6180140022 โนโซมิ ยามาดะ
NLP Final Project “ศัพท์หนังสือพิมพ์”

1 ภาพรวมและจุดประสงค์

หนังสือพิมพ์ไทยนิยมใช้คำศัพท์ที่ค่อนข้างประหลาดเพื่อทำให้ดึงดูดกว่า เช่น ตอนที่พูดถึงประเทศญี่ปุ่นชอบใช้คำว่า “แดนปลาดิบ” ในโครงการนี้ วิเคราะห์ว่ามีศัพท์บ่งชี้ประเทศอะไรบ้าง โดยใช้ machine learning
กระบวนการสี่ส่วนคือ

1. เก็บข้อมูล (headline, description, article) จาก **ไทยรัฐ** (https://www.thairath.co.th) โดยใช้ beautiful soup และบันทึกไว้เป็น .tsv
2. เอาบทความที่เกี่ยวกับ 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 เพราะฉะนั้นเปลี่ยยเป็นเก็บข้อมูลจากไทยรัฐแทน เก็บข้อมูลมาทั้งหมด มากกว่าแสนบทความ ส่วนใหญ่ไม่เกี่ยวกับประเทศ แต่บันทึกไว้ใน thairath.tsv เพื่อการวิจัยในอนาคต

2 วิธีเก็บข้อมูล

เว็บไซต์ไทยรัฐไม่ให้ API จึงต้องตัดแท็กด้วย beautiful soup
URL ของบทความในเว็บไซต์เป็น 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 ตร.เรียกดูง่าย คนขับเช็กใบสั่งสบาย",
    "alternativeHeadline": "รู้ยัง? เดือนหน้า คนไทยไม่ต้องพกใบขับขี่ ตร.เรียกดูง่าย คนขับเช็กใบสั่งสบาย",
    "image": [
        "https://www.thairath.co.th/media/4DQpjUtzLUwmJZZPFiS41hpWPlrOuaBAvVVKFNK1sav4.jpg"
    ],
    "datePublished": "2018-12-16T19:05:00+07:00",
    "description": "เดือนหน้า คนไทยไม่ต้องพกใบขับขี่ ขนส่งทางบกปิ๊งไอเดียดูดข้อมูลลงแอปฯ ใช้แสดงให้เจ้าหน้าที่ดูเมื่อถูกเรียกตรวจ ไฮเทคมีระบบแจ้งเตือนบัตรหมดอายุ เ
    "articleBody": "เดือนหน้า คนไทยไม่ต้องพกใบขับขี่ ขนส่งทางบกปิ๊งไอเดียดูดข้อมูลลงแอปฯ ใช้แสดงให้เจ้าหน้าที่ดูเมื่อถูกเรียกตรวจ ไฮเทคมีระบบแจ้งเตือนบัตรหมดอายุ 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 มีบทความจำนวนทั้งหมด 105,923 ข้อ (ID: 100000-1370000) แต่บทความที่ใช้จริงไม่มาก ดึงบทความที่มีห้าคำว่า *ญี่ปุ่น อเมริกา ประเทศจีน เยอรมัน เกาหลี* มาบันทึกเป็นไฟล์ใหม่ country.tsv สุดท้ายให้เป็น tokenized text และบันทึกเป็น headline.tsv

thairath.co.th $\xrightarrow{\text{get()}}$ thairath.tsv $\xrightarrow{\text{find\_article()}}$ country.tsv $\xrightarrow{\text{tokenize\_headline()}}$ 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	213	177	176	47	82

Accuracy, precision, recall, F-score เป็นดังนี้

```
Accuracy
0.6067226890756302

Confusion Matrix
[[149   7  37   1  19]
 [ 24  39  10   0   4]
 [ 40   7 108   2  19]
 [ 14   0  11  18   4]
 [ 20   1  14   0  47]]

Report
             precision    recall  f1-score   support

          0       0.60      0.70      0.65       213
          1       0.72      0.51      0.60        77
          2       0.60      0.61      0.61       176
          3       0.86      0.38      0.53        47
          4       0.51      0.57      0.54        82

avg / total       0.62      0.61      0.60       595
```

จำนวนข้อมูลที่ใช้การฝึกไม่ค่อยมากเท่าไร เพราะฉะนั้น ทั้ง Accuracy และ F-score ถึงแค่ 60% นอกจากนั้น

4 ผลการวิจัย

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

ญี่ปุ่น

'ยุ่น', 'ชิ', 'โตเกียว', 'เซเนกัล', 'โตโยต้า', 'ซัป', 'หาด', 'ตะ', 'ทิศทาง', 'ปล้น', 'แปลง', 'พระราม', 'เนียส', 'BNK48', 'ดิบ', 'บัลลังก์', 'ฮ่า', 'ลุม', 'ชั้นนำ', 'ธีร', 'ชนา', 'สื่อสาร', 'เจ', 'ซากุระ', 'โนะ', 'ซามูไร', 'ปินส์', 'ทวาย', 'SUZUKI', 'อลป.', 'SPCG', 'มหัศจรรย์', 'TOYOTA', 'พิรุณ', 'ตูน', 'วะ', 'มาร์ค', 'สร้างสรรค์', 'ศุภ', 'NISSAN', 'หม่อง', 'เมียน', 'โครงสร้าง', 'เทิร์น', 'บิ', 'นุ่น', 'มุ้ย', 'จูเนียร์', 'ผู้เชี่ยวชาญ', 'สัน', 'วาด', 'แม้ว', 'นิส', 'เมย์', 'เสร็จ', 'สกี', 'ศุกร์', 'เรโซ', 'ฟาร์ม', 'วิสาขบูชา', 'สวดมนต์', 'ANA', 'รีบ', 'ซูเปอร์', 'จ้าว', 'กลม', 'เบลเยียม', 'พรีม', 'โรด', 'แนะ', 'ฆู', 'กานต์', 'อ้อน', 'ไปรษณีย์', 'กะ', 'เดวิส', 'ผสม', 'ธัน', 'อีซี', 'ติว', 'ปล้ำ', 'บาย', 'ญี่', 'มิตซูบิชิ', 'แผนการ', 'AV', 'กิ', 'เฟล', 'โนเบล', 'ท้องถิ่น', 'ฮอกไกโด', 'ฟ้าแลบ', 'ศิลป์', 'ผอม', 'โอกินาวา', 'เกียว', 'อิ่ม', 'บัต', 'นครบาล', 'เล่า'

จีน

'ทุเรียน', 'หม่า', 'อู่ฮั่น', 'อาลี', 'ล่ม', 'สมุนไพร', 'ไช', 'กองทัพ', 'กาญจนา', 'รัล', 'น้ำป่า', 'เซี่ยงไฮ้', 'ท่วม', 'ปักกิ่ง', 'ฝน', 'WGP 2017', 'พรรค', 'โฟน', 'ปริมณฑล', 'พายุ', 'เมอร์', 'ไม้', 'สรรเสริญ', 'บัว', 'มังกร', 'อิตาลี', 'หางโจว', 'คอก', 'สะเก็ด', 'ฉ.', 'มาเก๊า', 'ขนุน', 'ทักษิณ', 'มติ', 'แพนด้า', 'ตี๋', 'ย้อนหลัง', 'มิลค์', 'สเวิลด์', 'ตรุษจีน', 'ยู', 'คอย', 'สะเดา', 'ศฤงคาร', 'ยาง', 'MH370', 'ตู้', 'พลิ้ว', 'ดำน้ำ', 'ศก.', 'วัว', 'กระจก', 'โขง', 'สงขลา', 'ดีเปรสชัน', 'ขวัญ', 'ถนน', 'ถึงแก่อนิจกรรม', 'คลื่นลม', 'พิพากษา', 'จม', 'ตั๊ก', 'องศา', 'ชำแหละ', 'แอร์', 'อยุธยา', 'อ่างทอง', 'หนาว', 'ตอน', 'จยย.', 'ประกาย', 'สิทธิชัย', 'อี้', 'ลอต', 'ลำไย', 'ร่าง', 'เอิร์น', 'อาชีพ', 'ขยะ', 'ปัก', 'กลอ', 'แผนงาน', 'อึ๊บ', 'ช่อ', 'ตัดสินใจ', 'แคระ', 'ขบวน', 'รังนก', 'เซอร์เบีย', 'บังคับการ', 'มะม่วง', 'เดียร์', 'อ่าวไทย', 'หนูน้อย', 'ต้า', 'แปรปรวน', 'รับมือ', 'ตลิ่ง', 'บี้', 'องค์กร'

อเมริกา

'อเมริกา', 'สหรัฐ', 'ศรีสะเกษ', 'ฟลอยด์', 'ไอซี', 'นิวยอร์ก', 'สำรวจ', 'พู', 'ดีเอ็นเอ', 'ช.', 'ลาออก', 'มะ', 'เณร', 'เปรู', 'USA', 'ธนบัตร', 'บาร์', 'ไดโนเสาร์', 'แหลม', 'ท้า', 'สหรัฐอเมริกา', 'ดวงตา', 'ซิช', 'ดอน', 'นศ.', 'ชิลี', 'ดวงจันทร์', 'สัญญาณ', 'อเมริกัน', 'ชนิด', 'ข้อสรุป', 'เขต', 'มิก', 'แคลิฟอร์เนีย', 'เสก', 'น้ำผึ้ง', 'อาคาร', 'ปารีส', 'เปลี่ยนแปลง', 'รี', 'ทวีป', 'พืช', 'มหึมา', 'คริส', 'ละเมิด', 'สามารถ', 'แคนาดา', 'ตัวละคร', 'สูญเสีย', 'ภาวะ', 'แซม', 'ไขมัน', 'นีย์', 'แอปเปิล', 'ป้องกัน', 'จ้า', 'แข็งค่า', 'มนุษย์', 'หรูหรา', 'ป้อม', 'ลำไส้', 'โบอิ้ง', 'พม.', 'กราด', 'นาซา', 'สุริยคราส', 'อริ', 'กพท.', 'มหาสมุทร', 'ตอัพ', 'เดือ', 'หลักสูตร', 'ใต้น้ำ', 'ปวดหัว', 'พฤหัสบดี', 'AI', 'กัวเตมาลา', 'ดาวเคราะห์', 'ศูนย์กลาง', 'แค่ไหน', 'อำลา', 'วิทย์', 'ออสการ์', 'เจ๊', 'แชม', 'คอลเลจ', 'รีญ', 'เอาอยู่', 'กติก', 'เด็ต', 'เคน', 'นายพล', 'สินสอด', 'ยาวนาน', 'สารพัน', 'ทศวรรษ', 'หอม', 'ตรวจเลือด', 'การศึกษา', 'เมิน'

เยอรมัน

'หงส์', 'คลอปป์', 'ซิล', 'ฮิตเลอร์', 'MERCEDES', 'ฟิลิป', 'ปืนใหญ่', 'ผู้ดี', 'ราชวงศ์', 'ฝรั่ง', 'เบียร์', 'เยอรมนี', 'พรหม', 'เอล', 'ไอน์สไตน์', 'ยุโรป', 'ม้าลาย', 'บุน', 'ฟินแลนด์', 'เมธี', 'ภูเก็ต', 'เก', 'หูหนวก', 'จิ้งจอก', 'ปืน', 'เสือ', 'PORSCHE', 'นู', 'อุส', 'เป๊ป', 'รัชสมัย', 'ชาน', 'ลงตัว', 'เทล', 'โรล', 'เยิร์น', 'อุตสาหกรรม', 'BMW', 'ฮูธ', 'เลอร์', 'อลิส', 'แลก', 'คาร์ล', 'ซัป', 'ฝึกงาน', 'เล', 'ปิศาจ', 'เวท', 'ลูกครึ่ง', 'ทลาย', 'นักแสดง', 'ลูกหนัง', 'ทรหด', 'สี่', 'เดส', 'ฟิลด์', 'นเกอร์', 'บี.', 'กริม', 'พัทยา', 'สภา', 'TGI', 'มะเร็ง', 'กีฬา', 'นา', 'ดาล', 'เงื่อนไข', 'n', 'ถ้วย', 'ฤดูกาล', 'ห่วย', 'เทศกาล', 'สงคราม', 'อินทรี', 'เบิร์ด', 'ลัก', 'นาย', 'อังกฤษ', 'ลงสนาม', 'สุนัข', 'เทพศิรินทร์', 'เจ็บ', 'วันนี้', 'ทองคำ', 'ทฤษฎี', 'เช่', 'ปุยส์', 'ดาวรุ่ง', 'ประตู', 'กุนซือ', 'บุก', 'ดวง', 'ดี้', 'ใกล้', 'เรฟ', 'ซเว', 'ซอย', 'หนู', 'ในประเทศ', 'จิ๊กซอว์'

เกาหลี

'เกาหลีใต้', 'โสม', 'ยอน', 'ศัลยกรรม', 'กง', 'เบลารุส', 'T-50TH', 'ปป้า', 'เรีย', 'ลีซอ', 'กิม', 'เกาหลีเหนือ', 'รู้ตัว', 'แตงโม', 'แย้', 'เค้ก', 'โซล', 'ติช', 'คิม', 'นัม', 'เกา', 'จว.', 'Money', 'ปันผล', 'นึก', 'ห', 'เงินเฟ้อ', 'เปรี้ยง', 'เมนู', 'ถ่ายภาพ', 'ควัน', 'แบม', 'เรื่อย', 'เกิร์ล', 'เบลเยี่ยม', 'เหล้า', 'เลิฟ', 'กำ', 'หลี', 'T-50', 'มวยสากล', 'ปัท', 'เตนล์', 'ลาว', 'ยุ', 'หลังฉาก', 'คัพ', 'BLACKPINK', 'มงคล', 'เนส', 'ความเคลื่อนไหว', 'เขื่อน', 'ทูต', 'อาชกาบัต', 'อั้ม', 'คิงส์', 'หุ่น', 'หู', 'ตเว่อร์', 'เฮือง', 'ใส', 'ชื่นมื่น', 'บุรีรัมย์', 'แท็ก', 'ชอน', 'ตลาดหุ้น', 'ชะตากรรม', 'ลาน', 'ยุ่ง', 'แต๊งค์', 'ภา', 'อินทรี', 'ธน', 'เลข', 'เสียดาย', 'ท้าทาย', 'รัว', 'ไถ่', 'ใบ้', 'ผูก', 'ฮ', 'the Gods:', 'ญ่า', 'ใบเตย', 'งา', 'ไขว่คว้า', 'เบิ้ล', 'สงกรานต์', 'ช', 'ขอบคุณ', 'คิ้ว', 'เม', 'อก', 'จมูก', 'เตะ', 'จริญ', 'แป้ง', 'ขุนพล', 'ซุง', 'แซบ'

ในกรณีนี้ใช้ข้อมูลไม่มาก แต่ได้คำที่บ่งชี้แต่ละประเทศ เช่น

ญี่ปุ่น โตโยต้า, ซากุระ, ซามูไร, ดิบ, AV

จีน หม่า, เซี่ยงไฮ้, มังกร

อเมริกา สหรัฐ, นิวยอร์ก, แซม

เยอรมัน ฮิตเลอร์ MERCEDES ฟิลิป

เกาหลี โสม, ศัลยกรรม, ปป้า(<อปป้า?),

ที่น่าสนใจก็คือ คำเหล่านี้มี metaphor และ metonymy ด้วย ตัวอย่างหัวข้อที่มีคำว่า *ดิบ มังกร แซม ม้าลาย โสม* คือ

1215108	5 แข้งตามสมทบ! ช้างศึกยู-16 วางคิวอุ่นโอกินาวา ก่อนลุยแดนปลา**ดิบ**	ญี่ปุ่น
1170393	**มังกร**สายพันธุ์ ใหม่ กวิน กาญจนพาสน์	จีน
1170842	’ช้างศึก ยู-23’ กับภารกิจลุยศึกชิงแชมป์เอเชีย ฉบับแดน**มังกร**	จีน
1197812	เศรษฐกิจดีแต่หุ้นร่วง บทเรียนจากเมืองลุง**แซม**	อเมริกา
1017750	’**ม้าลาย**’ เชือด ’เปแอสเช’ สุดมัน 3-2 ศึกไอซีซี	เยอรมัน (ทีมฟุตบอล)
1358598	**โสม**แดงถล่มซาอุฯ 3-0 ส่งช้างศึกตกรอบแรกบอลชาย	เกาหลีเหนือ (ทีมฟุตบอล)

คำเหล่านี้ทั้งหมดแสดงประเทศนั่นเอง แต่บางคำยังผิดอยู่ เช่น ผู้ดี = อังกฤษ (ไม่ใช่เยอรมัน) ถ้าให้ฝึกมากกว่านี้ ก็จะสามารถสร้างพจนานุกรม “ศัพท์หนังสือพิมพ์” ได้
นอกจากนี้ 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 “ปราจีนบุรี” ด้วย
- นอกจากนั้น บทความบางข้อ label ซ้ำกัน เช่นฟุตบอล ญี่ปุ่น vs จีน ตอนแรกไม่ได้สังเกต ก็เลย accuracy ต่ำมาก