



โครงการวิศวกรรมคอมพิวเตอร์

ระบบรวบรวมข้อมูลจาก Deep Web

Deep Web Crawler

โดย

นางสาวกัลยารัตน์

ศรีชัย

นายประพัฒน์

วิริยะรุ่งเรืองชัย

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ศรีราชา

มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา

ปีการศึกษา 2564

โครงการวิศวกรรมคอมพิวเตอร์

เรื่อง

ระบบรวบรวมข้อมูลจาก Deep Web

Deep Web Crawler

โดย

นางสาวกัลยารัตน์ ศรีชัย

นายประพัฒน์ วิริยะรุ่งเรืองชัย

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ศรีราชา

มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา

ปีการศึกษา 2564

คำนำ

โครงการฉบับนี้จัดทำขึ้นเพื่อรวบรวมข้อมูลที่ได้ทำการฝึกงานภาคฤดูร้อน ปีการศึกษา 2564 จากมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา (Kasetsart University Sriracha Campus) โครงการฉบับนี้จะกล่าวถึงระบบการรวบรวมข้อมูลจาก Deep Web กล่าวถึงวิธีการปฏิบัติงานหรือสิ่งที่ได้รับการฝึกงาน มีการบันทึกการปฏิบัติงานในวันต่าง ๆ โดยจะเก็บรายละเอียดของเนื้อหาที่ได้ทำการฝึกงานไว้ให้สามารถดูได้โดยง่ายเป็นระเบียบและตลอดจนประโยชน์ที่ได้รับจากการฝึกงานภาคฤดูร้อน ปีการศึกษา 2564 ซึ่งโครงการได้ทำการสรุปไว้เป็นอย่างดี

โครงการฉบับนี้อาจเป็นประโยชน์ต่อผู้ที่ได้อ่านไม่มากนักน้อย หากมีข้อผิดพลาดประการใดหรือตกหล่นข้อมูลใดก็ขออภัยมา ณ ที่นี้ด้วย

นางสาวกัลยารัตน์ ศรีชัย
นายประพัฒน์ วิริยะรุ่งเรืองชัย

มิถุนายน 2564

สารบัญ

	หน้า
สารบัญภาพ	5
บทที่ 1 บทนำ	6
ที่มาและความสำคัญ	6
เป้าหมายของโครงการ	6
ประโยชน์ของข้อมูลที่ได้จาก Deep Web	6
สถานที่ปฏิบัติงาน	7
ข้อมูลผู้บังคับบัญชา	7
ระบบโครงสร้างบุคคลกรในหน่วยงาน	8
บทที่ 2 การดำเนินงาน	11
งานที่ได้รับมอบหมาย	11
Flowchart อธิบายการทำงานของ Deep Web Crawler	11
โครงสร้างดาต้าเบส	12
Class Diagram	12
ขั้นตอนการพัฒนาระบบ	13
วิธีการทำงานของระบบ	13
รายงานการดำเนินงานประจำวัน	14
ปัญหาที่พบระหว่างฝึกงานและวิธีแก้ปัญหาที่เหมาะสม	18
สิ่งที่ประทับใจ	18
บทที่ 3 สรุป	19
บรรณานุกรม	20

สารบัญภาพ

ภาพที่ 1 ภาพแผนภูมิโครงสร้างองค์กร	8
ภาพที่ 2 ภาพแผนภูมิโครงสร้างการบริหารงาน	8
ภาพที่ 3 ภาพแผนภูมิโครงสร้างอัตรากำลัง (สายวิชาการ).....	9
ภาพที่ 4 ภาพแผนภูมิโครงสร้างอัตรากำลัง (สายสนับสนุน+ศูนย์วิจัยและบริการวิศวกรรม)	9
ภาพที่ 5 ภาพ Flowchart อธิบายการทำงานของ Deep Web Crawler	11
ภาพที่ 6 ภาพโครงสร้างดาต้าเบส	12
ภาพที่ 7 ภาพ Class Diagram	12
ภาพที่ 8 ภาพขั้นตอนการทำงานของ Deep Web Crawler.....	13
ภาพที่ 9 ภาพขั้นตอนการทำงานของ Deep Web Crawler.....	13
ภาพที่ 10 ภาพขั้นตอนการทำงานของเว็บไซต์สำหรับแสดงข้อมูลที่ Crawl.....	14
ภาพที่ 11 ภาพข้อมูลหน้าตาของเว็บเพจที่เก็บไว้มาแสดง	14

บทที่ 1

บทนำ

ที่มาและความสำคัญ

Deep Web เป็นข้อมูลหรือเว็บไซต์ที่ไม่ถูกจัดทำ Index โดย Search Engine มาตรฐานทั่วไป เช่น Google หรือ Yahoo นั้นหมายความว่า ผู้ใช้ทั่วไปจะไม่สามารถค้นหาเว็บไซต์เหล่านั้นได้เจอผ่าน Search Engine แต่ยังคงสามารถเข้าผ่าน URL ได้ตามปกติ เว็บไซต์เหล่านั้นได้แก่ ฐานข้อมูลของผู้ใช้, เว็บไซต์ที่จำเป็นต้อง Login ก่อน, เว็บเมล, เพจที่อยู่ด้านหลัง Firewall เป็นต้น ซึ่งส่วนใหญ่จะเก็บข้อมูลเกี่ยวกับเอกสารการศึกษา บันทึกการแพทย์ เอกสารกฎหมาย รายงานทางวิทยาศาสตร์ ข้อมูลรัฐบาล หรือคลังข้อมูลขององค์กร

ผู้จัดทำเห็นความมีประโยชน์ของ Deep Web จึงต้องการจัดทำโครงการนี้ขึ้นมา

เป้าหมายของโครงการ

1. ออกแบบและพัฒนาระบบ Deep Web โดยใช้ภาษา Python
2. พัฒนา Web Site ที่สามารถเรียกดูข้อมูลจาก Database

ประโยชน์ของข้อมูลที่ได้จาก Deep Web

ข้อมูลที่ได้จาก Deep Web สามารถนำข้อมูลมารวบรวมสร้างเป็น Web Portal ได้ไม่ว่าจะเป็นข้อมูลเกี่ยวกับร้านอาหาร เพลง หนังสือ และซีรี่ย์ภาพยนตร์

1. ข้อมูลที่ได้เกี่ยวกับร้านอาหาร เราสามารถนำข้อมูลเหล่านั้น มาแยกหมวดหมู่ร้านอาหารประเภทต่าง ๆ ได้ เช่น หมวดหมู่อาหาร หมวดหมู่ร้านกาแฟหรือของหวาน หรือผู้ใช้อาจสามารถตรวจสอบได้ว่าร้านอาหารไหนอยู่บริเวณพื้นที่ใกล้เคียงกับผู้ใช้ ทำให้สามารถค้นหาร้านอาหารที่ต้องการได้อย่างสะดวก รวดเร็ว เพื่อความสะดวกสบายแก่ผู้ใช้งาน
2. ข้อมูลที่ได้เกี่ยวกับเพลง เราสามารถนำข้อมูลเหล่านั้น มาจัดสรรแยกหมวดหมู่ของเพลง ไม่ว่าจะเป็นแยกเพลงตามรูปแบบเนื้อหาของเพลง เช่น เพลงเกาหลี เพลงสากล และเพลงไทย หรือแยกเป็นรูปแบบ

เพลงตามจังหวะ เช่น Classic pop jazz R&B Rap Hip hop Rock และอื่น ๆ และแยกหมวดหมู่เพลงตามรายชื่อศิลปินที่ผู้ใช้ชื่นชอบก็ตาม ก็ช่วยให้ผู้ใช้ได้เลือกฟังเพลงที่ตนเองชื่นชอบ ไม่ต้องคอยเสียเวลาค้นหา เพื่อความสะดวกสบายแก่ผู้ใช้งาน

3. ข้อมูลที่ได้เกี่ยวกับหนังสือ เราสามารถนำข้อมูลเหล่านี้มาแยกหมวดประเภทต่าง ๆ เช่น สารคดี บันเทิงคดี สิ่งพิมพ์ ตำรา วารสาร นิตยสาร เป็นต้น หรือจะแยกเป็นหมวดหมู่หนังสือยอดนิยม เพื่อให้ผู้ใช้งานสามารถค้นหาหนังสือที่สนใจ อย่างสะดวก
4. ข้อมูลที่ได้เกี่ยวกับซีรี่ย์ภาพยนตร์ สามารถนำข้อมูลมาแยกประเภทของภาพยนตร์ต่าง ๆ ได้ เช่น ภาพยนตร์ดราม่า ภาพยนตร์วิทยาศาสตร์ ภาพยนตร์ครอบครัว ภาพยนตร์ระทึกขวัญ ภาพยนตร์อาชญากรรม ภาพยนตร์สารคดี ภาพยนตร์การ์ตูน เป็นต้น หรือจะจัดอันดับซีรี่ย์ยอดนิยมประจำสัปดาห์ หรือจะแยกเป็นหมวดหมู่ภาพยนตร์ไทย ซีรี่ย์เกาหลี ซีรี่ย์จีนก็ได้ เพื่อความหลากหลายของผู้ใช้

สถานที่ปฏิบัติงาน

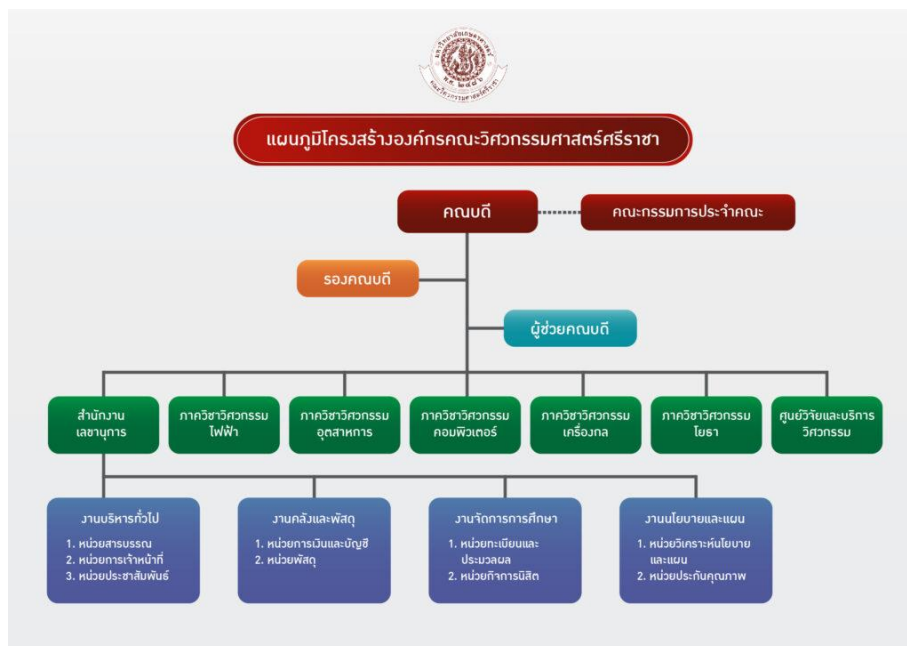
Work from home.

ข้อมูลผู้บังคับบัญชา

ผู้ที่ทำการควบคุมและสอนงานให้กับข้าพเจ้าคือ ผศ.ดร.กุลวดี สมบูรณ์วิวัฒน์ มีตำแหน่งเป็นผู้ช่วยศาสตราจารย์ โดยปฏิบัติงานอยู่ที่ คณะวิศวกรรมศาสตร์ศรีราชา มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา อีเมลที่สามารถติดต่อได้ kulwadee@eng.src.ku.ac.th

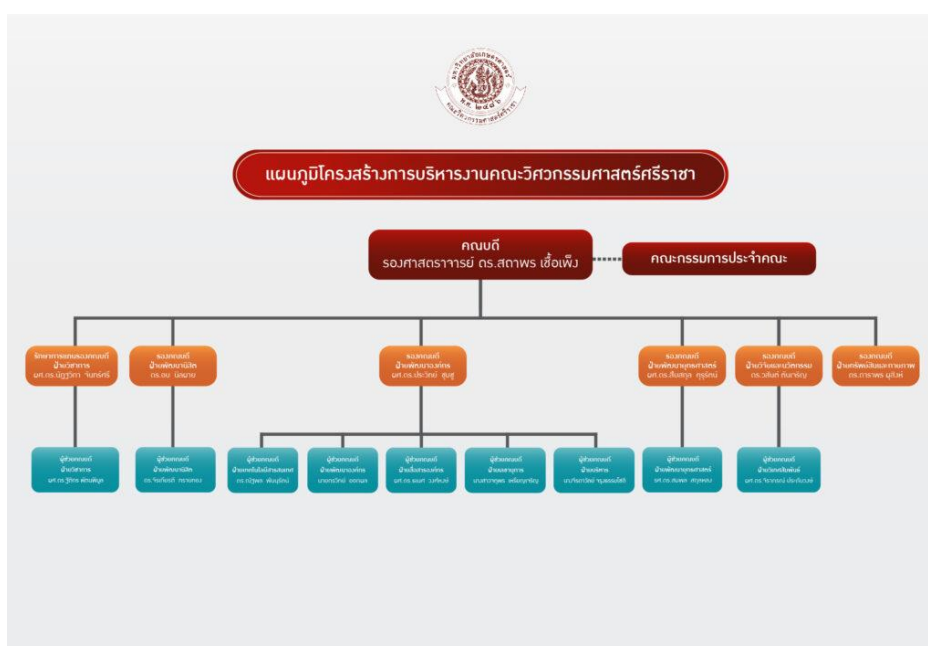
ระบบโครงสร้างบุคคลกรในหน่วยงาน

แผนภูมิโครงสร้างองค์กร



ภาพที่ 1 ภาพแผนภูมิโครงสร้างองค์กร

แผนภูมิโครงสร้างการบริหารงาน



ภาพที่ 2 ภาพแผนภูมิโครงสร้างการบริหารงาน

จากแผนภาพดังกล่าว บ่งบอกถึงระบบโครงสร้างภายในมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา โดยมีการไล่ระบบจากตำแหน่งสูงสุดลงมาตำแหน่งต่ำกว่า และมีการใช้สีที่แตกต่างกันเพื่อให้สามารถสื่อสารมองเห็น และเข้าใจง่าย

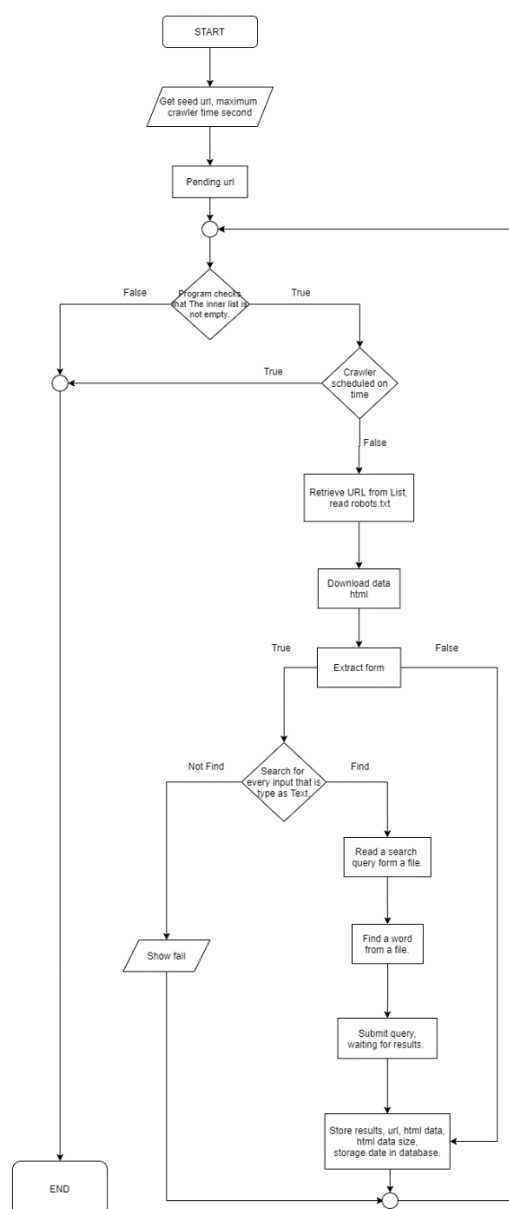
บทที่ 2

การดำเนินงาน

งานที่ได้รับมอบหมาย

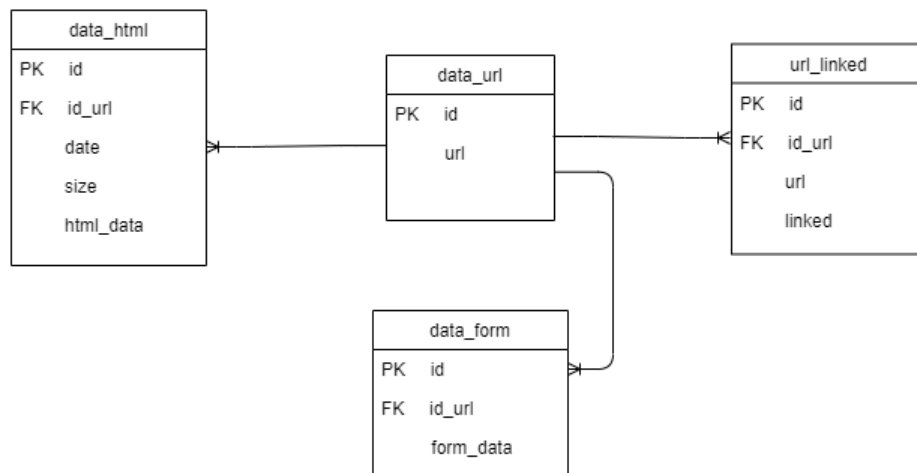
ได้รับมอบหมายงานให้พัฒนาระบบ Deep Web และสร้าง Web Site สำหรับอ่านข้อมูลในดาต้าเบส

Flowchart อธิบายการทำงานของ Deep Web Crawler



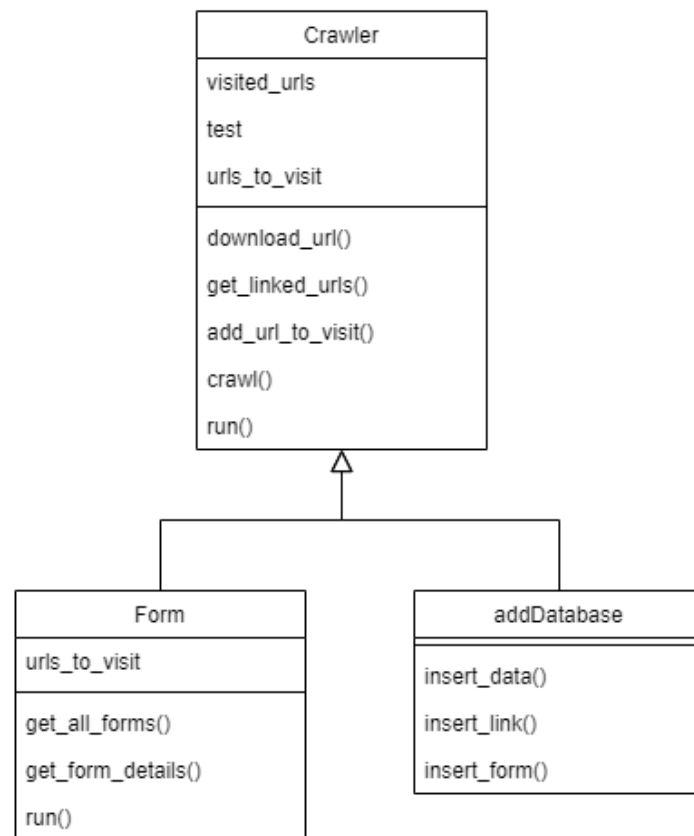
ภาพที่ 5 ภาพ Flowchart อธิบายการทำงานของ Deep Web Crawler

โครงสร้างดาต้าเบส



ภาพที่ 6 ภาพโครงสร้างดาต้าเบส

Class Diagram



ภาพที่ 7 ภาพ Class Diagram

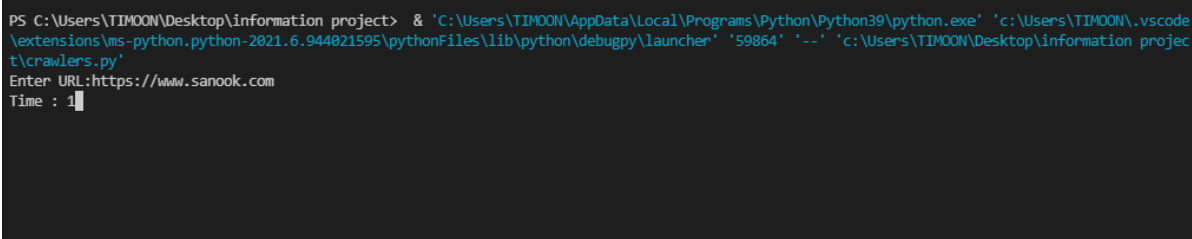
ขั้นตอนการพัฒนาระบบ

1. หา code basic ของ crawler มาเป็นขั้นเริ่มต้น
2. ทำความเข้าใจการทำงานของโค้ด และเขียนออกมาเป็นflowchart
3. ทำการ Crawler ข้อมูลในเว็บไซต์ และดึงข้อมูลเก็บลงในดาต้าเบส
4. เขียนโปรแกรมให้สามารถค้นหาคำค้นหาจากไฟล์ได้แบบอัตโนมัติ และดึงข้อมูลเก็บลงในดาต้าเบส
5. สร้างโครงสร้างดาต้าเบส
6. สร้าง Class Diagram
7. พัฒนาเว็บไซต์เพื่อแสดงข้อมูลในดาต้าเบส

วิธีการทำงานของระบบ

ขั้นตอนการทำงานของ Deep Web Crawler

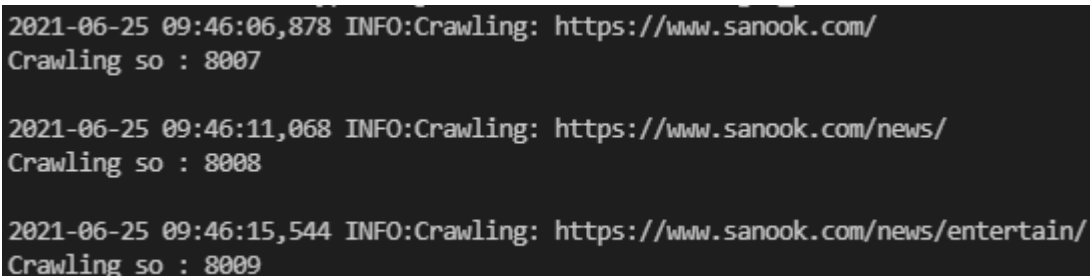
1. ใส่ URLและเวลา(ชั่วโมง) ที่ต้องการ Crawl



```
PS C:\Users\TIMOON\Desktop\information project> & 'C:\Users\TIMOON\AppData\Local\Programs\Python\Python39\python.exe' 'c:\Users\TIMOON\.vscode\extensions\ms-python.python-2021.6.944021595\pythonFiles\lib\python\debugpy\launcher' '59864' '--' 'c:\Users\TIMOON\Desktop\information project\crawlers.py'
Enter URL:https://www.sanook.com
Time : 1
```

ภาพที่ 8 ภาพขั้นตอนการทำงานของ Deep Web Crawler

2. โปรแกรมจะทำการ Crawl url, data ของ html, size ของ html, form ของ html, วันเวลาที่ Crawl เก็บลงในดาต้าเบส



```
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
```

ภาพที่ 9 ภาพขั้นตอนการทำงานของ Deep Web Crawler

3. โปรแกรมจะหยุดทำงานตามเวลาที่ใส่ไว้ตามขั้นตอนที่ 1 หรือ ไม่มีข้อมูลให้ Crawl

รายงานการดำเนินงานประจำวัน

วัน/เดือน/ปี	งานที่ได้รับมอบหมาย
17/พ.ค./64	<ul style="list-style-type: none"> ● ได้รับมอบหมายงานให้ศึกษาข้อมูลเกี่ยวกับ Deep Web Crawler ● ศึกษา Paper ที่อาจารย์ได้มอบหมายให้
18/พ.ค./64	ได้รับมอบหมายงานให้ค้นคว้า Code เกี่ยวกับ Deep Web Crawler โดยใช้ภาษา Python
19/พ.ค./64	<ul style="list-style-type: none"> ● นำเสนอข้อมูลที่ไปศึกษามา ● และ Code ที่ไปศึกษาค้นคว้ามา
20/พ.ค./64	ได้รับมอบหมายงานให้ศึกษา Code ที่ค้นคว้ามา
21/พ.ค./64	<ul style="list-style-type: none"> ● นำเสนอ Code ที่ศึกษามา ● และได้รับมอบหมายให้ Crawler URL ของเว็บไซต์
24/พ.ค./64	<ul style="list-style-type: none"> ● นำเสนอการ Crawler URL ของเว็บไซต์ ● ได้รับมอบหมายงานให้ดึงข้อมูลของเว็บไซต์
25/พ.ค./64	<ul style="list-style-type: none"> ● ได้รับมอบหมายงานให้ดึงข้อมูลของเว็บไซต์
26/พ.ค./64	<ul style="list-style-type: none"> ● นำเสนองานที่ดึงข้อมูลของเว็บไซต์ ● ได้รับมอบหมายงานให้เก็บข้อมูลลงในดาต้าเบส
27/พ.ค./64	เก็บข้อมูลที่ดึงจากเว็บไซต์ลงในดาต้าเบส MySQL
28/พ.ค./64	<ul style="list-style-type: none"> ● แสดงข้อมูลที่เก็บลงดาต้าเบส MySQL ● ได้รับคำแนะนำจากอาจารย์ให้ใช้ดาต้าเบส PostgreSQL เพราะสะดวกและใช้งานง่ายกับการใช้ภาษา Python

31/พ.ค./64	ได้รับมอบหมายงานให้โปรแกรมค้นหาเป็นคีย์เวิร์ดคำอื่นได้
1/มิ.ย./64	<ul style="list-style-type: none"> ● ทำให้โปรแกรมสามารถค้นหาคำค้นหาจากไฟล์ Text ● เริ่มเขียน Flowchart การทำงานของ Deep Web
2/มิ.ย./64	<ul style="list-style-type: none"> ● ทำการเขียนโปรแกรมให้ค้นหา input ทุกตัวที่มีชนิดเป็น Text ● แก้ไข Flowchart
3/มิ.ย./64	แก้ไขโปรแกรมให้ทำการค้นหาคีย์เวิร์ดเพียงรอบเดียว ถ้าหากเป็นกรณีหน้าเว็บไซต์เดียวกัน
4/มิ.ย./64	<ul style="list-style-type: none"> ● ได้รับมอบหมายงานให้สร้างโครงสร้างของดาต้าเบส ● เก็บข้อมูลลงดาต้าเบส 1000 เว็บ
7/มิ.ย./64	<ul style="list-style-type: none"> ● ได้รับมอบหมายงานให้แก้ไข Flowchart การทำงานของ Deep Web Crawler ● อธิบายฐานข้อมูลโครงสร้างดาต้าเบส ● พิกคำค้นหาแล้วให้โปรแกรมค้นหาซื้อร้านอาหารมาได้แบบอัตโนมัติ
8/มิ.ย./64	<ul style="list-style-type: none"> ● แก้ไข Flowchart การทำงานของ Deep Web Crawler ● สร้างโครงสร้างดาต้าเบส ● พัฒนาระบบให้สามารถค้นหาได้แบบอัตโนมัติ
9/มิ.ย./64	<ul style="list-style-type: none"> ● อ่านไฟล์ Robot.text ก่อน Crawler ทุกครั้ง ● และใส่ขั้นตอนอ่านไฟล์ลงใน Flowchart
10/มิ.ย./64	เริ่มทำการเก็บข้อมูลเว็บไซต์อื่น
11/มิ.ย./64	แก้ไข Flowchart
14/มิ.ย./64	แก้ไข Flowchart ใส่ดีเลย์ก่อน Crawler

15/มิ.ย./64	ได้รับมอบหมายงานให้สร้างเว็บไซต์
16/มิ.ย./64	เพิ่มเงื่อนไขในโปรแกรม หยุดเมื่อ Crawler ได้ครบจำนวนหน้าที่กำหนด
17/มิ.ย./64	โปรแกรมหยุดเมื่อครบเวลาที่เรากำหนด
18/มิ.ย./64	ได้รับมอบหมายงานให้แก้ไขโปรแกรม
21/มิ.ย./64	<ul style="list-style-type: none"> ● ได้รับมอบหมายงานให้แก้ไข โครงสร้างดาต้าเบส ● และดึงข้อมูลจากดาต้าเบสลงบน Web Site
22/มิ.ย./64	ได้รับมอบหมายงานให้ดึงข้อมูลจากดาต้าเบสลงบน Web Site
23/มิ.ย./64	ดึงข้อมูลจากดาต้าเบสลงบน Web Site
24/มิ.ย./64	ได้รับมอบหมายงานให้ทำรายงาน पोस्เตอร์
25/มิ.ย./64	นำเสนอโครงการ Deep Web Crawler

ปัญหาที่พบระหว่างฝึกงานและวิธีแก้ปัญหที่เหมาะสม

จากการฝึกงานนั้นช่วงเวลาในการฝึกฝนได้เจอปัญหาหนึ่งคือการจะดึงข้อมูลจากเว็บไซต์ต่าง ๆ แล้วพบว่าบางเว็บไซต์ไม่อนุญาตให้ใช้ข้อมูลบางส่วน ซึ่งในตอนแรกทางผู้จัดทำไม่ทราบว่าต้องทำการศึกษาไฟล์ Robot.txt มารายทของการจะดึงข้อมูลเว็บไซต์คือ หากเราจะดึงข้อมูลของเว็บไซต์ใด ให้ทำการอ่านไฟล์ Robot.txt ของเว็บไซต์นั้นเสียก่อน (ไฟล์ Robot.txt คือ ข้อตกลงที่เว็บไซต์จะบอกเราว่าอนุญาตให้เราใช้ข้อมูลใด และไม่อนุญาตให้ใช้ข้อมูลใด) ทางเราได้ทำการพยายามจะดึงข้อมูลเว็บไซต์ Wongnai โดยไม่ได้ศึกษาไฟล์ Robot.txt ผลสรุปว่าดึงข้อมูลได้มา 10 ข้อมูล ทางผู้จัดทำโดนเว็บไซต์ Wongnai บล็อก ทำให้ไม่สามารถดึงข้อมูลได้อีก และหลังจากนั้นผู้จัดทำจะดึงข้อมูลของเว็บไซต์ใดก็จะทบทวนการอ่านศึกษาไฟล์ Robot.txt ก่อนทุกครั้งมาตลอด จึงจะไม่โดนบล็อกหรือปิดกั้นอีก

สิ่งที่ประทับใจ

จากการฝึกงานครั้งนี้ ข้าพเจ้าได้รับความรู้ความแนะนำจากอาจารย์มาโดยตลอด เมื่อเจอปัญหาในสิ่งที่ไม่เคยรู้มาก่อน ก็มีอาจารย์ที่เคยช่วยบอกช่วยสอน ต้องขอขอบคุณเป็นอย่างยิ่ง

บทที่ 3

สรุป

Deep Web เป็นเว็บไซต์ที่ผู้ใช้ทั่วไปจะไม่สามารถค้นหาเว็บไซต์เหล่านั้นได้เจอผ่าน Search Engine แต่ยังคงสามารถเข้าผ่าน URL ได้ตามปกติ ประโยชน์ของ Deep Web นอกจากจะสามารถเข้าถึงได้เฉพาะบุคคล บุคลากรขององค์กรนั้นๆ เพื่อความปลอดภัยของฐานข้อมูลของผู้ใช้, หรือจะเก็บข้อมูลเกี่ยวกับเอกสาร การศึกษา บันทึกการแพทย์ เอกสารกฎหมาย รายงานทางวิทยาศาสตร์ ข้อมูลรัฐบาลแล้ว ประโยชน์ของมันยังสามารถที่จะรวบรวมข้อมูลต่าง ๆ เพื่อสร้างขึ้นเป็น Web Portal ได้เพื่อความสะดวกสบายแก่ผู้ใช้งาน

การจัดทำโครงการนี้ขึ้น เพื่อสามารถนำประโยชน์ของ Deep Web Crawler มาใช้ให้เกิดประโยชน์สูงสุด และเพื่ออำนวยความสะดวกสบายของมนุษย์

บรรณานุกรม

Techtalkthai. 21 กุมภาพันธ์ พ.ศ. 2559. เราจะค้นหาข้อมูลกว่า 96% บนโลกอินเทอร์เน็ตใน Deep Web ได้อย่างไร. แหล่งที่มา: <https://www.techtalkthai.com/how-to-search-websites-in-deep-web/#:~:text=Deep%20Web%20เป็นข้อมูลหรือ,ที่จำเป็นต้อง%20Login%20ก่อน%2C,24 มิถุนายน 2564>

คณะวิศวกรรมศาสตร์ศรีราชา. โครงสร้างองค์กรวิศวกรรมศาสตร์ศรีราชา. แหล่งที่มา: https://www.eng.src.ku.ac.th/th/โครงสร้างองค์กร/?fbclid=IwAR0VB9rk9mGcmp7CjflxEn5Et6swPabHXjSgP867tYKSQ-30Tall_rSVdvc,24 มิถุนายน 2564