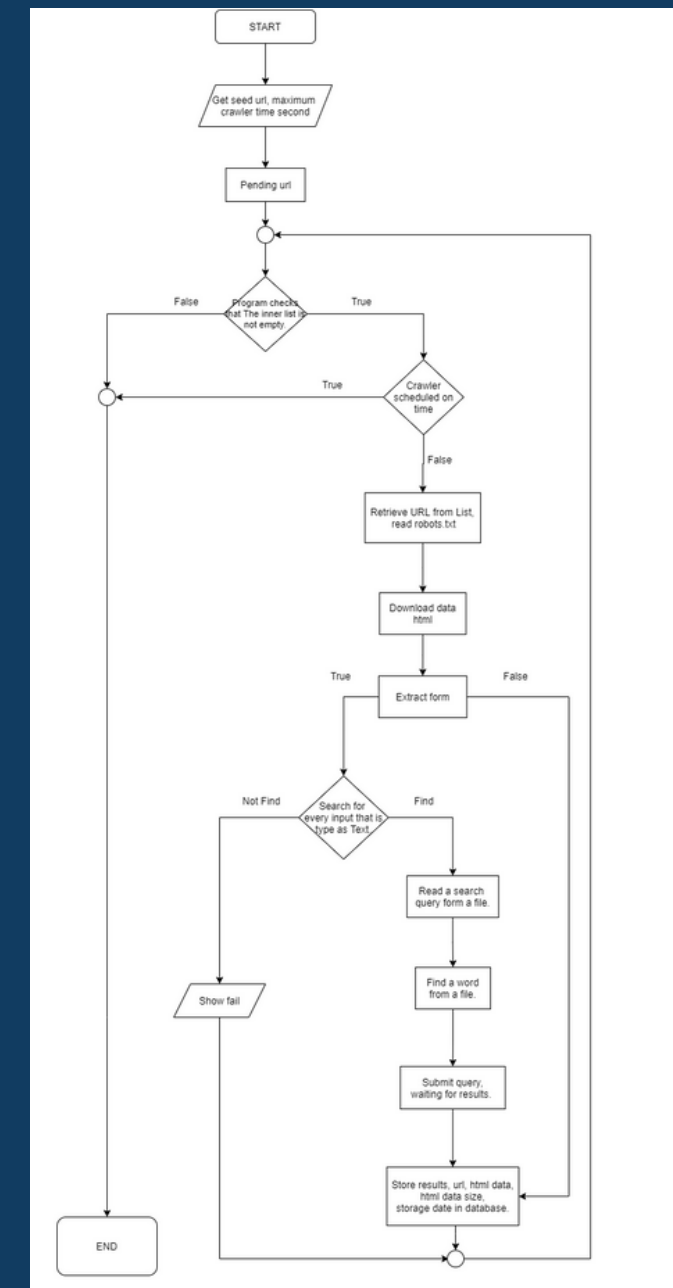


เป้าหมายของโครงการ Deep Web Crawler

1. ออกแบบและพัฒนาระบบ Deep Web
2. พัฒนา Web Site เรียกดูข้อมูลจาก Database

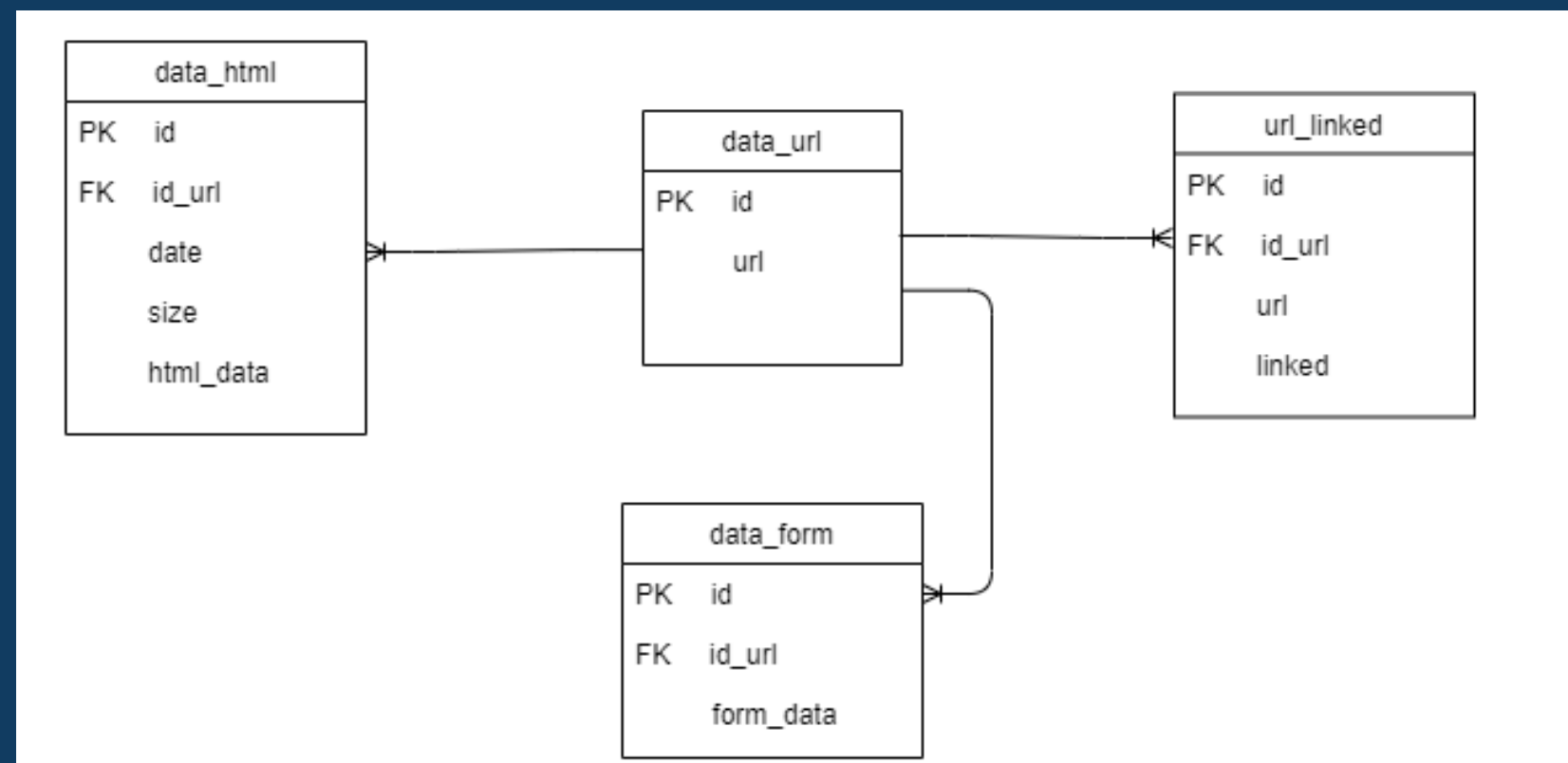
Flowchart อธิบายการทำงานของ Deep Web Crawler

1. ดาวน์โหลดเว็บเพจที่ต้องการจะ Crawler
2. สกัดฟอร์ม
3. ถ้ามีฟอร์มก็ส่ง Submit Query
4. จัดเก็บข้อมูลลงดาต้าเบส

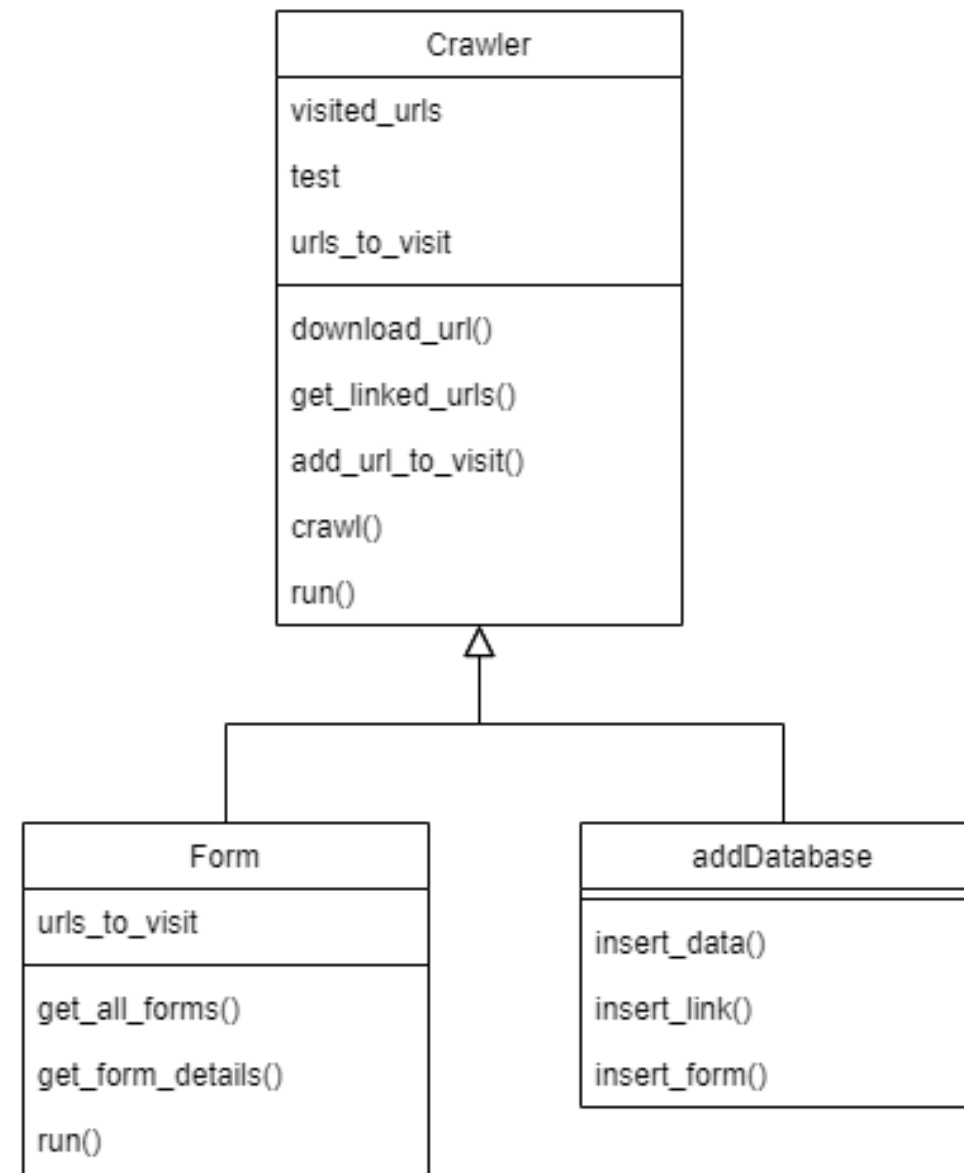


”

โครงสร้างดาต้าเบส



CLASS DIAGRAM



UI ของโปรแกรม Crawler และ Web page

```
2021-06-25 09:46:06,878 INFO:Crawling: https://www.sanook.com/  
Crawling so : 8007
```

```
2021-06-25 09:46:11,068 INFO:Crawling: https://www.sanook.com/news/  
Crawling so : 8008
```

```
2021-06-25 09:46:15,544 INFO:Crawling: https://www.sanook.com/news/entertain/  
Crawling so : 8009
```

คำค้น = <http://guru.sanook.com/>

date	detail	url
Tue, 15 Jun 2021 17:44:10 GMT	ความรู้ เทคโนโลยี สารานุกรม สารานุกรมออนไลน์ ความรู้รอบตัว ความรู้ทั่วไป พจนานุกรม เทมส์ เพลงใหม่ เพลง	http://guru.sanook.com/

Crawler

ค้นหา

คำค้น = ข้าว

id	detail	url
15	ค้นหาร้านอาหาร โรงแรม ที่พัก สถานที่ท่องเที่ยว ร้านอาหารสวยและสปา ใน กรุงเทพฯและปริมณฑลภาพถ่ายโดย Pednoi AhHaค้นหาร้านอาหารใกล้เคียงร้านอาหารร้านสวย และ สปาที่พักสถานที่ท่องเที่ยวที่บันทึกไว้ร้านอาหารกาแฟ / ของหวานรับที่ร้านดีสุดคุ้มสุดคุ้มUsers' Choiceสิ่งเคลียร์เพิ่มเติม..สั่งผ่านแอปฯ แวะรับที่ร้านได้เลย!ดูทั้งหมด »Blue Elephant Cooking School and Restaurant สAVORได้4.4240 รีวิว88888เปิดอยู่ทั้งปีเดอะโกลด์ ลีแอนด์	https://www.wongnai.com/9681-กรุงเทพฯและปริมณฑล

เก็บข้อมูลทั้งหมด 5923 เว็บเพจ จาก 12 เว็บไซต์
ขนาดข้อมูล 2500 MB
เก็บฟอร์มได้ 5435 ฟอร์ม
พบ 1 ฟอร์มต่อ1เว็บเพจ
เวลา 1 ชั่วโมง สามารถเก็บข้อมูลได้700ถึง800
เว็บเพจ

สิ่งที่ระบบทำได้และทำไม่ได้

สิ่งที่ทำได้

- ระบบรวบรวมข้อมูลเป็นแบบ SINGLE THREADS
- การค้นหาFORMและใส่ข้อมูลลงในFORMจากไฟล์ ยังเป็นแบบTEXT อย่างเดียว และคีย์เวิร์ดค้นหาจากไฟล์ยังพึกไว้อยู่
- เว็บสามารถแสดงข้อมูลที่เก็บไว้ในดาต้าเบสได้

สิ่งที่ทำไม่ได้

- ระบบรวบรวมข้อมูลที่เป็นแบบMULTI THREADS
 - การค้นหาFORMและใส่ข้อมูลลงในFORM
- รูปแบบอื่นยังทำไม่ได้