Deep Web Crawler

(ระบบรวบรวมข้อมูลจาก Deep Web)

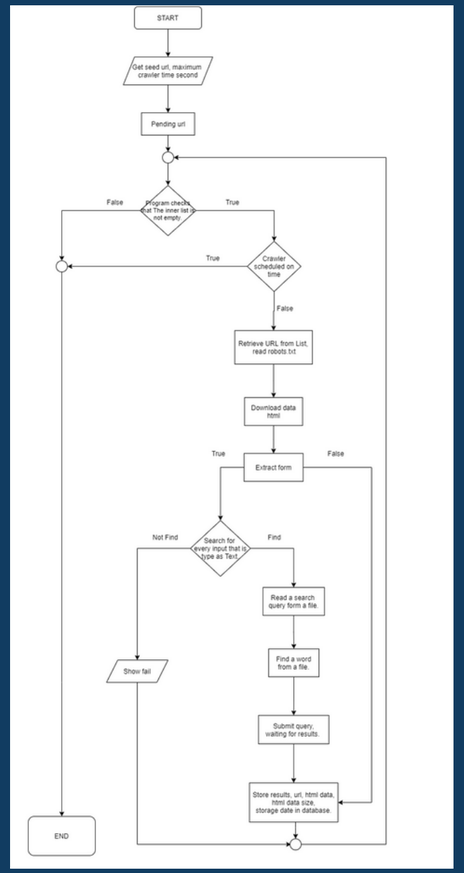
Deep Web คืออะไร

เป้าหมายของโครงการ Deep Web Crawler

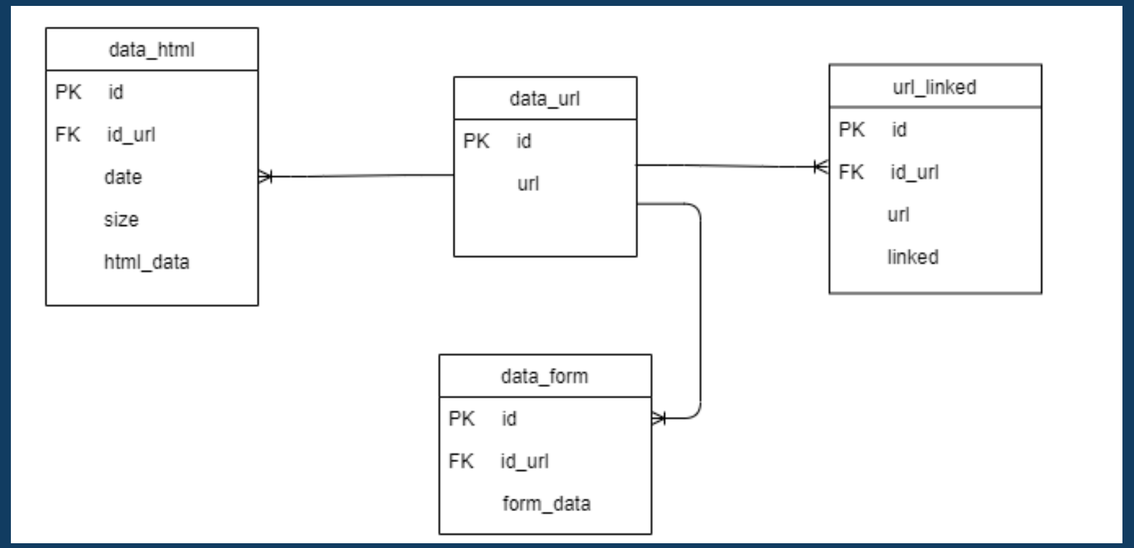
1. ออกแบบและพัฒนาระบบ Deep Web
2. พัฒนา Web Site เรียกดูข้อมูลจาก Database

Flowchart อธิบายการทำงานของ Deep Web Crawler

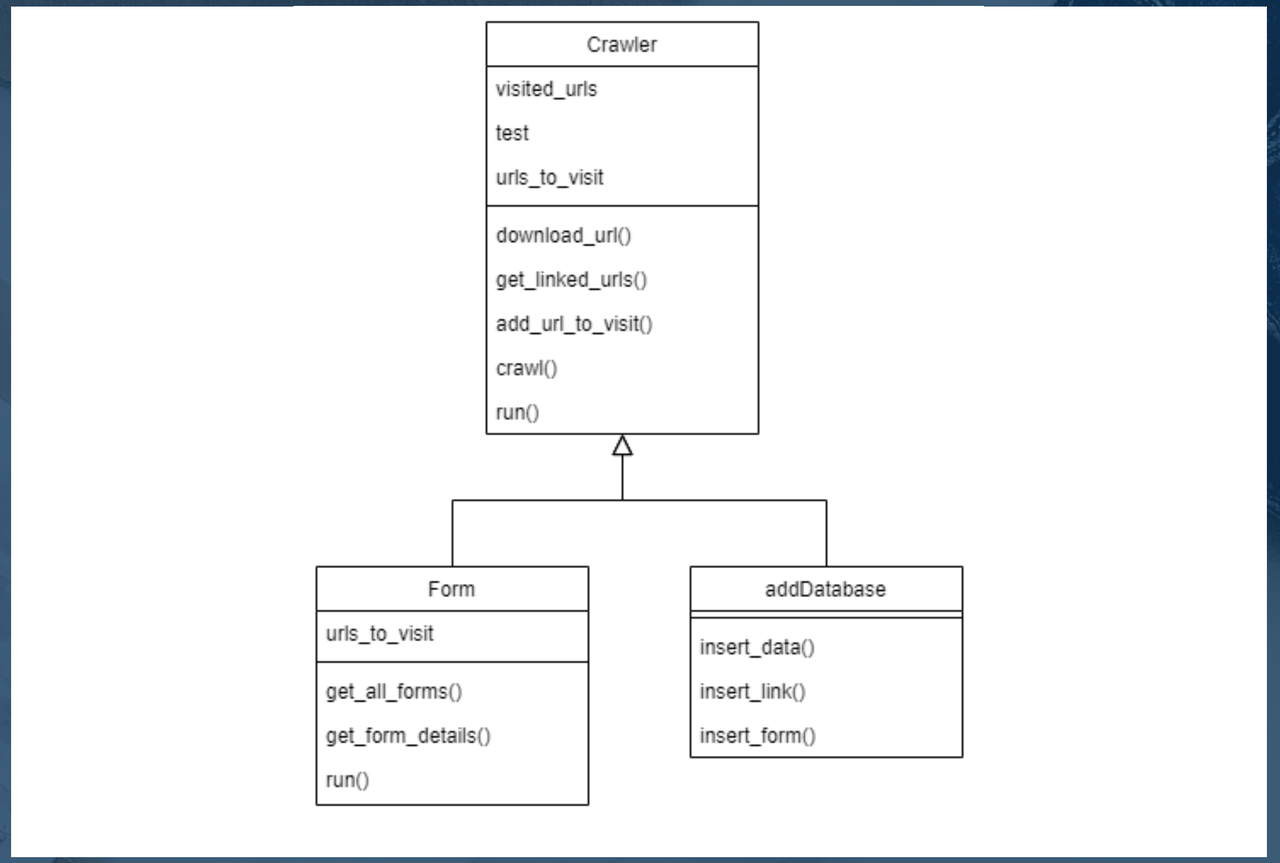
1. ดาวน์โหลดเว็บเพจที่ต้องการจะ Crawler
2. สกัดฟอร์ม
3. ถ้ามีฟอร์มก็สั่ง Submit Query
4. จัดเก็บข้อมูลลงดาต้าเบส



โครงสร้างดาต้าเบส



CLASS DIAGRAM

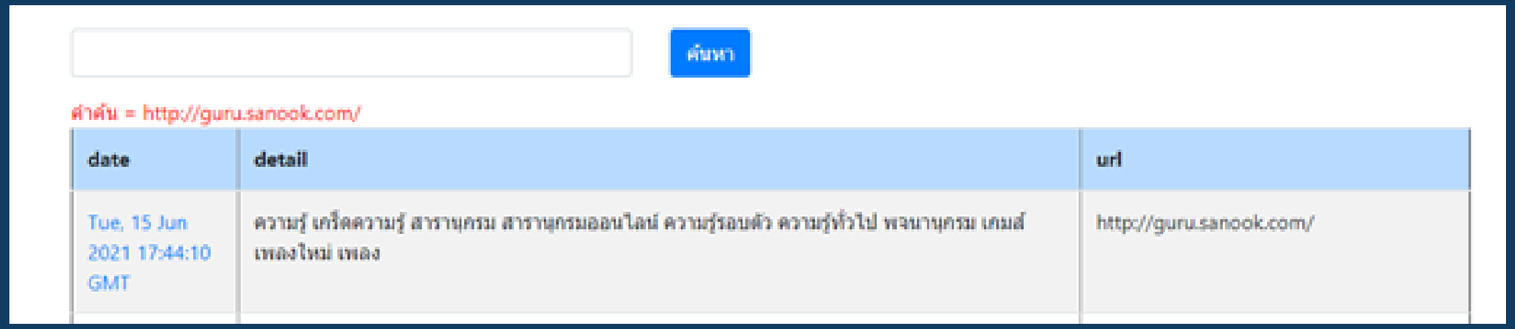


UI ของโปรแกรม Crawler และ Web page

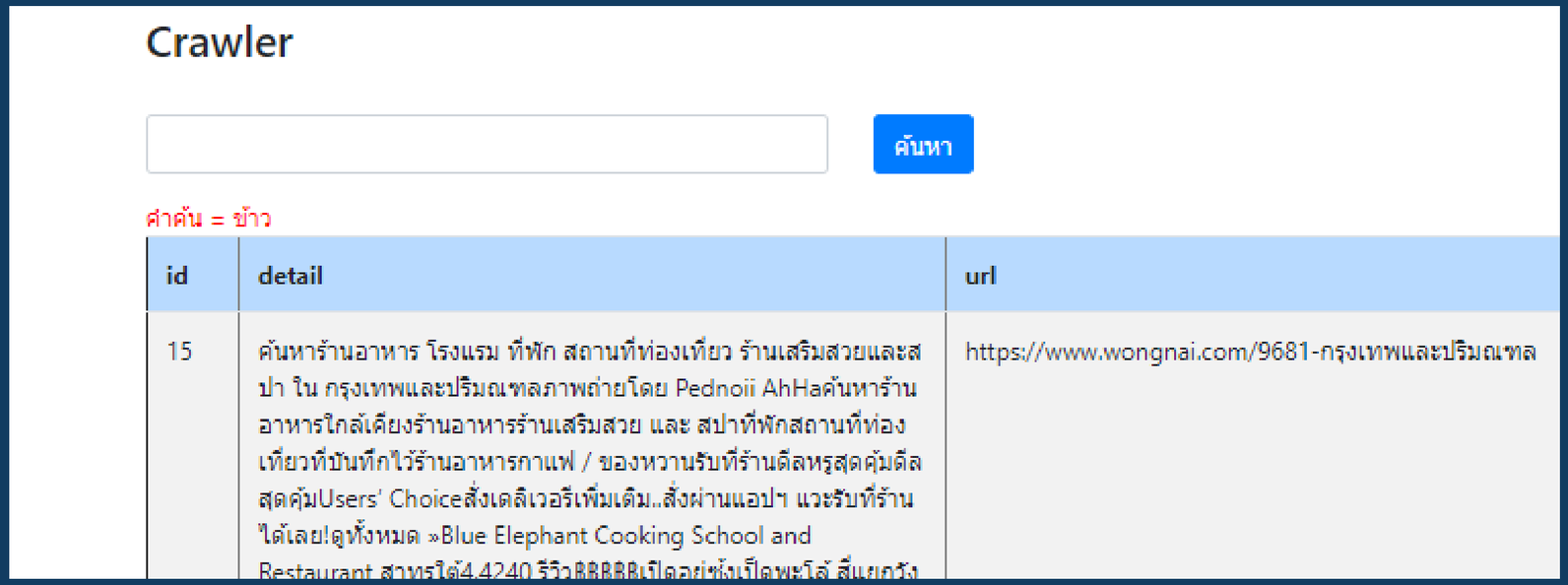
```
2021-06-25 09:46:06,878 INFO:Crawling: https://www.sanook.com/
Crawling so : 8007

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
```

1.กรอก urlและเวลาที่ต้องการค้นหา
2.โปรแกรมจะทำการ Crawl url, data ของ html, size ของ html,
form ของ html, วันเวลาที่ Crawl เก็บลงในดาต้าเบส
3.โปรแกรมจะหยุดทำงานตามเวลาที่ใส่ไว้ตามขั้นตอนที่ 1 หรือ ไม่มีข้อมูลให้Crawl

UI ของโปรแกรม Web page



1. ใส่ URL, สิ่งที่เราต้องการค้นหา เช่น เพลง , https://www.sanook.com
2. เว็บจะนำข้อมูลจากดาต้าเบสมาแสดงบนหน้าเว็บไซต
3. เมื่อเรากดที่วันเวลา เว็บจะดึงข้อมูลหน้าตาของเว็บเพจที่เก็บไว้มาแสดง



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 และคีย์เวิร์ดค้นหาจากไฟล์ยังฟิกไว้อยู่
-เว็บสามารถแสดงข้อมูลที่เก็บไว้ในดาต้าเบสได้

สิ่งที่ทำไม่ได้
-ระบบรวบรวมข้อมูลที่เป็นแบบMULTI THREADS
-การค้นหาFORMและใส่ข้อมูลลงในFORM รูปแบบอื่นยังทำไม่ได้