

DS4Biz - Assignment 2 (25 Marks)

Text Scraping & Classification

Deadline: วันศุกร์ที่ 27 พฤศจิกายน 2563 เวลา 24.00 น.

Submission: *Two*-People Group github repository – stage, commit with message, and push

Overview

จุดประสงค์ของ Assignment นี้ คือ ให้นัก. ทำการ scrape (รวบรวม) ข้อมูลจากหน้าเว็บไซต์ จำนวน 2 เว็บ ได้แก่

1. เว็บจำลองเก็บรวบรวมข่าวเก่า

http://www.it.kmitl.ac.th/~teerapong/news_archive/index.html

ให้นัก. ทำการ scrape บทความข่าวจากหน้าเว็บไซต์ (มีจำนวนมากกว่า 1 หน้าเว็บ) **ทั้งหมดทุกบทความ** โดยมีเป้าหมายเพื่อพัฒนาระบบที่มี Machine Learning/Data Science ในการช่วยบรรณาธิการข่าว ในการจัดหมวดหมู่ข่าวอัตโนมัติ เช่น Technology, Business และ Sport โดยปัญหานี้ให้นัก.มองเป็น **Multi-class classification problem** โดยข่าว 1 บทความ สามารถอยู่ได้ในหมวดหมู่เดียวจาก 3 classes โดยขณะนั้น. ทำการ scrape ข้อมูล ให้ทำการเก็บหมวดหมู่ข่าวมาเป็นตัวอย่างในการ train/fit model ของนศ. โดยนศ. สามารถทดลองใช้เฉพาะเนื้อหาข่าว (body/content) หรือหัวข้อข่าว (heading/title) หรือรวมกันก็ได้ เพื่อให้ได้ประสิทธิภาพสูงสุด (ความแม่นยำ (accuracy) ปัจจุบันอยู่ที่ ~0.975) รวมถึงขั้นตอนการทำ Text Processing อยู่นัก.จะเลือกทดลองใช้ และทำการตั้งค่าการทดลอง (Experimental Settings) อย่างถูกต้องตามหลักวิชาการ (เช่น การแบ่ง train/test dataset การใช้ hold-out sampling หรือ cross validation ตามที่นศ. เห็นเหมาะสม และต้องให้เหตุผลการเลือกใช้) ให้ได้ชนิดของโมเดลและพารามิเตอร์ที่ทำให้โมเดลมีความแม่นยำที่สุด



News Article Archive

Archive of all news headlines and stories, organised per month.

Articles — [January](#) [118]

Articles — [February](#) [124]

Articles — [March](#) [116]

Articles — [April](#) [118]

Articles — [May](#) [115]

Articles — [June](#) [115]

Articles — [July](#) [122]

Articles — [August](#) [116]

Articles — [September](#) [113]

Articles — [October](#) [124]

Articles — [November](#) [122]

Articles — [December](#) [115]

2. เว็บไซต์การ scrape ข้อมูล ซึ่งแสดงคำกล่าว (Quote) ของผู้มีชื่อเสียง

<https://quotes.toscrape.com/>

ให้นัก. ทำการ scrape “quote” จากหน้าเว็บไซต์ (มีจำนวนมากกว่า 1 หน้าเว็บ) **ทั้งหมดทุก ๆ quote** โดยมีเป้าหมายเพื่อพัฒนาระบบที่มี Machine Learning/Data Science ในการช่วยติด Tag แก่ quote อย่างอัตโนมัติ เช่น Love, Inspirational, Abilities, Deep-thoughts เป็นต้น โดยปัญหานี้ให้นัก.มองเป็น **Binary classification problem** ว่า quote นี้ควรทำการติด tag เช่น “humor” หรือไม่ (Yes/No)

โดย 1 quote สามารถติด tag ได้มากกว่าหรือเท่ากับ 1 tag โดยขณะนั้น.ทำการ scrape ข้อมูลให้ทำการเก็บ tags มาเป็นตัวอย่างในการ train/fit model ของนศ. โดยนศ.สามารถทดลองใช้เฉพาะเนื้อคำของ quote (body/content) หรือชื่อผู้แต่ง (author) หรือรายละเอียดของผู้แต่ง (about author/author detail) หรือนำมารวมกันก็ได้ เพื่อให้ได้ประสิทธิภาพสูงสุด รวมถึงขั้นตอนการทำ Text Processing อยู่ทีนศ. จะเลือกทดลองใช้และทำการตั้งค่าการทดลอง (Experimental Settings) อย่างถูกต้องตามหลักวิชาการ (เช่น การแบ่ง train/test dataset การใช้ hold-out sampling หรือ cross validation ตามทีนศ.เห็นเหมาะสม และต้องให้เหตุผลการเลือกใช้) ให้ได้ชนิดของโมเดลและพารามิเตอร์ที่ทำให้โมเดลมีความแม่นยำที่สุด

Quotes to Scrape

Login

“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”

by Albert Einstein (about)

Tags: change deep-thoughts thinking world

“It is our choices, Harry, that show what we truly are, far more than our abilities.”

by J.K. Rowling (about)

Tags: abilities choices

“There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”

by Albert Einstein (about)

Tags: inspirational life live miracle miracles

“The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.”

by Jane Austen (about)

Tags: aliteracy books classic humor

Top Ten tags

love
inspirational
life
humor
books
reading
friendship
science
truth
death

สำหรับ Assignment นี้ นศ.จะต้องถูกทำลงใน Jupyter Notebook โดยแยกการทำการทดลองของนศ. เป็น 2 Notebook โดยแต่ละ Notebook ใช้กับคนละเว็บกัน

โดยเว็บข่าว ให้นัก.ตั้งชื่อ Notebook ว่า

1. news.ipynb
2. quote.ipynb

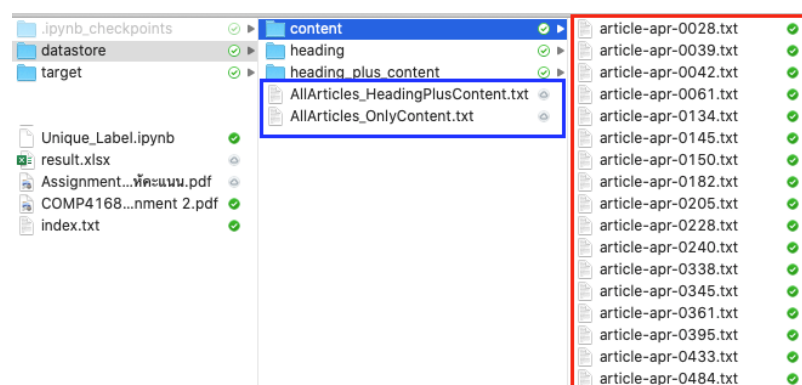
โค้ดที่นศ. เขียนจะ**ต้อง**มีการเขียนอธิบายที่ชัดเจนโดยใช้ Markdown cells เพื่ออธิบายแต่ละ Code cells ที่อยู่ด้านล่าง (ลำดับถัดไป) ของแต่ละ Markdown cell รวมทั้งใช้ inline (#) or block comments (""" ... """) ประกอบการอธิบายโค้ดและผลลัพธ์ของการวิเคราะห์ข้อมูลของนศ.

โดยทุก ๆ Code cells จะต้องมีการเขียนอธิบายโดย Markdown cells ว่า แต่ละ Code cells นั้นทำอะไรและแปลความหมายผลลัพธ์ว่าอะไร

Part 1: Data Collection (30%)

เป้าหมายของส่วนที่ 1 นี้ คือ ให้นักศึกษาฝึกการ scrape ข้อมูลจากหน้าเว็บไซต์ โดยทำการรวบรวมข้อมูล text ที่ถูก labelled (news category or quote tags) แล้ว**ทั้งหมด** ที่ถูกแสดงตามหมวดหมู่ของข่าวหรือการติด tags โดยในส่วนนั้นศ.จะต้องทำรายการงาน ทั้งหมดดังต่อไปนี้

1. ระบุหา URLs และ news category and quote tag labels ทั้งหมด สำหรับข้อมูล text ทั้งหมดที่มีอยู่ในเว็บไซต์ข้างล่างนี้ โดยโค้ดของนศ.จะต้องใช้ root URL เริ่มต้นข้างล่างนี้ในการหาข้อมูล text ที่เหลือทั้งหมด โดยเริ่มจาก URL ตั้งต้นที่ให้
http://www.it.kmitl.ac.th/~teerapong/news_archive/index.html
<https://quotes.toscrape.com/>
2. ให้นศ.ค้นคืนหาข้อมูล text ทั้งหมดตาม URLs ที่ได้จากการสกัดจากหน้าเว็บไซต์ที่ได้จากตั้งแต่ root URL ที่ได้ข้างต้นและ crawl ตาม anchor links ที่ปรากฏในหน้าโดย scrape เฉพาะเนื้อหาข้อมูล text ที่จะนศ.จะใช้ในการทดลองเท่านั้น และทำการสกัดส่วนที่เป็น body text ซึ่งมีเนื้อหาของข้อมูล text และทำการ save ส่วนของเนื้อหา text เหล่านั้นเป็น plain text แยกกันของแต่ละเว็บไซต์ โดยแต่ละเว็บไซต์ใช้เพียง 1 ไฟล์ ซึ่งเนื้อหาของแต่ละบทความจะถูกบันทึกไว้ในแต่ละบรรทัด (1 บรรทัด : 1 บทความ/quote) โดยให้ทำการ save เนื้อหาข้อมูลไว้ที่ folder ชื่อ ../datastore โดยตั้งชื่อไฟล์ให้สื่อความหมาย และใช้นามสกุล *****.txt เช่น จากรูปตัวอย่างข้างล่าง ผมทำการ crawl ข้อมูล แล้วเก็บไว้ทั้ง 2 แบบ แบบที่เก็บ body text + title และ เฉพาะ body text อย่างเดียว โดยตั้งชื่อไฟล์ว่า “AllArticles_HeadingPlusContent.txt” และ “AllArticles_OnlyContent.txt” และ ตามลำดับ (ในกรอบสีน้ำเงิน)



โดยตัวอย่างไฟล์ที่ save เป็น plain text นั้น อาจใช้ตามตัวอย่างที่ให้ข้างล่างนี้ก็ได้ ซึ่ง 1 บรรทัด คือ text ของ 1 บทความ หรือจะใช้ในรูปแบบ csv

1	21st-Century Sports: How Digital Technology Is Changing the Face Of The Sporting Industry The sporting industry has come a long way since the '60s. It has carved out for itself a niche with its roots so deep that I cannot fathom the sports industry showing any sign of decline any time soon - or later. The reason can be found in this seemingly subtle difference - other industries have customers; the sporting industry has fans. Vivek Ranadivé, leader of the ownership group of the NBA's Sacramento Kings, explained it beautifully, "Fans will paint their face purple, fans will evangelize. ... Every other CEO in every business is dying to be in our position - they're dying to have fans." While fan passion alone could almost certainly keep the industry going, leagues and sporting franchises have decided not to rest on their laurels. The last few years have seen the steady introduction of technology into the world of sports - amplifying fans' appreciation of games, enhancing athletes' public profiles and informing their training methods, even influencing how contests are waged. Also, digital technology in particular has helped to create an alternative source of revenue, besides the games themselves - corporate sponsorship. They achieved this by capitalizing on the ardor of their customer base - sorry, fan base. Return to article search results
2	Asian quake hits European shares Shares in Europe's leading reinsurers and travel firms have fallen as the scale of the damage wrought by tsunamis across south Asia has become apparent. More than 23,000 people have been killed following a massive underwater earthquake and many of the worst hit areas are popular tourist destinations. Reinsurance firms such as Swiss Re and Munich Re lost value as investors worried about rebuilding costs. But the disaster has little impact on stock markets in the US and Asia. Currencies including the Thai baht and Indonesian rupiah weakened as analysts warned that economic growth may slow. "It came at the worst possible time," said Hans Goetti, a Singapore-based fund manager. "The impact on the tourist industry is pretty devastating, especially in Thailand." Travel-related shares dropped in Europe, with companies such as Germany's TUI and Lufthansa and France's Club Mediterranée sliding. Insurers and reinsurance firms were also under pressure in Europe. Shares in Munich Re and Swiss Re - the world's two biggest reinsurers - both fell 1.7% as the market speculated about the cost of rebuilding in Asia. Zurich Financial, Allianz and Axa also suffered a decline in value. However, their losses were much smaller, reflecting the market's view that reinsurers were likely to pick up the bulk of the costs. Worries about the size of insurance liabilities dragged European shares down, although the impact was exacerbated by light post-Christmas trading. Germany's benchmark Dax index closed the day 16.29 points lower at 3,817.69 while France's Cac index of leading shares fell 5.07 points to 3,817.69. Investors pointed out, however, that declines probably would be industry specific, with the travel and insurance firms hit hardest. "It's still too early for concrete damage figures," Swiss Re's spokesman Floiran Woest told Associated Press. "That also has to do with the fact that the damage is very widely spread geographically." The unfolding scale of the disaster in south Asia had little immediate impact on US shares, however. The Dow Jones index had risen 20.54 points, or 0.2%, to 10,847.66 by late morning as analysts were cheered by more encouraging reports from retailers about post-Christmas sales. In Asian markets, adjustments were made quickly to account for lower earnings and the cost of repairs. Thai Airways shed almost 4%. The country relies on tourism for about 6% of its total economy. Singapore Airlines dropped 2.6%. About 5% of Singapore's annual gross domestic product (GDP) comes from tourism. Malaysia's budget airline, AirAsia fell 2.9%. Resort operator Tanco Holdings slumped 5%. Travel companies also took a hit, with Japan's Kinki Nippon sliding 1.5% and HIS dropping 3.3%. However, the overall impact on Asia's largest stock market, Japan's Nikkei, was slight. Shares fell just 0.03%. Concerns about the strength of economic growth going forward weighed on the currency markets. The Indonesian rupiah lost as much as 0.6% against the US dollar, before bouncing back slightly to trade at 9,300. The Thai baht lost 0.3% against the US currency, trading at 39.10. In India, where more than 2,000 people are thought to have died, the rupee shed 0.1% against the dollar. Analysts said that it was difficult to predict the total cost of the disaster and warned that share prices and currencies would come under increasing pressure as the bills mounted. Comments are closed for this article.
3	BT offers free net phone calls BT is offering customers free internet telephone calls if they sign up to broadband in December. The Christmas give-away entitles customers to free telephone calls anywhere in the UK via the internet. Users will need to use BT's internet telephony software, known as BT Communicator, and have a microphone and speakers or headset on their PC. BT has launched the promotion to show off the potential of a broadband connection to customers. People wanting to take advantage of the offer will need to be a BT Together fixed-line customer and will have to sign up to broadband

- (optional) heading/title พาดหัวข่าว ของแต่ละบทความ หรือ about author/author detail รายละเอียดของผู้แต่ง สามารถนำมารวมกับส่วนของ body ได้ แต่ถ้านศ.เลือกวิธีนี้ร่วมด้วย ต้องมีกระบวนการในการใช้งานคำที่ปรากฏใน title แตกต่างจาก คำที่ปรากฏอยู่ใน body และมีการทำการทดลองเปรียบเทียบว่า การใช้ title ประกอบในการทำ Text Classification นั้น ช่วยในเพิ่มประสิทธิภาพของการทำ Text Classification โดยใช้คำจาก body เพียงอย่างเดียวหรือไม่ และเพิ่มหรือลดประสิทธิภาพอย่างไร
- ให้ศ.ทำการบันทึก category/tag labels ของข้อมูล text ทั้งหมด ในไฟล์ที่แยกจาก body เพื่อใช้ในการทำ target variable โดยให้ทำการ save เนื้อหาข้อมูลไว้ที่ folder ชื่อ ../target โดยตั้งชื่อไฟล์ให้สื่อความหมาย และใช้นามสกุล category.txt (1 column) หรือ tags.txt (n columns according to the number of total tags)

Part 2: Text Classification (50%)

เป้าหมายของส่วนที่ 2 นี้ คือ ให้นักศึกษาทำการวิเคราะห์จำแนกหมวดหมู่ของข้อความ (text classification) จากข้อมูลจากที่ได้มาใน part 1 โดยในส่วนนั้นศ.จะต้องทำรายการงาน ทั้งหมดดังต่อไปนี้

- โหลดชุดของไฟล์ที่เราสร้างขึ้นใน part 1 ลงใน Jupyter Notebook ของตัวเอง (ในแต่ละไฟล์ต้องมี class label โดยยึดตาม category/tag label ที่นศ. ได้การ save ไฟล์แยกเอาไว้จากบทความข่าว โดย class label นั้นจะต้องสัมพันธ์กับข่าวหรือ quote ที่ระบุเอาไว้)
- จากข้อมูล text (raw documents) ที่โหลดมาข้างต้น ให้ศ.สร้าง document-term matrix โดยใช้วิธีที่เหมาะสมในแต่ละขั้นตอน ในการประมวลข้อความเบื้องต้น (text pre-processing) และการถ่วงน้ำหนักของคำ (term weighting) ซึ่งนำไปสู่ประสิทธิภาพของการทำ classification ในลำดับถัดไป พร้อมอธิบายเทคนิคที่นศ.เลือกใช้ในการประมวลผลข้อความเบื้องต้นและการถ่วงน้ำหนักคำ
- ให้ศ. สร้าง multi-class classification models (ข่าว) และ binary classification problem (quote) อย่างน้อย 2 โมเดล หรือมากกว่า โดยใช้ classifiers ต่างประเภทกัน อย่างน้อย 3 ประเภท หรือมากกว่า (ยังทำการทดลองครบสมบูรณ์ หลายโมเดล ยิ่งได้คะแนนมาก) และทำการ Tune โมเดลให้ได้ประสิทธิภาพสูงที่สุด

(จะดูขั้นตอนการ Tune ว่าทำถูกต้องตามหลักวิชาการหรือไม่ หากถูกต้อง และสมบูรณ์ มีวิธีการ Tune ที่ทันสมัย ยิ่งได้คะแนนมาก) และอธิบายเหตุผลของการเลือกประเภทโมเดลมาในการทดลอง พร้อมวิธีการ Tune โมเดลที่คิดว่าจะทำให้ได้ประสิทธิภาพสูงที่สุด

4. ให้นัก. ทำการเปรียบเทียบผลลัพธ์ที่ได้จากแต่ละ models ที่นัก. ได้เลือกไว้ในข้อ 3 โดยนัก. จะต้องเลือกวิธีการหรือการนำเสนอที่เหมาะสมในการประเมินประสิทธิภาพและเปรียบเทียบระหว่างโมเดล นัก. จะต้องรายงานและอภิปรายผลลัพธ์การประเมินที่ได้ลงใน Markdown cells ของ Jupyter Notebook ของตนเอง (นำผลการทดลองของแต่ละโมเดลที่ทำการ Tune แล้ว มาเปรียบกันระหว่างโมเดล เช่นรูปแบบตารางหรือ graph ต่าง ๆ ยิ่งสมบูรณ์ ดูทำความเข้าใจง่าย ยิ่งได้คะแนนมาก)

Code quality and explanation text (20%)

- นัก. ต้องใช้ Markdown cells ในการอธิบายแต่ละขั้นตอนของกระบวนการ โดยนัก. ควรแยกส่วน ของ Part 1 – data collection และ Part 2 – classifier evaluation ให้เด่นชัด แต่ยังคงอยู่ใน Notebook เดียวกันของแต่ละเว็บ
- โค้ดที่นัก. เขียนจะต้องอ่านและเข้าใจได้โดยง่าย ชัดเจน และไม่คลุมเครือ โดยควรมี comment อธิบายที่เพียงพอในการทำให้เข้าใจโค้ดได้โดยง่าย แต่ไม่มากเกินไป
- ความซับซ้อนของโค้ดอยู่ในระดับเท่าที่จำเป็น และมีการใช้ Package ต่าง ๆ ที่เหมาะสม โดยเกณฑ์หลัก ๆ จะดูจากวิธีการเขียนการทดลองเชิงเปรียบเทียบว่าสมเหตุสมผลหรือไม่ สอดคล้องตามเป้าหมายตามที่นัก. มีเจตนาธรรม ในการทำหรือไม่ ซึ่งการให้คะแนนจะพิจารณาประสิทธิภาพของ classifier เป็นเรื่องรอง และความเร็วในการประมวลผลเป็นเรื่องรอง (หากการทดลองไม่เว้นเว้อจนเกินจุดประสงค์)

Guidelines:

- สำหรับ assignment นี้ อนุญาตให้นัก. ใช้เฉพาะ third-party packages เหล่านี้เท่านั้นในการทำ assignment ได้แก่: NumPy, Pandas, Scikit-learn, NLTK, SciPy, Requests, BeautifulSoup, Scrapy, Matplotlib, Seaborn หากใครใช้มากกว่านี้ ต้องขออนุญาตก่อน มิฉะนั้นจะหักคะแนน package ละ 10% และหากมีการอนุญาต ก็จะประชาสัมพันธ์ให้นัก. คนอื่นใช้ได้ด้วย เช่นเดียวกัน
- ให้นัก. ทำการส่ง Assignment ซึ่งคือ
 1. ไฟล์ .ipynb ของ Jupyter Notebook ของนัก. พร้อมข้อมูลที่รวบรวมมา ใน Github repository ของนัก. แต่ละคน โดยในแต่ละ Jupyter Notebook ของนัก. นัก.จะต้องเขียน ชื่อ นามสกุล รหัสนัก. ลงใน Markdown cell แรกของ Notebook
 2. Snapshots ของไฟล์ .ipynb ที่มี output ของทุก cells เป็น
 - ไฟล์ .html โดยการ download as ไฟล์ .ipynb ของนัก. เป็น html
 - ไฟล์ .pdf โดยการ download as ไฟล์ .ipynb ของนัก. เป็น pdf

ทั้งนี้เพื่อป้องกันการนำไฟล์มารันใหม่แล้วได้ผลลัพธ์ไม่เหมือนเดิม จะได้ตรวจจากผลลัพธ์ที่นัก.ทำได้
- Assignment นี้เป็นงานกลุ่ม โดยแต่ละกลุ่มมีนัก. 1-2 คน หากมีการตรวจสอบพบการคัดลอก (Plagiarism) จะได้ 0 คะแนนในส่วนของ Assignment นี้ หากมีข้อสงสัย และหากมีหลักฐานชัดเจนว่ามีการคัดลอกงานจากแหล่งใด ๆ ก็ตาม นัก. จะได้เกรด F ในวิชานี้ และส่งเรื่องต่อให้กับทางคณะฯ และสถาบันฯ ต่อไป

- Hard deadline: วันศุกร์ที่ 27 พฤศจิกายน 2563 เวลา 24.00 น.
 - ส่งช้า 1-5 วัน: ลด 20% จากคะแนนตรวจที่ได้ (ขอเปิด Github ให้ส่งช้า)
 - ส่งช้า 6-10 วัน: ลด 40% จากคะแนนตรวจที่ได้ (ขอเปิด Github ให้ส่งช้า)
 - จะไม่มีการรับตรวจ Assignment หากส่งช้าเกิน 10 วัน โดยปราศจากหลักฐานชี้แจงเหตุผลในการส่งงานช้า ได้แก่ หลักฐานด้านการแพทย์ว่าเข้านอนโรงพยาบาลเพื่อรับการรักษา