รายงาน

การค้นหารายชื่อวัดโดยการใช้ Regular Expression

จัดทำโดย

กลุ่ม
ใครจะเพิ่มเดี๋ยวเตะภูออกให้

สมาชิกกลุ่ม
63010139 จิรภัทร แก้วส่งส่งแสง
63010177 ชญานิน เลียงจินดาถาวร
63010231 ชินพัฒน์ ศิริยาใจ
63010256 ฐานพัฒน์ สิทธิพรชัยสกุล
63010279 ณภัทร จิรารัตนกุลชัย
63010382 ทิวัตถ์ โพธิ์ครี
63010445 ธรณินท์ พงษ์สถิตย์พร
63010522 นาวีวัฒน์ พฤกพัฒนาชัย
63010548 บูรพา ทิมแดง
63010630 พชรพล จารุวนาวัฒน์
63010750 ภาสกร คงบุญเกียรติ

เสนอ
รศ.ดร.เกียรติกูล เจียรนัยธนะกิจ

รายงานนี้เป็นส่วนหนึ่งของ
วิชา Theory of Computer
รหัสวิชา 01076013
ภาคเรียนที่ 2 ปีการศึกษา 2565

สรุป Regular Expression

จังหวัดที่ทำ web crawler

1. จังหวัดนครศรีธรรมราช
2. วัดในจังหวัดนครสวรรค์
3. วัดในจังหวัดนนทบุรี
4. วัดในจังหวัดนราธิวาส
5. วัดในจังหวัดอุทัยธานี

พบว่าจะมี 2 pattern คือ 1. ใช้ Table 2. ไม่ใช้ Table

1. แบบใช้ Table (นนทบุรี)

```
▶ <div id="contentSub">…</div>
```

```
<a href="/wiki/%E0%B8%9B%E0%B8%A3%E0%B8%B0%E0%B9%80%E0%B8%97%E0%B8%A8%E0%B9%84%E0%B8%97%E0%B8%A2" title="ประเทศไทย">ประเทศไทย</a>
"แบ่งตามจังหวัด"
```

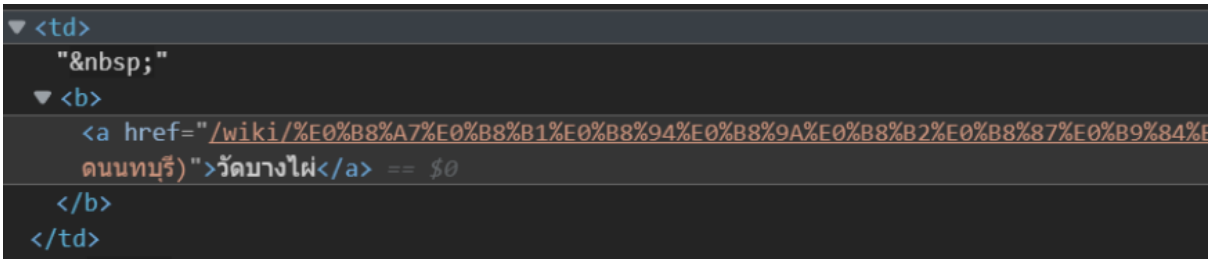
ชื่อวัดระหว่าง Tag div และ Tag a

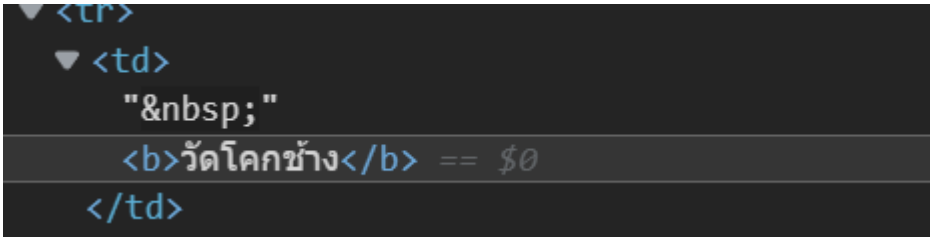
RE 1 :

```
re.compile('<div id=[\u0000-\uFFFF]*title="ประเทศไทย">ประเทศไทย</a>แบ่งตามจังหวัด</div>')
```

คำอธิบาย : \u0000-\uFFFF คือ unicode ของ string ทั้งหมด Regular Expression [\u0000-\uFFFF]* จะนำตัวอักษรทั้งหมดที่อยู่ระหว่าง ประเทศไทยแบ่งตามจังหวัด มา

มี pattern เป็น <div id= …….title=”ประเทศไทย>ประเทศไทย แบ่งตามจังหวัด





RE 2 :

```
re.compile('<td>.*</td>')
```

คำอธิบาย : เป็นการกรองเอาเฉพาะข้อมูลที่อยู่ภายใน table ของหน้าเว็บ ซึ่งอยู่ระหว่าง <td> กับ </td> เท่านั้น .* คือเอาตัวอักษรซ้ำกี่ก็ได้ (ตัวอักษรทั้งหมด)

มี pattern เป็น <td>....</td>

RE 3 :

```
re.compile('>วัด[\u0E01-\u0E5B]*')
```

คำอธิบาย : ใน tag td ตัวอักษรของวัดที่ต้องการเริ่มต้นจะเป็น “>” เพราะใน tag td จะมีสิ่งที่ไม่ต้องการขึ้นต้นคำว่า “วัด” โดย \u0E01-\u0E5B เป็น unicode ของภาษาไทย ซึ่ง [\u0E01-\u0E5B]* คือเอาตัวอักษรภาษาไทยทั้งหมด ตัดคำหลังจากตัวอักษรภาษาไทยตัวสุดท้าย

มี pattern เป็น >วัด….ตามด้วยภาษาไทย….

RE 4 :

```
re.sub(">", "", i)
```

คำอธิบาย : หลังจากที่กรองออกมาแล้ว จะพบว่าวัดที่กรองได้มาจะมีเครื่องหมาย “>” ติดออกมาด้วย ซึ่งเป็นสิ่งที่ไม่ต้องการ ทำให้ต้องมาลบเครื่องหมายออก ด้วยคำสั่ง re.sub

2. ไม่ใช้ Table (อุทัยธานี,นราธิวาส,นครศรีธรรมราช,นครสวรรค์)

RE 1 ใช้

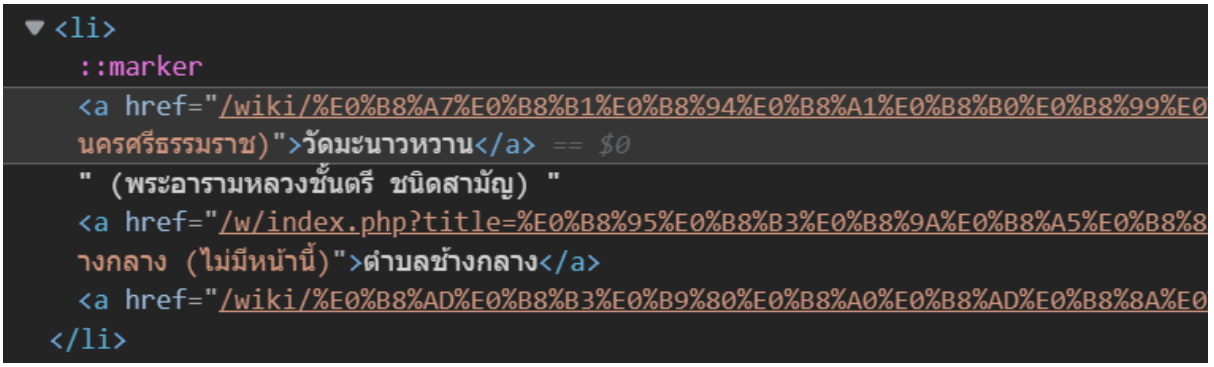
```
<main id="content" class="mw-body" role="main">
<span class="mw-headline" id="ดูเพิ่ม">ดูเพิ่ม</span>
```

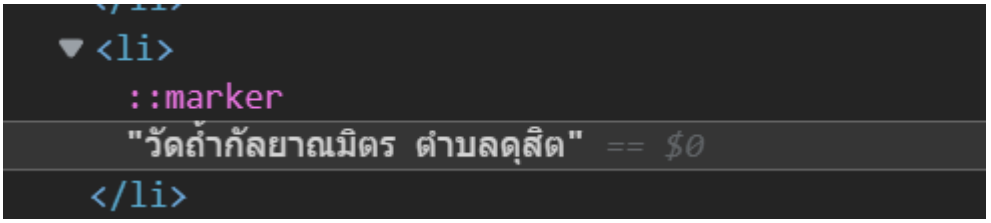
ชื่อวัดระหว่าง Tag main และ Tag span

```
re.compile('<main[\u0000-\uFFFF]*id="ดูเพิ่ม">ดูเพิ่ม</span>')
```

คำอธิบาย : \u0000-\uFFFF คือ unicode ของ string ทั้งหมด Regular Expression [\u0000-\uFFFF]* จะนำตัวอักษรทั้งหมดที่อยู่ระหว่าง <main กับ id=”ดูเพิ่ม”>ดูเพิ่ม”

มี pattern เป็น <main…….id=”ดูเพิ่ม”>ดูเพิ่ม





Note: “” มีตอนกด inspect ในตอน request จริงไม่มีใน request

RE 2 ใช้

```
re.compile('<li>.*</li>')
```

คำอธิบาย :เป็นการกรองเอาเฉพาะข้อมูลที่อยู่ภายใน table ของหน้าเว็บ ซึ่งอยู่ระหว่าง <li> กับ เท่านั้น .* คือเอาตัวอักษรซ้ำกี่ก็ได้ (ตัวอักษรทั้งหมด)

มี pattern เป็น <li>....

RE 3 :

```
re.compile('>วัด[\u0E01-\u0E5B]*')
```

คำอธิบาย : คำอธิบาย : ใน tag td ตัวอักษรของวัดที่ต้องการเริ่มต้นจะเป็น “>” เพราะใน tag td จะมีสิ่งที่ไม่ต้องการขึ้นต้นคำว่า “วัด” โดย \u0E01-\u0E5B เป็น unicode ของภาษาไทย ซึ่ง [\u0E01-\u0E5B]* คือเอาตัวอักษรภาษาไทยทั้งหมด ตัดคำหลังจากตัวอักษรภาษาไทยตัวสุดท้าย

มี pattern เป็น >วัด….ตามด้วยภาษาไทย….

RE 4 :

```
re.sub(">", "", i)
```

คำอธิบาย : หลังจากที่กรองออกมาแล้ว จะพบว่าวัดที่กรองได้มาจะมีเครื่องหมาย “>” ติดออกมาด้วย ซึ่งเป็นสิ่งที่ไม่ต้องการ ทำให้ต้องมาลบเครื่องหมายออก ด้วยคำสั่ง re.sub