

1 ภาพรวมและจุดประสงค์

หนังสือพิมพ์ไทยนิยมใช้คำศัพท์ที่ค่อนข้างประหลาดเพื่อให้ดึงดูดกว่า เช่น ตอนที่พูดถึงประเทศญี่ปุ่น ชอบใช้คำว่า “แดนปลาดิบ” ในโครงการนี้ วิเคราะห์ว่ามีศัพท์บ่งชี้ประเทศอะไรบ้าง โดยใช้ machine learning

กระบวนการส่วนคือ

1. เก็บข้อมูล (headline, description, article) จาก **ไทยรัฐ** (<https://www.thairath.co.th>) โดยใช้ beautiful soup และบันทึกไว้เป็น .tsv
2. เอาบทความเกี่ยวกับ 5 ประเทศ: ญี่ปุ่น จีน อเมริกา เยอรมัน เกาหลี
3. ตัดเป็นคำโดยใช้ tltk แล้วฝึกด้วย headline description หรือ article โดยใช้ sklearn
4. วัด F-score และให้แสดง feature ที่ parameter สูง

ตอนแรกเลือก matichon แต่ใช้คำที่เป็นทางการ เพราะฉะนั้นเปลี่ยนเป็นเก็บข้อมูลจากไทยรัฐแทน เก็บข้อมูลมาทั้งหมด มากกว่าแสนบทความ ส่วนใหญ่ไม่เกี่ยวกับประเทศ แต่บันทึกไว้ใน thairath.tsv เพื่อการวิจัยในอนาคต

2 วิธีเก็บข้อมูล

เว็บไซต์ไทยรัฐไม่ให้ API จึงต้องตัดแท็กด้วย beautiful soup

URL ของบทความในเว็บไซต์เป็น [https://www.thairath.co.th/content/\(7เลข\)/](https://www.thairath.co.th/content/(7เลข)/) แต่บางเลขไม่มีบทความ เพราะฉะนั้นเขียนโปรแกรมว่า request.get แล้ว ถ้า status code == 200 จะเอา html มา โครงสร้าง html เป็นอย่างนี้

```
<script><script type="application/ld+json" async="" class="next-head">{
  "@context": "http://schema.org",
  "@type": "NewsArticle",
  "headline": "รู้ยัง? เดือนหน้า คนไทยไม่ต้องพกใบขับขี่ ดร.เรียกดูง่าย คนขับเช็คใบสั่งสบาย",
  "alternativeHeadline": "รู้ยัง? เดือนหน้า คนไทยไม่ต้องพกใบขับขี่ ดร.เรียกดูง่าย คนขับเช็คใบสั่งสบาย",
  "image": [
    "https://www.thairath.co.th/media/4DQpjUtzLUwmJZZPFiS41hpWPlr0uaBAvVVKFNK1sav4.jpg"
  ],
  "datePublished": "2018-12-16T19:05:00+07:00",
  "description": "เดือนหน้า คนไทยไม่ต้องพกใบขับขี่ ขนส่งทางบกปีงไอเดียดูข้อมูลลงแอปฯ ใช้แสดงให้เจ้าหน้าที่ดูเมื่อถูกเรียกตรวจ โสเทคมระบบแจ้งเตือนบัตรหมดอายุ เ",
  "articleBody": "เดือนหน้า คนไทยไม่ต้องพกใบขับขี่ ขนส่งทางบกปีงไอเดียดูข้อมูลลงแอปฯ ใช้แสดงให้เจ้าหน้าที่ดูเมื่อถูกเรียกตรวจ โสเทคมระบบแจ้งเตือนบัตรหมดอายุ เ",
  "author": "Thairath",
  "publisher": {
    "@type": "Organization",
    "name": "ไทยรัฐออนไลน์",
    "logo": {
      "@type": "ImageObject",
      "url": "https://www.thairath.co.th/_next/static/images/online-logo-47b15ae139ec1909179a41203b0cd550.png"
    }
  },
  "dateModified": "2018-12-16T19:05:00+07:00",
  "mainEntityOfPage": {
    "@type": "WebPage",
    "@id": "https://www.thairath.co.th/content/1446488"
  }
}</script><link rel="preload" href="/_next/static/ENsnhDCjAEN2AxUFNDzIA/pages/content.js" as="script"><link rel="preload" href=
```

ใน <script> ~ </script> มี headline description และ articleBody จึงเก็บเหล่านี้มาเป็น json และใช้ json.loads() ทำให้เป็น dict และบันทึกไว้ใน thairath.tsv

thairath.tsv มีบทความทั้งหมด 105,923 ข้อ (ID: 100000-1370000) แต่บทความที่ใช้จริงไม่มาก ดึงบทความที่มีคำว่า ญี่ปุ่น อเมริกา ประเทศจีน เยอรมัน เกาหลี มาบันทึกเป็นไฟล์ใหม่ country.tsv สุดท้ายให้เป็น tokenized text และบันทึกเป็น headline.tsv

```
thairath.co.th --get()--> thairath.tsv --find_article()--> country.tsv --tokenize_headline()--> headline.tsv
```

นอกจากนี้ เก็บข้อมูลมาเพิ่ม 5558 ข้อ (ID: 1400000-1420000) เพื่อประเมิน model โดยบันทึกเป็น thairath_test.tsv, country_test.tsv, headline_test.tsv เช่นเดียวกัน

3 ML ด้วยหัวข้อ

ข้อมูลที่เก็บมาต้องตัดเป็นคำ ในโครงการนี้ใช้ `tltk.nlp.word_segment()` เพื่อตัด ตอนแรกให้ฝึกโดยใช้เนื้อหาบทความ (article) แต่ใช้เวลามากเกินไปจึงใช้แค่ headline อย่างเดียว จำนวนข้อมูลที่ใช้คือ

	ญี่ปุ่น	จีน	อเมริกา	เยอรมัน	เกาหลี
การฝึก: headline.tsv	5662	2047	4559	1036	3902
การประเมิน: headline_test.tsv	293	103	218	54	146

Accuracy, precision, recall, F-score เป็นดังนี้

Accuracy
0.5343980343980343

Confusion Matrix
[[181 14 66 1 31]
[32 42 21 0 8]
[48 13 133 2 22]
[17 0 14 18 5]
[48 7 30 0 61]]

Report

	precision	recall	f1-score	support
0	0.56	0.62	0.58	293
1	0.55	0.41	0.47	103
2	0.50	0.61	0.55	218
3	0.86	0.33	0.48	54
4	0.48	0.42	0.45	146
avg / total	0.55	0.53	0.53	814

จำนวนข้อมูลที่ใช้การฝึกไม่ค่อยมากเท่าไร เพราะฉะนั้น ทั้ง Accuracy และ F-score ไม่ถึง 60%

4 ผลการวิจัย

แม้ว่า F-score ไม่ค่อยสูงก็ตาม บางส่วนสามารถเดาถูก ที่นี้ ให้แสดง Top 100 feature ของแต่ละ label เป็นผลการวิจัย

ญี่ปุ่น

'ยุ่น', 'ชิ', 'โตเกียว', 'เซเนกัล', 'โตโยต้า', 'ซัป', 'หาด', 'ตะ', 'ทิศทาง', 'ปล้น', 'แปลง', 'พระราม', 'เนียส', 'BNK48', 'ดิบ', 'บัลลังก์', 'ฮา', 'ลุม', 'ชั้นนำ', 'ธีร์', 'ชนา', 'สื่อสาร', 'เจ', 'ซากุระ', 'โนะ', 'ซามูไร', 'ปิ่น', 'ทวาย', 'SUZUKI', 'อลป.', 'SPCG', 'มหัศจรรย์', 'TOYOTA', 'พิรุณ', 'ตูน', 'วะ', 'มาร์ค', 'สร้างสรรค์', 'ศุภ', 'NISSAN', 'หม่อง', 'เมียน', 'โครงสร้าง', 'เทิร์น', 'ปิ', 'นุ่น', 'มัย', 'จูเนียร์', 'ผู้เชี่ยวชาญ', 'สัน', 'วาด', 'แม้ว', 'นิส', 'เมย์', 'เสร็จ', 'สกี', 'ศุภ', 'เรโซ', 'ฟาร์ม', 'วิสาขบูชา', 'สวดมนต์', 'ANA', 'รีบ', 'ซูเปอร์', 'จ้าว', 'กลม', 'เบลเยียม', 'พริ้ม', 'โรด', 'เนะ', 'ฮู', 'กานต์', 'อ้อน', 'ไปรษณีย์', 'กะ', 'เดวิส', 'ผสม', 'ธัน', 'อิชิต', 'ติว', 'ปล้ำ', 'บาย', 'ญี่ปุ่น', 'มิทซูบิชิ', 'แผนการ', 'AV', 'กิ', 'เฟล', 'โนเบล', 'ท้องถิ่น', 'ฮอกไกโด', 'ฟ้าแลบ', 'ศิลป์', 'ผอม', 'โอกินาวา', 'เกี่ยว', 'อิม', 'บัต', 'นครบาล', 'เล่า'

จีน

'ทุเรียน', 'หมา', 'อู๋ฮั่น', 'อาลี', 'ลุ่ม', 'สมุนไพรร', 'ไซ', 'กองทัพ', 'กาญจนา', 'รัล', 'น้ำป่า', 'เซียงไฮ้', 'ท่วม', 'ปักกิ่ง', 'ฝน', 'WGP 2017', 'พรรค', 'โพน', 'ปริณทล', 'พายู', 'เมอร์', 'ไม้', 'สรรเสริญ', 'บัว', 'มังกร', 'อิตาลี', 'ทางโจว', 'คอก', 'สะกิด', 'ฉ.', 'มาเก๊า', 'ขนุน', 'ทักษิณ', 'มติ', 'แพนด้า', 'ตี้', 'ย้อนหลัง', 'มิลค์', 'สเวลด', 'ตรุษจีน', 'ยู', 'คอย', 'สะเดา', 'ศฤงคาร', 'ยาง', 'MH370', 'ตู้', 'พลั่ว', 'ดำน้า', 'ศก.', 'ว้าว', 'กระจก', 'โขง', 'สงขลา', 'ดีเปรสชัน', 'ขวัญ', 'ถนน', 'ถึงแก่นนิจกรรม', 'คลื่นลม', 'พิพากษา', 'จม', 'ตึก', 'องศา', 'ข้าทะเล', 'แอร์', 'อยุธยา', 'อ่างทอง', 'หนาว', 'ตอน', 'จยย.', 'ประกาย', 'สิทธิชัย', 'อี่', 'ลอต', 'ลำไย', 'ร่าง', 'เอิร์น', 'อาชีพ', 'ขยะ', 'ปัก', 'กลอ', 'แผนงาน', 'อื้อ', 'ช่อ', 'ตัดสินใจ', 'แคะ', 'ขบวน', 'ริงนก', 'เซอร์เบีย', 'บังคับการ', 'มะม่วง', 'เดียร์', 'อ่าวไทย', 'หนูน้อย', 'ต้า', 'แปรปรวน', 'รับมือ', 'ตลิ่ง', 'ปี่', 'องค์กร'

อเมริกา

'อเมริกา', 'สหรัฐ', 'ศรีสะเกษ', 'ฟลอยด์', 'ไอซี', 'นิวยอร์ก', 'สำรวจ', 'พู', 'ดีเอ็นเอ', 'ช.', 'ลาออก', 'มะ', 'เณร', 'เปรู', 'USA', 'ธนบัตร', 'บาร์', 'ไดโนเสาร์', 'แหลม', 'ท้า', 'สหรัฐอเมริกา', 'ดวงตา', 'ซีซ', 'ดอน', 'นศ.', 'ซีลี', 'ดวงจันทร์', 'สัญญาณ', 'อเมริกัน', 'ชนิด', 'ข้อสรุป', 'เขต', 'มิก', 'แคลิฟอร์เนีย', 'เสก', 'น้ำผึ้ง', 'อาคาร', 'ปารีส', 'เปลี่ยนแปลง', 'รี', 'ทวีป', 'พีช', 'มหิมา', 'คริส', 'ละเมียด', 'สามารถ', 'แคนาดา', 'ตัวละคร', 'สูญเสีย', 'ภาวะ', 'แซม', 'ไขมัน', 'นีย', 'แอปเปิล', 'ป้องกัน', 'จำ', 'แข่งค่า', 'มนุษย์', 'หรรษา', 'ป้อม', 'ลำไส้', 'โบอิง', 'พม.', 'กราด', 'นาซา', 'สุริยคราส', 'อริ', 'กพท.', 'มหาสมุทร', 'ตอ', 'เดือ', 'หลักสูตร', 'ได้นำ', 'ปวดหัว', 'พฤษภดี', 'AI', 'กัฒมาลา', 'ดาวเคราะห์', 'ศูนย์กลาง', 'แคไหนด', 'อาลา', 'วิทย', 'ออสการ์', 'เจ้', 'แซม', 'คอลเลจ', 'ริญ', 'เอาอยู่', 'กติก', 'เด็ด', 'เคน', 'นายพล', 'สินสอด', 'ยาวนาน', 'สารพัน', 'ทศวรรษ', 'หอม', 'ตรวจเลือด', 'การศึกษา', 'เมิน'

เยอรมัน

'หงส์', 'คลอปป์', 'ซีล', 'ฮิตเลอร์', 'MERCEDES', 'ฟิลิป', 'ปืนใหญ่', 'ผู้ดี', 'ราชวงศ์', 'ฝรั่ง', 'เปียร์', 'เยอรมนี', 'พรหม', 'เอล', 'ไอน์สไตน์', 'ยุโรป', 'มัลลาย', 'บุณ', 'ฟินแลนด์', 'เมธิ', 'ภูเก็ด', 'เก', 'ทูนวอก', 'จิ้งจอก', 'ปืน', 'เสือ', 'PORSCHÉ', 'นู', 'อุส', 'เป็ป', 'รัชสมัย', 'ชาน', 'ลงตัว', 'เทล', 'โรล', 'เยิร์น', 'อุตสาหกรรม', 'BMW', 'ฮูธ', 'เลอร์', 'อลิส', 'แลก', 'คาร์ล', 'ซิป', 'ฝึกงาน', 'เล', 'ปิศาจ', 'เวท', 'ลูกครึ่ง', 'ทลาย', 'นักแสดง', 'ลูกหนัง', 'ทรหด', 'สี่', 'เดส', 'ฟิลด์', 'นเกอร์', 'ปี', 'กริม', 'พทยา', 'สภา', 'TGI', 'มะเร็ง', 'กีฬา', 'นา', 'ดาล', 'เงื่อนไข', 'ก', 'ถั่ว', 'ฤดูกาล', 'ห่วย', 'เทศกาล', 'สงคราม', 'อินทรี', 'เบิร์ด', 'ลัก', 'นาย', 'อังกฤษ', 'ลงสนาม', 'สุนัข', 'เทพศิรินทร์', 'เจ็บ', 'วันนี้', 'ทองคำ', 'ทฤษฎี', 'เซ', 'ปุยส์', 'ดาร์ง', 'ประตุ', 'กุนชือ', 'บุก', 'ดวง', 'ดี', 'ใกล้', 'เรฟ', 'ชเว', 'ชอย', 'หนู', 'ในประเทศ', 'จิกซอร์'

เกาหลี

'เกาหลีใต้', 'โสม', 'ยอน', 'ศัลยกรรม', 'กง', 'เบลารุส', 'T-50TH', 'ปป้า', 'เรีย', 'ลีซอ', 'กิม', 'เกาหลีเหนือ', 'รู้ตัว', 'แดงโม', 'แย', 'เค็ก', 'โซล', 'ดิช', 'คิม', 'นม', 'เกา', 'จว.', 'Money', 'ปันผล', 'นีก', 'ห', 'เงินเฟ้อ', 'เปรี้ยว', 'เมนู', 'ถ่ายภาพ', 'ควัน', 'แบบ', 'เรื่อย', 'เกิร์ล', 'เบลเยียม', 'เหล่า', 'เลิฟ', 'กำ', 'หลี่', 'T-50', 'มวยสากล', 'ปีท', 'เดนล', 'ลาว', 'ยุ', 'หลังฉาก', 'คัพ', 'BLACKPINK', 'มงคล', 'เนส', 'ความเคลื่อนไหว', 'เขื่อน', 'ทูต', 'อาชกา', 'บัต', 'อัม', 'คิงส์', 'หุ่น', 'หู', 'ตเวอร์', 'เฮือง', 'ใส', 'ขึ้นมื่น', 'บุริรัมย์', 'แท็ก', 'ซอน', 'ตลาด', 'หุ่น', 'ชะตากรรม', 'ลาน', 'ยุง', 'แต่งค์', 'ภา', 'อินทรี', 'ธน', 'เลข', 'เสียดาย', 'ทำทาย', 'ร้ว', 'ไถ', 'ใบ', 'ผูก', 'ฮ', 'the Gods:', 'ญา', 'ไบเตย', 'งา', 'ไขว่คว้า', 'เบิ้ล', 'สงกรานต์', 'ช', 'ขอบคุณ', 'คิว', 'เม', 'อก', 'จุมุก', 'เตะ', 'จริญ', 'แป้ง', 'ขุนพล', 'ซุง', 'แซบ'

ในกรณีนี้ใช้ข้อมูลไม่มาก แต่ได้คำที่บ่งชี้แต่ละประเทศ เช่น

ญี่ปุ่น โตโยต้า, ซากุระ, ซามูไร, ดิบ, AV

จีน หม่า, เชียงไฮ้, มังกร

อเมริกา สหรัฐ, นิวยอร์ก, แชม

เยอรมัน ฮิตเลอร์ MERCEDES ฟิลิป

เกาหลี โสม, ศัลยกรรม, ป้า(<อป้า),

ที่น่าสนใจก็คือ คำเหล่านี้มี metaphor metonymy ด้วย ตัวอย่างหัวข้อที่มีคำว่า ดิบ มังกร แชม ม้าลาย โสม คือ

1215108	5 แข่งตามสมทบ! ช้างศึกยู-16 วางคิวอุ่นโอกินาวา ก่อนลุยแดนปลาดิบ	ญี่ปุ่น
1170393	มังกรสายพันธุ์ใหม่ กวิน กาญจนพาสน์	จีน
1170842	'ช้างศึก ยู-23' กับภารกิจลุยศึกชิงแชมป์เอเชีย ฉบับแดนมังกร	จีน
1197812	เศรษฐกิจดีแต่หุ้นร่วง บทเรียนจากเมืองลุงแซม	อเมริกา
1017750	'ม้าลาย' เชือด 'เปแอสเช' สุดมัน 3-2 คิกไอซีซี	เยอรมัน (ทีมฟุตบอล)
1358598	โสมแดงถล่มซาอุฯ 3-0 ส่งช้างศึกตรอบแรกบอลชาย	เกาหลีเหนือ (ทีมฟุตบอล)

คำเหล่านี้ทั้งหมดแสดงประเทศตัวเอง แต่บางคำยังผิดอยู่ เช่น ผู้ดี = อังกฤษ (ไม่ใช่เยอรมัน) ถ้าให้ฝึกมากกว่านี้ ก็จะสามารถสร้างพจนานุกรม “ศัพท์หนังสือพิมพ์” ได้นอกจากนี้ ข้อมูลที่เก็บไว้ครั้งนี้ ใช้กับอย่างอื่นได้ด้วย ในโครงการนี้ label เป็นประเทศเท่านั้น ถ้าใส่ label ตาม “หมวดหมู่” ของข่าว เช่น การเมือง ต่างประเทศ กีฬา ก็จะได้ news categorizer

5 ปัญหาและอุปสรรค

- ปัญหาใหญ่ที่สุดคือความแม่นยำของ tokenization (tltk)
- ปัญหาใหญ่ที่สองคือเวลา (request.get, tokenize)
- บทความที่ใช้คือระหว่าง 2560-2561 ซึ่งค่อนข้างใหม่และ topic ซ้ำกัน เพราะฉะนั้น ศัพท์ที่ “temporary trend” ก็ออกมา เช่น ‘เซเนกัล’ ในญี่ปุ่นคือเรื่องฟุตบอล
- ถ้าจำนวน label เปลี่ยน ผลลัพธ์ก็จะเปลี่ยน ครั้งนี้มีแค่ 5 ประเทศ แต่ถ้ามีหลายสิบประเทศ น่าจะเดายาก
- ยากที่จะหาบทความที่เกี่ยวข้อง ถ้าใส่ “จีน” ก็จะเก็บบทความเกี่ยวกับ “ปราชญ์บุรี” ด้วย
- นอกจากนั้น บทความบางข้อ label ซ้ำกัน เช่นฟุตบอล ญี่ปุ่น vs จีน