# โปรแกรมและเครื่องมือเพื่องานวิจัยภาษาศาสตร์

วิโรจน์ อรุณมานะกุล

draft : ๒๕ ม.ค. ๒๕๖๓

ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

# บทนำ

เอกสารนี้ใช้ประกอบการสอนวิชาเทคนิคการวิจัยทางภาษาศาสตร์ ของภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย โดยเน้นที่เรื่องการจัดเก็บข้อมูล และการใช้เครื่องมือหรือโปรแกรมต่างๆ เพื่อประโยชน์ในการวิจัย นอกจากการใช้ประกอบวิชาการวิจัย ในรายวิชาอื่น ๆ ที่เกี่ยวข้องกับการใช้คลังข้อมูลภาษาและการวิเคราะห์ ก็สามารถใช้ประโยชน์จากบทที่เกี่ยวข้องได้

เมื่อกล่าวถึงข้อมูลสำหรับงานวิจัย เราสามารถแยกเป็น ข้อมูลปฐมภูมิ (Primary sources) กับข้อมูลทุติยภูมิ (Secondary sources) ข้อมูลปฐมภูมิ คือข้อมูลดิบที่ นักวิจัยรวบรวมมาเพื่อใช้ในงานวิจัย ส่วนข้อมูลทุติยภูมิ คือข้อมูลที่มีผู้เก็บรวบรวมหรือวิเคราะห์มาแล้ว ถ้าพูดถึงเอกสาร (document) ก็จะเป็น primary sources เป็นเอกสารที่ให้ข้อมูลดิบเพื่อการวิจัย ส่วน secondary sources เป็นเอกสารที่ให้ข้อมูลเกี่ยวกับ primary sources อีกทีหนึ่ง

สำหรับข้อมูลในงานวิจัยทางด้านภาษาศาสตร์อาจจะแยกข้อมูลออกเป็นข้อมูลภาษา กับ ข้อมูลงานวิจัย ข้อมูลภาษาอาจแยกเป็นข้อมูลประเภทต่างๆ เช่น ข้อมูลภาษาเขียน ภาษาพูด, ภาษาปัจจุบัน ภาษาเก่า, ภาษามาตรฐาน ภาษาถิ่น, เป็นต้น ข้อมูลภาษามีความสำคัญมากเพราะเป็นตัวแทนของปรากฏการณ์ทางภาษาที่เราต้องการศึกษา จึงต้องมีปริมาณที่มากพอจะศึกษาได้และเป็นตัวแทนของภาษาที่ต้องการศึกษาอย่างแท้จริง งานวิจัยจำนวนมากจึงอาศัยการวิเคราะห์ข้อมูลจากคลังข้อมูลภาษาที่เป็นตัวแทนของภาษาที่ต้องการศึกษา

ข้อมูลอาจเก็บจากเอกสารต่างๆ หรือในบางกรณีก็เป็นการเก็บข้อมูลภาคสนาม สิ่งที่ต้องคำนึงคือทำอย่างไรให้เก็บข้อมูลที่ต้องการให้ได้ตามจำนวนที่เพียงพอ ไม่มีผลกระทบจากปัจจัยเกี่ยวกับผู้

เก็บข้อมูลและวิธีการเก็บข้อมูล  การควบคุมตัวแปรต่างๆ  หากมีข้อมูลที่คนอื่นเก็บมาหรือสร้างไว้แล้ว เราต้องพิจารณาว่าข้อมูลนั้นตรงกับที่เราต้อการหรือไม่  มีรายละเอียดที่มาครบถ้วน  และสามารถนำมาใช้ได้โดยตรงหรือไม่  การทำแบบสอบถามผ่านทางอินเทอร์เน็ตก็เป็นวิธีการหนึ่งที่เราสามารถใช้ได้หากต้องการสำรวจข้อมูลต่างๆ เพื่อใช้ในงานวิจัย

นอกจากข้อมูลภาษา ข้อมูลงานวิจัยก็เป็นสิ่งสำคัญ เป็นส่วนของการทบทวนวรรณกรรม ในระหว่างการสำรวจงานต่างๆ ผู้วิจัยจึงต้องเรียนรู้ที่จะเก็บข้อมูลงานวิจัยต่างๆ รู้จักการจัดการข้อมูลบรรณานุกรม การใช้โปรแกรม Endnote เพื่อทำบรรณานุกรม หากไม่มีโปรแกรม Endnote การใช้โปรแกรม Zotero ร่วมกับ web browser ต่างๆ ก็เป็นเครื่องมือหนึ่งที่ช่วยในการเก็บระเบียนข้อมูลระหว่างการค้นคว้าได้ดี ในเอกสารนี้จะกล่าวถึงการสร้างรายการบรรณานุกรมใหม่, การโหลดข้อมูลจากห้องสมุด, การเลือกรูปแบบ, การคัดลอกบรรณานุกรมใส่รายงาน โดยใช้โปรแกรม Endnote

การค้นข้อมูลงานวิจัย ปัจจุบันมีฐานข้อมูลจำนวนมากที่มหาวิทยาลัยรับเป็นสมาชิก ในจุฬาฯ หากเข้าไปที่เว็บห้องสมุดจะเห็น CU-reference database ที่เชื่อมไปยังฐานข้อมูลต่างๆ ทำให้สามารถค้น บทคัดย่อ บทความวิจัย วารสาร หนังสือออนไลน์ นอกจากนี้ ยังมีแหล่งข้อมูลบนอินเทอร์เน็ตอีกเป็นจำนวนมากที่สามารถใช้ประโยชน์ค้นข้อมูลวิจัย เช่น www.scribd.com, google search, scholar.google.co.th,

นอกจากโปรแกรมที่ใช้ในการจัดการข้อมูลต่างๆ แล้ว  โปรแกรมสำหรับใช้ในงานทางภาษาศาสตร์นั้นมีมากมาย ได้แก่ โปรแกรมวาดต้นไม้ โปรแกรมวิเคราะห์เสียง โปรแกรมสำหรับทำการทดลองทางจิตวิทยา  โปรแกรมกำกับข้อมูลภาษา เป็นต้น การทบทวนวรรณกรรมจะช่วยให้เราเห็นว่ามีเครื่องมืออะไรบ้างที่สามารถใช้ได้ และเครื่องมือไหนเป็นที่ยอมรับใช้กันแพร่หลายในสาขาที่เราสนใจ  โปรแกรมที่ไม่ได้ครอบคลุมในหนังสือเล่มนี้ และมีการใช้งานแพร่หลาย คือโปรแกรม Concordance เช่น AntConc  ซึ่งมีกล่าวถึงไว้อยู่ในหนังสือ ภาษาศาสตร์คลังข้อมูล

นอกเหนือจากโปรแกรมสำเร็จรูปที่มีคนทำให้เราได้ใช้ในเรื่องต่าง ๆ บางครั้ง เราก็จำเป็นที่จะต้องเขียนโปรแกรมขึ้นมาเพื่อประมวลผลภาษาเอง  ในหนังสือแยกต่างหากอีกเล่ม คือ R ในงานสถิติ

และการทำเหมืองข้อความ ก็เป็นอีกคู่มือหนึ่งสำหรับผู้สนใจที่ต้องการเขียน code หรือใช้ code เพื่องานทางวิจัยได้ โดยเฉพาะในการคำนวณสถิติ  การสร้างกราฟแสดงผล และการทำเหมืองข้อความ

นอกจากนี้ สิ่งสำคัญในการวิจัยคือการเก็บรักษาข้อมูลให้ปลอดภัย การสำรองข้อมูล การรักษาความปลอดภัยข้อมูลจึงเป็นสิ่งสำคัญและจำเป็น  เพราะเราไม่รู้ว่าพรุ่งนี้จะเปิดคอมพิวเตอร์มาแล้วใช้ได้ไหม จึงต้องมีการสำรองข้อมูลที่ดี หากข้อมูลสูญหายไม่ว่าด้วยเหตุใด งานวิจัยนั้นก็ล้มเหลวก่อนจะถึงปลายทาง

CHAPTER 1

# โปรแกรม Endnote

โปรแกรม Endnote เป็นโปรแกรมสำหรับช่วยทำบรรณานุกรม ข้อดีของโปรแกรมนี้คือสามารถจัดรูปแบบบรรณานุกรมและการอ้างอิงตามที่แต่ละวารสาร หรือแต่ละสมาคมต้องการได้ โดยที่ผู้วิจัยไม่ต้องนั่งพิมพ์หรือจัดรูปแบบใหม่ทุกครั้งที่ต้องการใช้ โปรแกรมยังสามารถคัดลอกข้อมูลบรรณานุกรมจากภายนอก เช่น ห้องสมุด ฐานข้อมูลงานวิจัยต่างๆ หรือจากโปรแกรมอื่นๆ ได้ โดยไม่ต้องเสียเวลาพิมพ์รายการบรรณานุกรมเข้าไปใหม่

# การติดตั้งโปรแกรม EndNote

**การติดตั้งโปรแกรม EndNote**

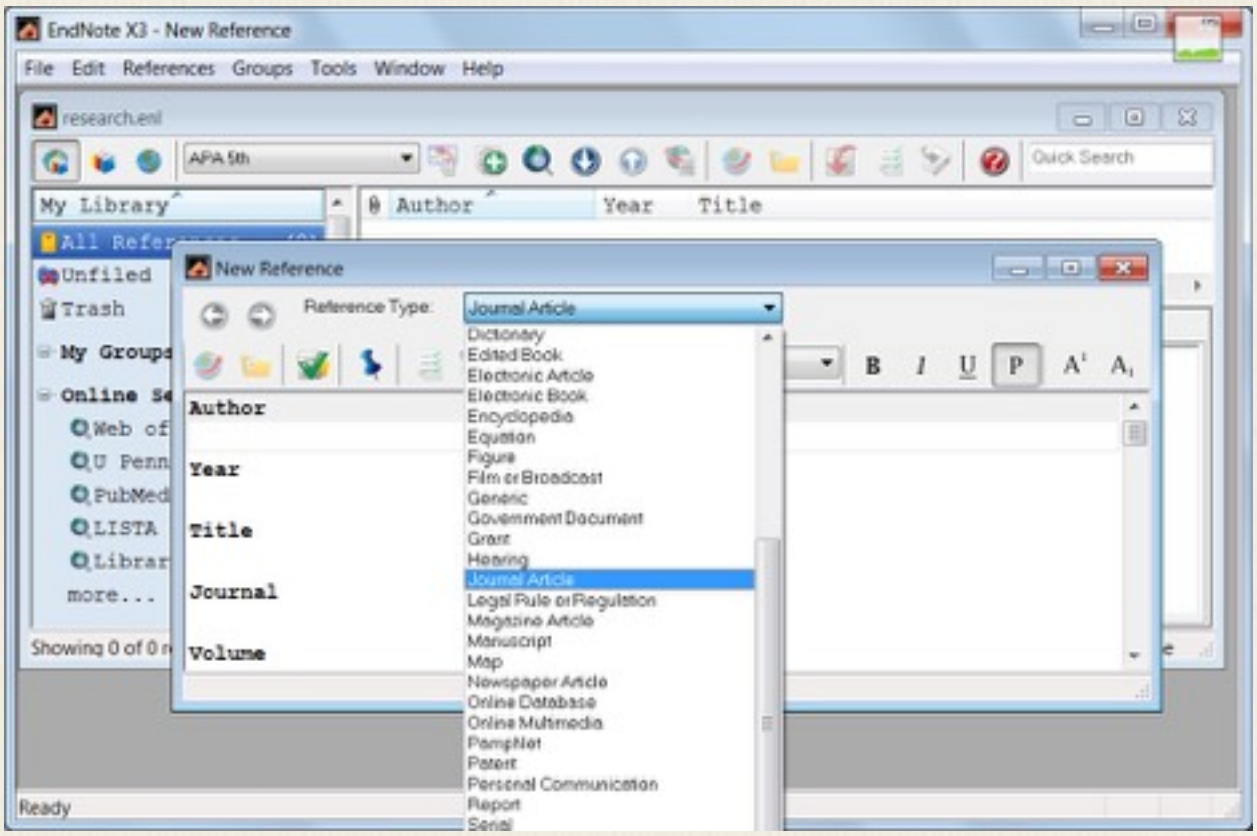
ให้เข้าไปโหลดโปรแกรมได้ที่ https://www.car.chula.ac.th/endnote.php เนื่องจากจุฬาฯ ได้ซื้อสิทธิ์การใช้โปรแกรม Endnote X9 มาให้อาจารย์และนิสิตได้ใช้ จึงต้องขอ password จากเจ้าหน้าที่ตามที่ระบุไว้ในหน้าเว็บ (โปรแกรมนี้ให้ใช้สำหรับอาจารย์และนิสิตจุฬาเท่านั้น ห้ามมิให้เผยแพร่แก่บุคคลภายนอก และ password จะเปลี่ยนทุกภาคการศึกษา)  โปรแกรมมีทั้งที่เป็นเวอร์ชั่นสำหรับ Windows หรือ Mac

# การตั้งฐานข้อมูลใหม่

การตั้งฐานข้อมูลใหม่

ให้เลือก File – New ตั้งชื่อตามที่ตัวเองต้องการ เช่น research การสร้างข้อมูลบรรณานุกรมใหม่

เลือก References – New Reference แล้วเลือกประเภทงานที่ตรงกับงานชิ้นนั้น เช่น เป็น Journal Article, Book, Edited Book, etc. แล้วกรอกข้อมูลใน field ต่างๆ เข้าไปเอง

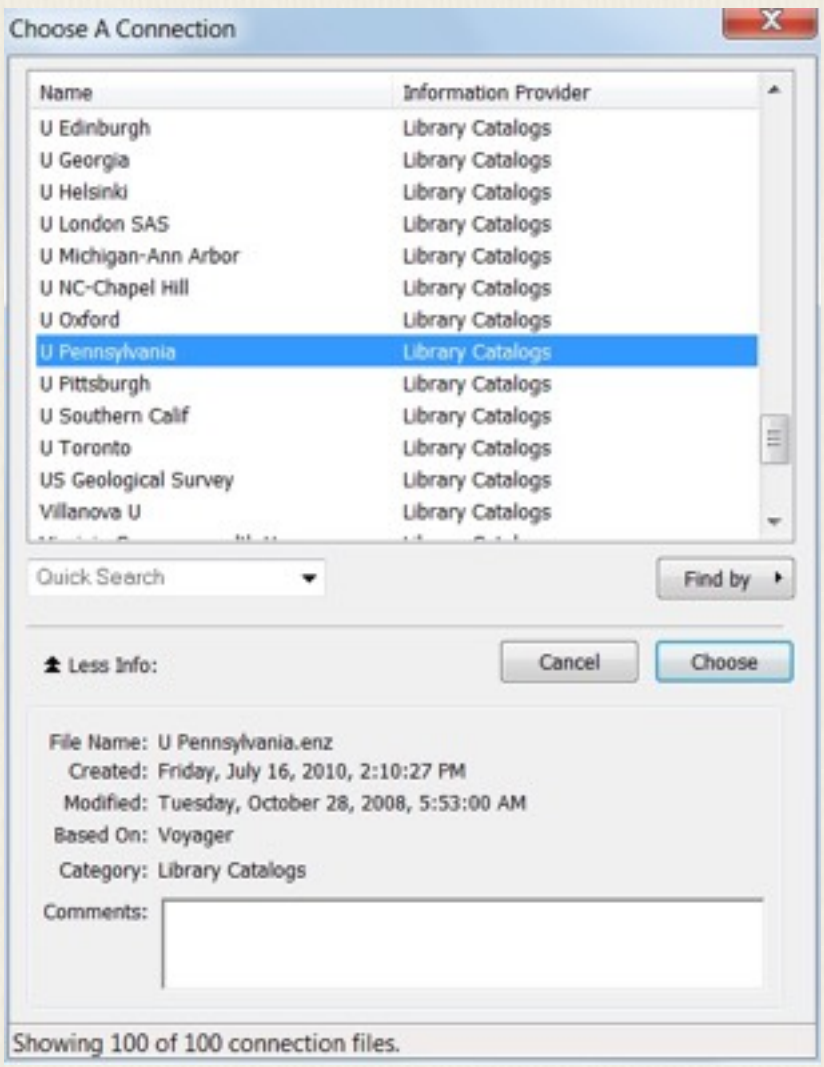



กรอกข้อมูลต่างๆ ใน field ที่เห็น หากเลือกประเภทรายการนั้นผิด ก็สามารถเลือกเปลี่ยนประเภทใหม่ได้ภายหลัง ข้อมูลใน field ต่างๆจะไปอยู่ใน field เดียวกันของประเภทรายการใหม่

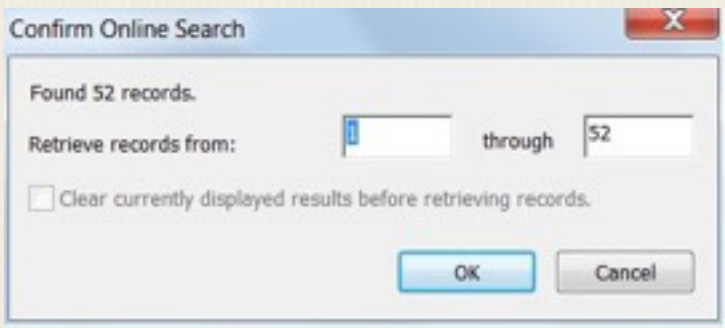
# การคัดลอกข้อมูลบรรณานุกรมจากห้องสมุดออนไลน์

การคัดลอกข้อมูลบรรณานุกรมจากห้องสมุดออนไลน์

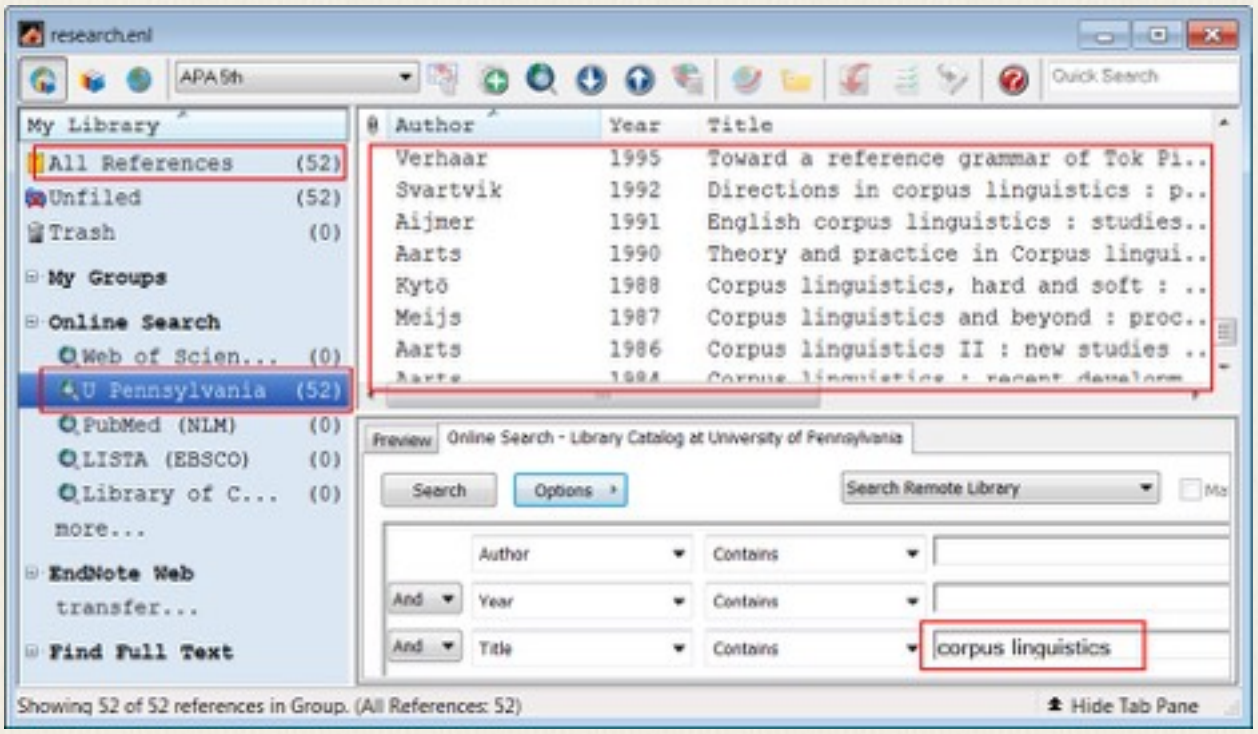
เลือก Tools – Online Search จะเห็นว่ามี online library ที่เราสามารถระบุให้โปรแกรม End-Note เข้าไปค้นรายการในห้องสมุดนั้นได้



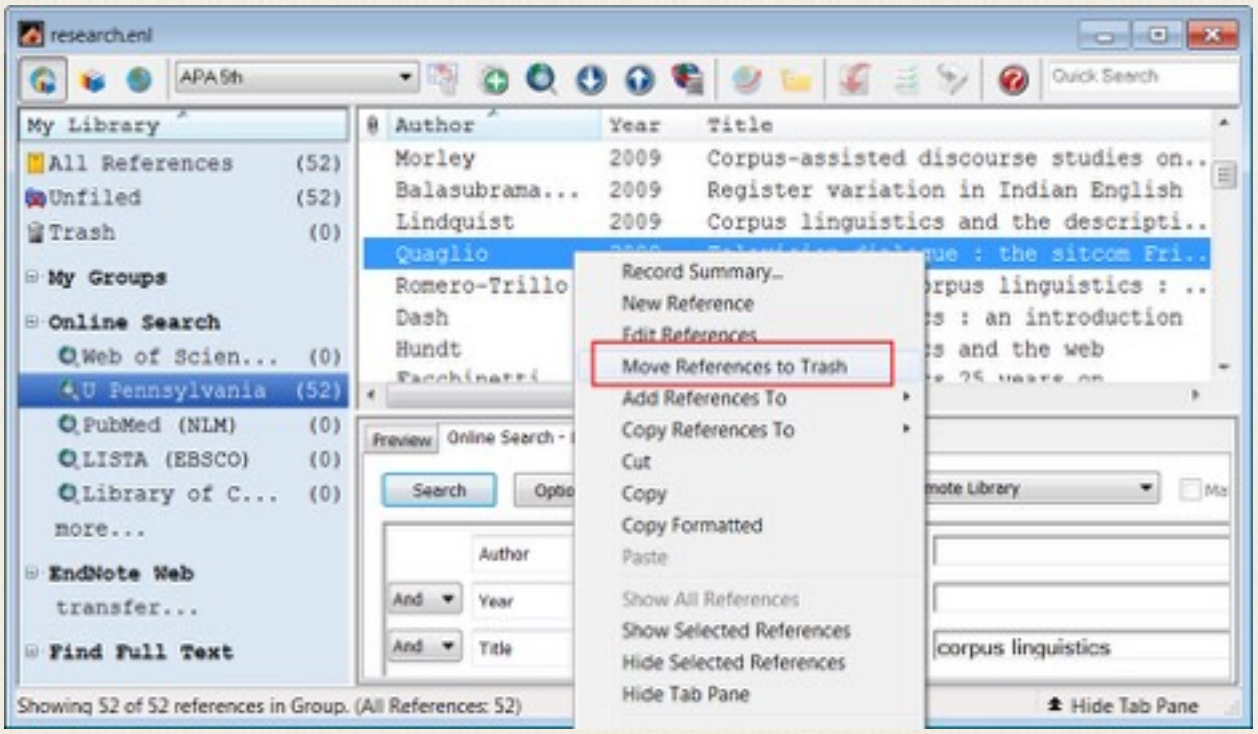

เมื่อเชื่อมต่อได้แล้ว ให้เลือกค้นรายการตามปกติ จะค้นตามชื่อผู้แต่ง ชื่อเรื่อง หรือคำสำคัญก็ได้ หากค้นพบรายการที่ต้องการจะมีหน้าจอบอกจำนวนรายการที่ค้นได้ออกมา



รายการที่ค้นมาได้จะถูกเพิ่มเข้าไปในฐานข้อมูลปัจจุบันที่ใช้อยู่



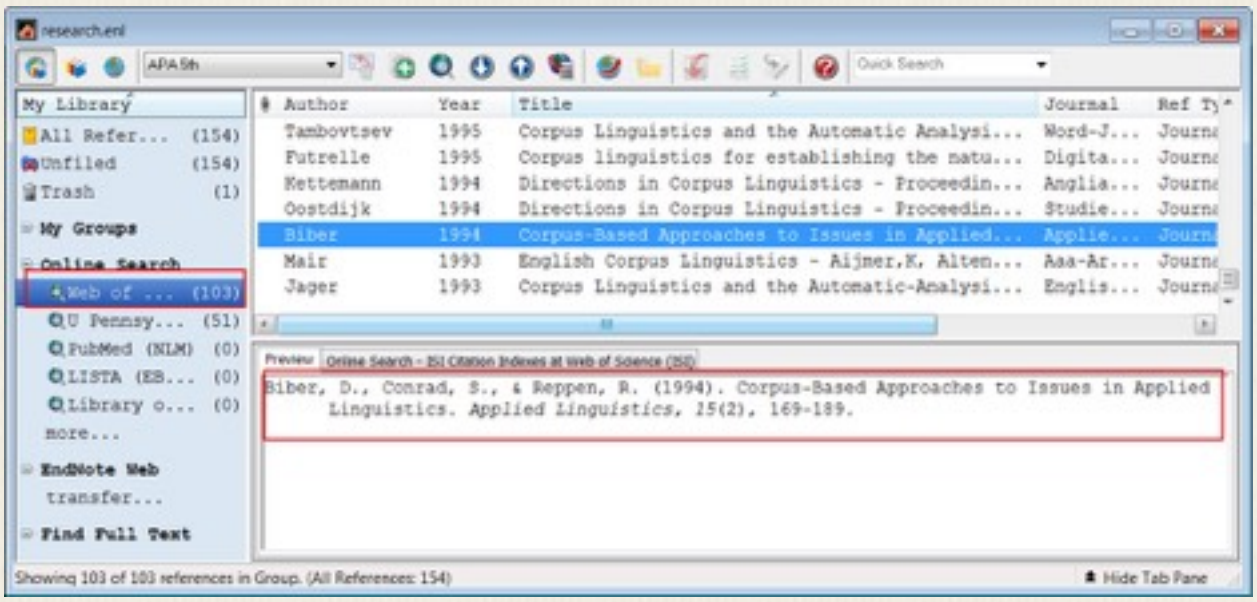

สามารถลบรายการที่ไม่ต้องการจากฐานทิ้งด้วยการเลือกย้ายรายการนั้นลงขยะ หรือกดปุ่ม Ctrl+D หรือ DEL



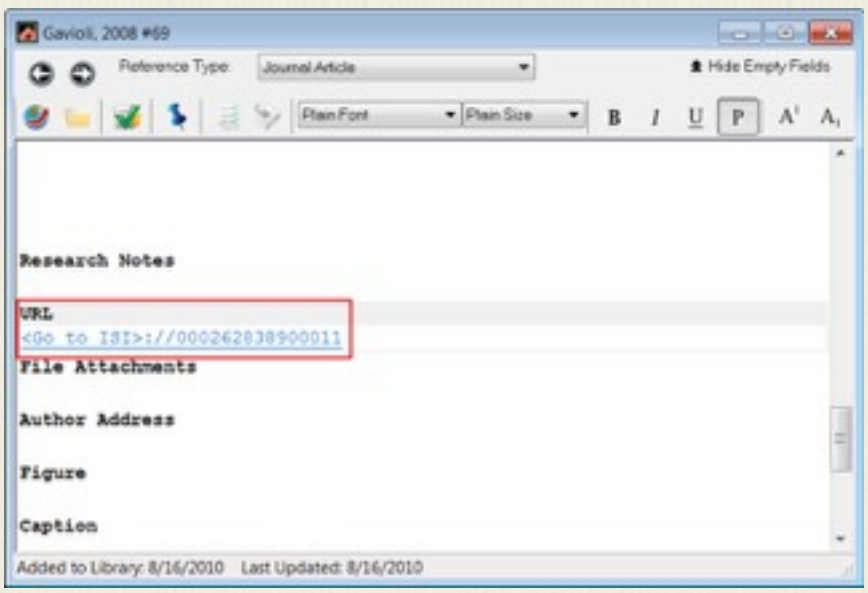

สามารถจัดเรียงข้อมูลตาม field ต่างๆได้ โดยการกดที่ชื่อ field นั้น เช่น กด title จะเป็นการเรียงตามชื่อเรื่อง

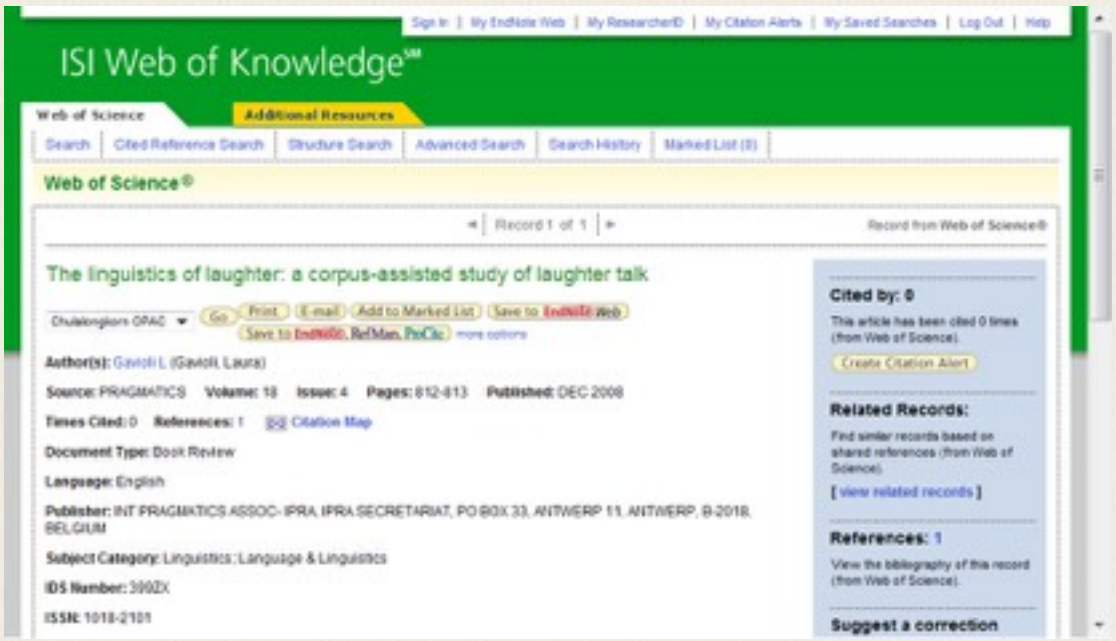
| Author | Year | Title | Journal | Ref Ty |
|---|---|---|---|---|
| Parodi | 2010 | Academic and professional discourse genres i... | Studie... | Book |
| Aijmer | 2004 | Advances in corpus linguistics : papers from... | Langua... | Book |
| Yamamoto | 1999 | Animacy and reference : a cognitive approach... | Studie... | Book |
| Connor | 2004 | Applied corpus linguistics : a multidimensio... | Langua... | Book |
| Renouf | 2006 | The changing face of corpus linguistics | Langua... | Book |
| Murphy | 2010 | Corpus and sociolinguistics : investigating ... | Studie... | Book |
| Facchinetti | 2007 | Corpus linguistics 25 years on | Langua... | Book |
| Dash | 2008 | Corpus linguistics : an introduction | | Book |

หากต้องการค้นหนังสือเราสามารถเลือก Tools - Online Search แล้วเลือกห้องสมุดต่างๆ ที่เปิดให้บุคคลภายนอกค้นผ่านอินเทอร์เน็ตได้ แต่หากต้องการค้นบทความในวารสารให้ไปค้นจากฐานข้อมูลที่มหาวิทยาลัยบอกรับไว้ เช่น Web of Sciences (ISI) ก็จะได้รายการบทความต่างๆ มาเก็บใน End-Note ทันที

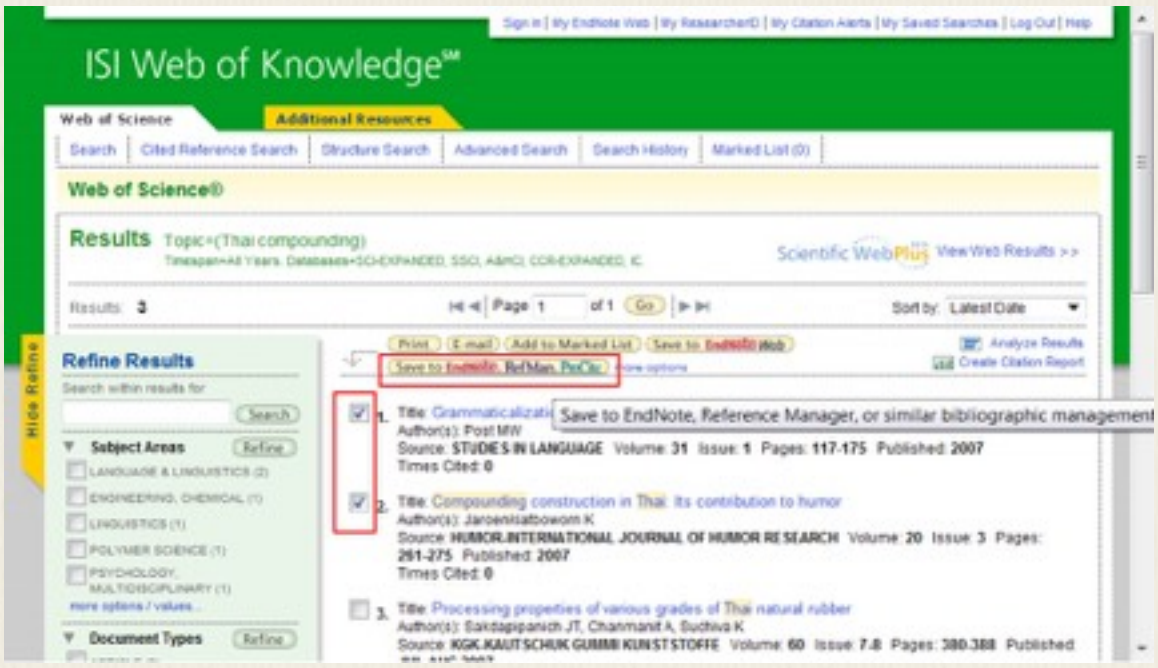


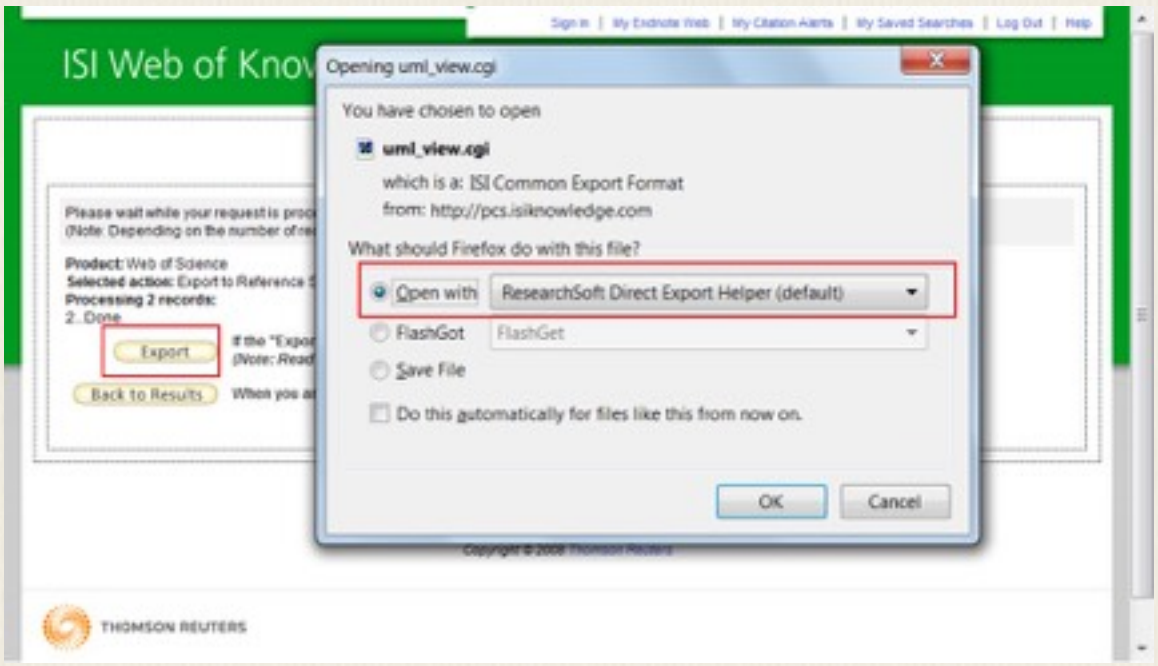

หาก click เข้าไปดูในแต่ละรายการ เราจะเห็นว่ามี link ที่จะโยงเราไปยังเว็บ ISI ได้เลย

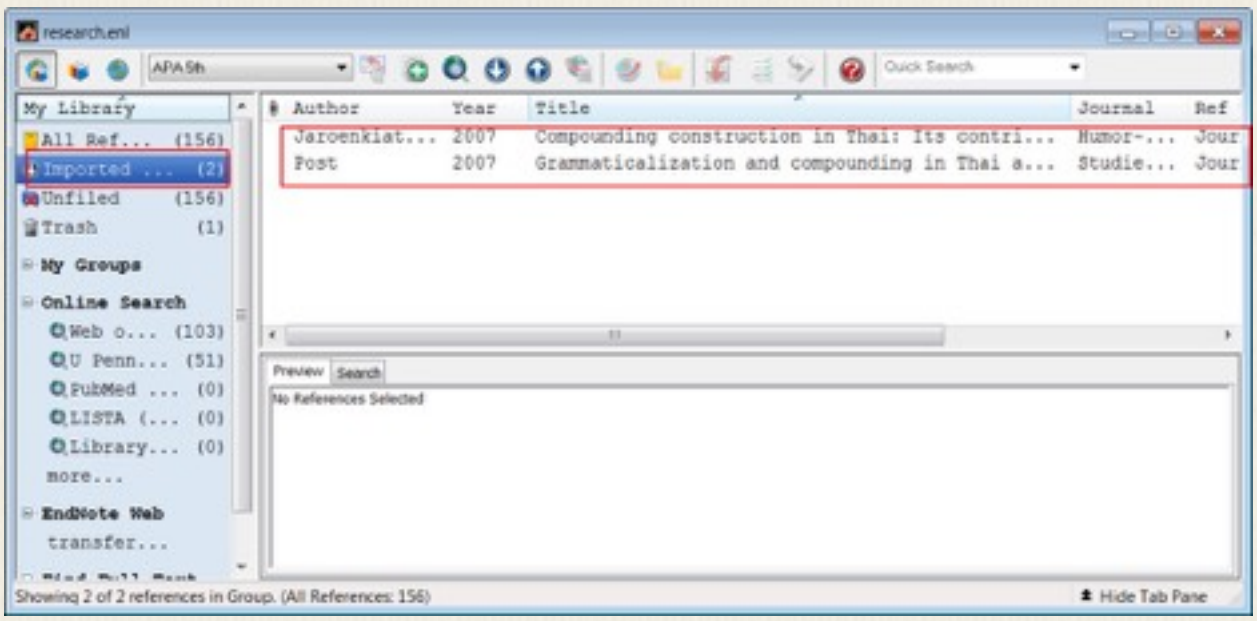




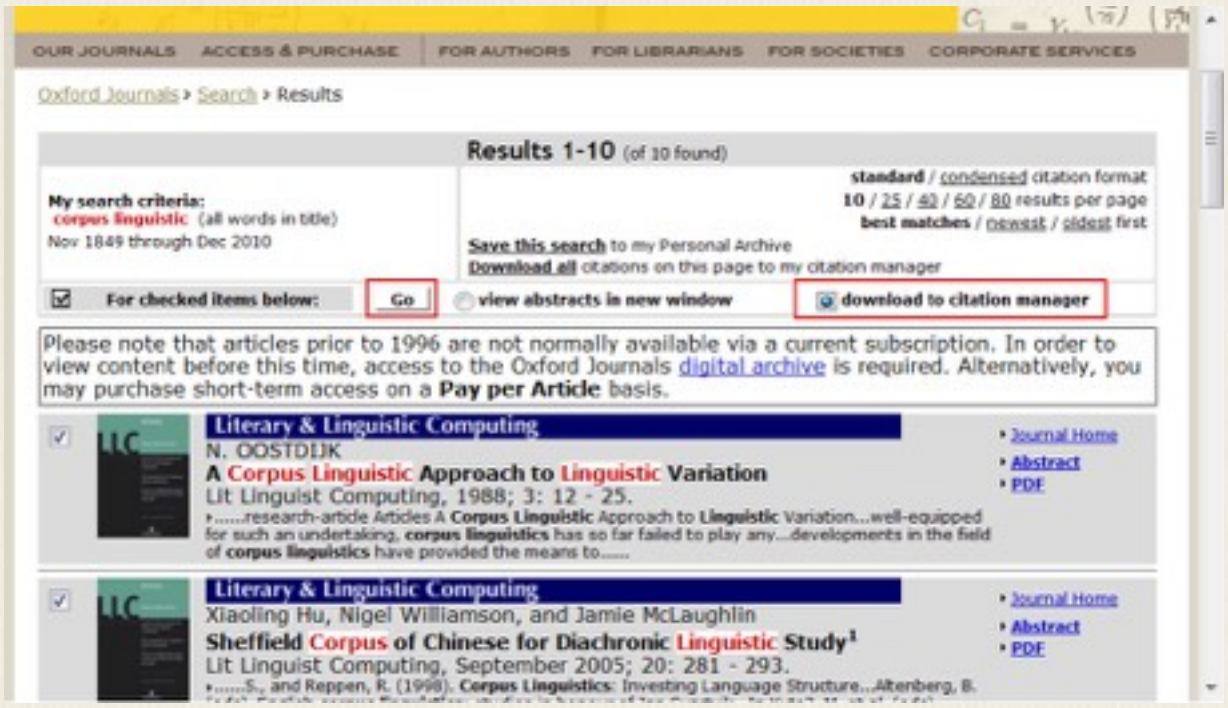

ในฐานข้อมูลบางฐาน อย่างเช่น ISI จะมีตัวเลือกให้เรา export ข้อมูลมา EndNote ได้ หลังจากค้นข้อมูลที่ต้องการได้แล้ว ให้ mark เลือกรายการที่ต้องการ export แล้วให้ open ด้วย ResearchSoft Direct Export ข้อมูลจะถูกนำเข้าใส่ใน EndNote ในฐานข้อมูลที่กำลังเปิดอยู่ทันที

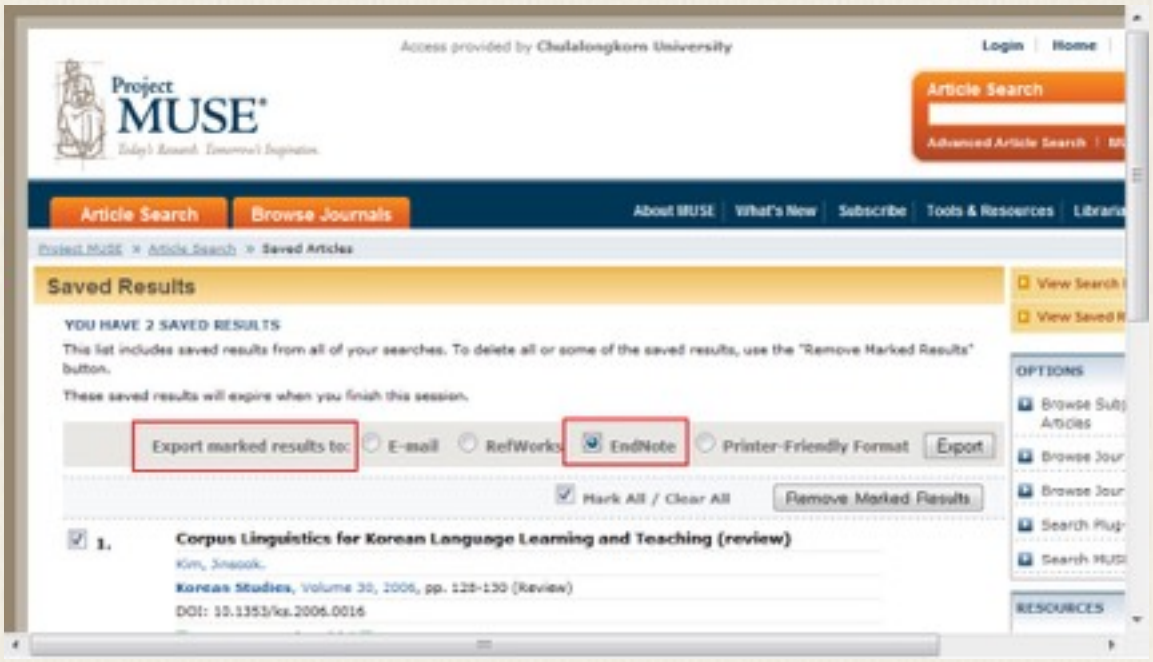


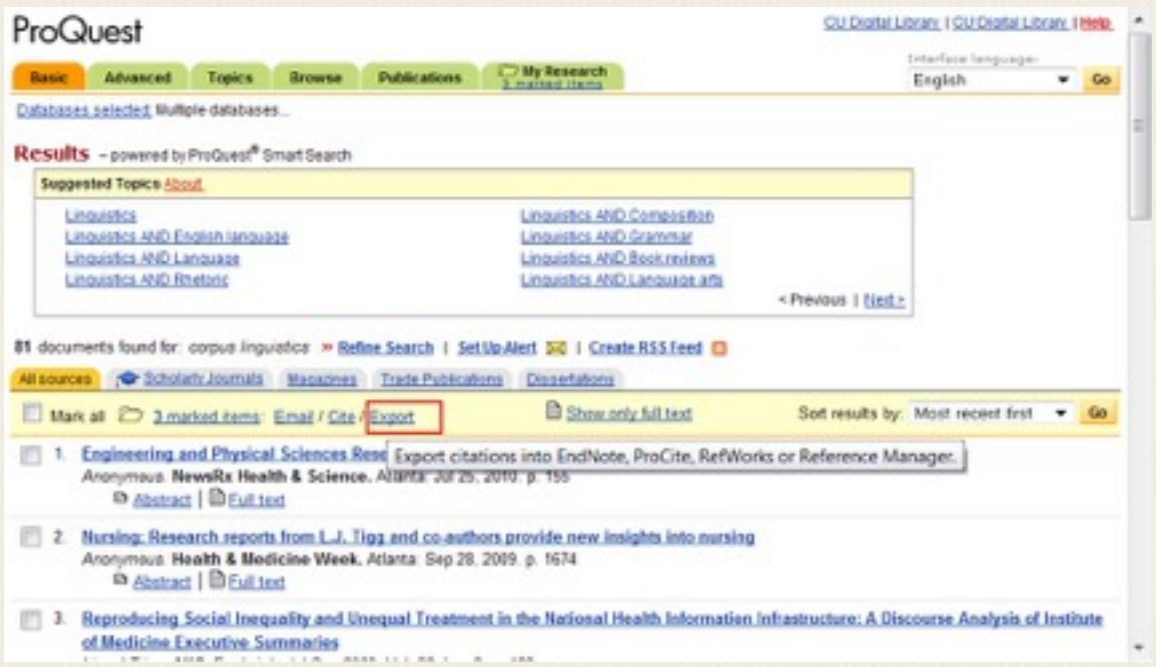



ฐานข้อมูลต่างๆ ตลอดจนวารสารต่างๆ ที่สามารถสืบค้นได้ มักจะมีทางเลือกให้เรา export ข้อมูลบรรณานุกรมออกมาเข้า EndNote ได้ ตัวอย่างเช่น





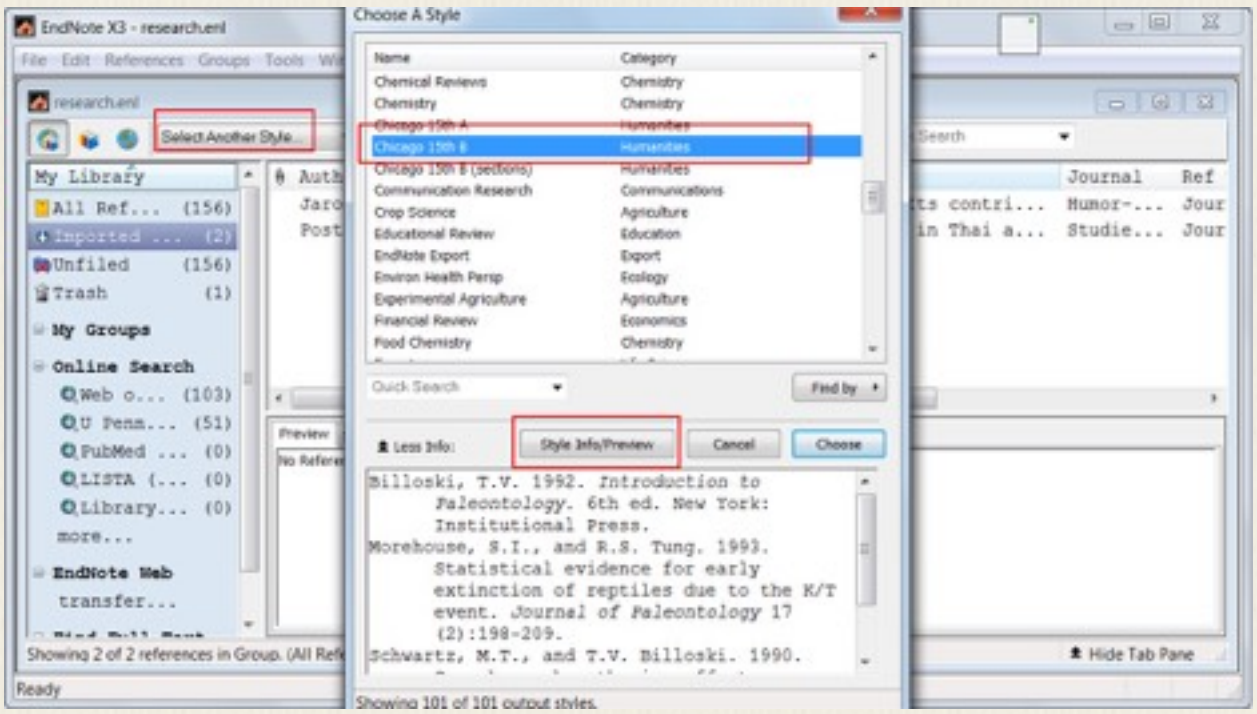




กรณีที่ export เข้า EndNote แบบอัตโนมัติไม่สำเร็จ ให้เลือก save เป็นไฟล์ไว้ ไฟล์ที่ได้ควรมีนามสกุล .enw ถ้าไม่ได้ไฟล์แบบนี้ให้ลองเปลี่ยนนามสกุลเอาเองเป็น .enw เมื่อเปิดไฟล์นี้ ข้อมูลจะถูกนำเข้า EndNote ให้เอง

# การเลือกแบบบรรณานุกรม

การเลือกแบบบรรณานุกรม

เราสามารถเลือกแบบบรรณานุกรมที่เราต้องการได้ ขึ้นอยู่กับว่าต้องการทำรายการอ้างอิงเพื่อตีพิมพ์ในวารสารใด เช่น Language, IEEE Proceedings ฯลฯ ให้เลือก select another style แล้วเลือกรายการแบบที่ต้องการ กดปุ่ม Style Info/Preview เพื่อดูตัวอย่างที่ออกมาว่าใช่แบบที่ต้องการหรือไม่

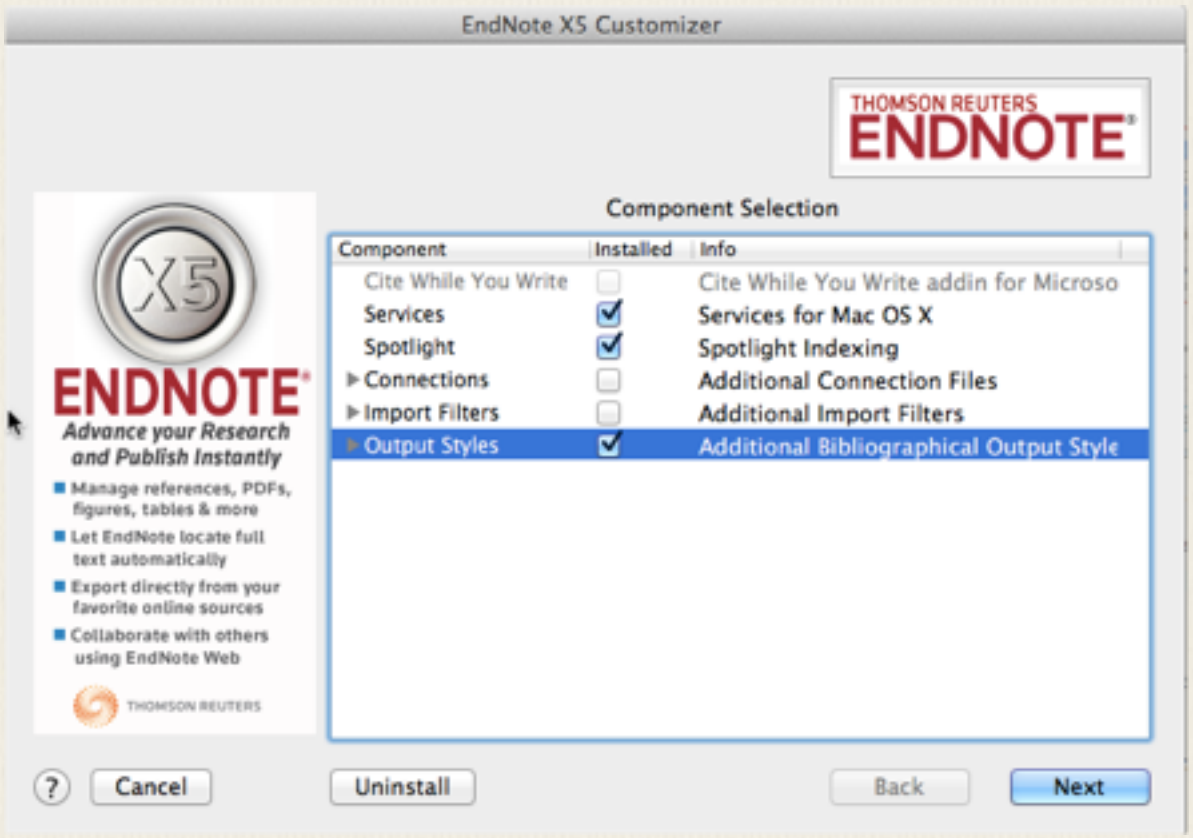



จำนวน style ที่มาเริ่มต้นมี 101 แบบ หากต้องการติดตั้ง style ใหม่เพิ่ม ให้เข้าไปที่ Control Panel ไปในส่วนการ uninstall โปรแกรม เลือก EndNote แล้วเลือก modify โปรแกรม และ add style ในสาขาที่ต้องการเพิ่ม

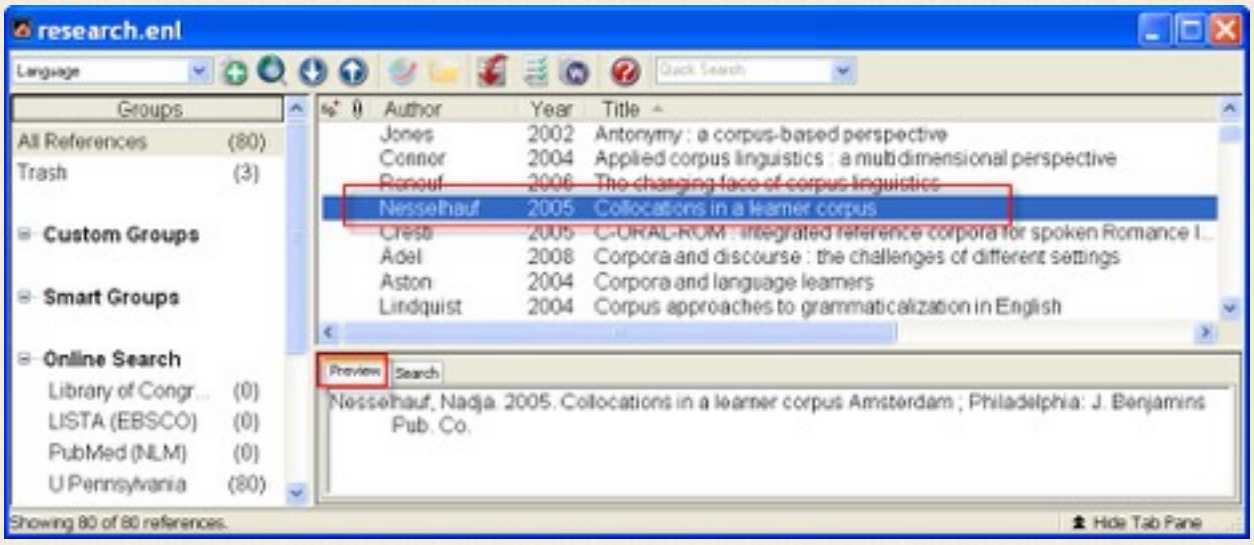


ในระหว่างการติดตั้งนี้ โปรแกรมจะถามหาไฟล์ที่ใช้ในการติดตั้งอย่าง ENX3Inst.msi ซึ่งอาจถูกลบทิ้งไปหลังการติดตั้ง หากเป็นเช่นนั้น ให้ install โปรแกรม EndNote ใหม่ และในระหว่างติดตั้งนี้ ให้ดูว่า temporary folder อยู่ที่ไหน ให้เข้าไป copy ไฟล์นี้ออกมาก่อนที่จะถูกลบทิ้ง จากนั้นจึงเริ่มขั้นตอนการ modify โปรแกรมใหม่

สำหรับ EndNote X5 (Mac) ให้เลือก EndNote - Customizer จะมีหน้าจอขึ้นให้เลือกเพิ่ม Output Style




สไตล์ต่างๆ ที่เลือกใช้สามารถ preview ดูได้ว่าใช่แบบการอ้างอิงที่ต้องการหรือไม่ โดยเลือกแท็บ pre-view จะเห็นรายการสิ่งที่ถูกเลือกไว้



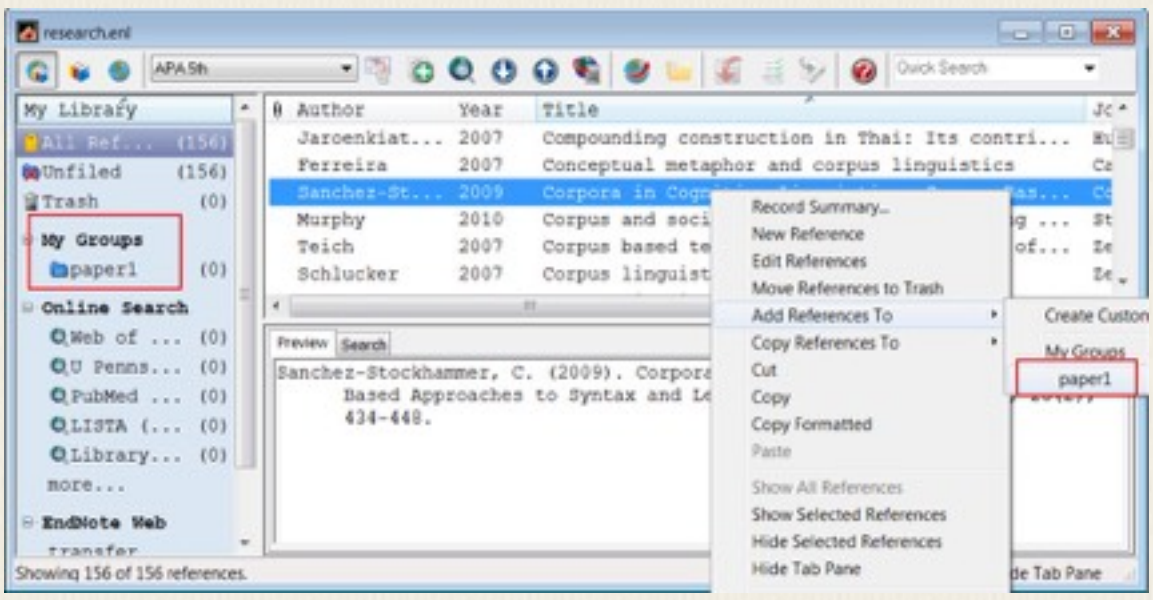
# การสร้างกลุ่มรายการอ้างอิง

การสร้างกลุ่มรายการอ้างอิง

ในการเขียนแต่ละบทความเรามักจะอ้างอิงเฉพาะบางรายการที่เกี่ยวข้อง เราจึงควรสร้าง group สำหรับแต่ละงานขึ้นมา รายการบรรณานุกรมที่อยู่ใน reference สามารถ copy ไปยัง group ต่างๆ ได้

เลือก Groups – Create Group จะเห็นว่าเราสามารถใส่ชื่อ group ใหม่ได้ เช่น ใส่ว่า paper1 สำหรับเก็บ references ที่ใช้ในการเขียน paper1

เลือกใส่รายการที่ต้องการจาก main references โดยเลือกรายการที่ต้องการ กดเม้าส์ปุ่มขวา เลือก Add reference to – paper1 จะเห็นว่า paper1 จะมีจำนวนรายการเพิ่มขึ้นอีก 1 ทำแบบนี้กับทุกรายการที่ต้องการใช้สำหรับ paper1 (สามารถ add หลายรายการพร้อมกันก็ได้ ให้เลือกทีเดียวหลายรายการ กด Ctrl+ mouse click)




รายการที่เลือก add เข้า group ก็จะยังคงอยู่ใน All references เราสามารถลบรายการที่ไม่ต้องการใน group ออกได้ โดยไม่มีผลไปการลบรายการออกจากฐานข้อมูลหลัก ถ้าต้องการลบจากฐานข้อมูลหลักต้องเข้าไปที่ All references แล้วเลือกรายการที่ต้องการลบ

Group ใดที่เมื่อเลิกใช้แล้ว ก็สามารถลบทิ้งได้ เพราะจะไม่มีผลต่อรายการที่เก็บในฐานข้อมูลหลัก (All References)

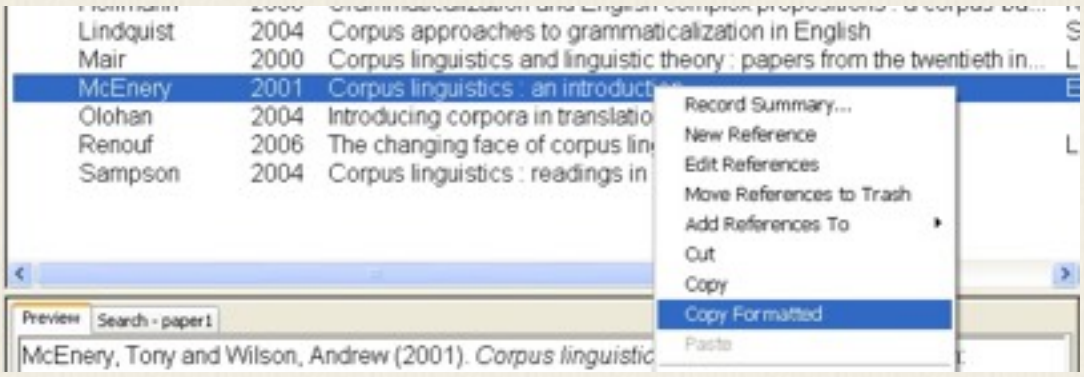
# การสร้างรายการบรรณานุกรม

การสร้างรายการบรรณานุกรม

เลือก group ที่ต้องการสร้างรายการบรรณานุกรม เช่น paper1

เลือก Edit – Select All และ Edit – Copy Formatted จากนั้นไปยัง paper ที่เราเขียนอยู่ paste ลงตำแหน่งที่ต้องการ

ถ้าต้องการเลือกเฉพาะบางรายการ กด right-click ที่รายการนั้น เลือก Copy Formatted




การเลือกแต่ละ style จะให้ format ที่แตกต่างกัน เช่น ถ้าเลือกวารสาร Language in Society จะได้

McEnery, Tony and Wilson, Andrew (2001). Corpus linguistics : an introduction. Edinburgh: Edinburgh University Press.

ถ้าเลือกวารสาร language จะได้

McEnery, Tony & Andrew Wilson. 2001. Corpus linguistics : an introduction Edinburgh: Edinburgh University Press.

เมื่อเขียนบทความเสร็จและได้รายการบรรณานุกรมที่เกี่ยวข้องลงใน group แล้ว ให้ไปที่ group นั้น เลือก Edit-Select All แล้วเลือก Edit-Copy Formatted จากนั้นไปที่ไฟล์ MS Word ของ

บทความนั้น แล้ว paste ลงในส่วนบรรณานุกรมของบทความก็จะได้รายการทั้งหมดใน format ที่ถูกต้องทันที

# การใช้งานร่วมกับโปรแกรม Word

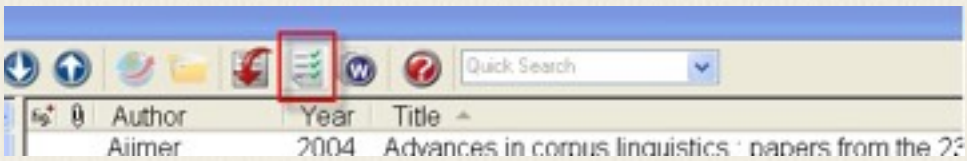
การใช้งานร่วมกับโปรแกรม WORD

เราสามารถใส่รายการอ้างอิงและบรรณานุกรมไประหว่างการทำงานได้ทีละรายการ โปรแกรม EndNote version นี้สามารถทำงานร่วมกับ MS Word 2003 และ 2007 ได้

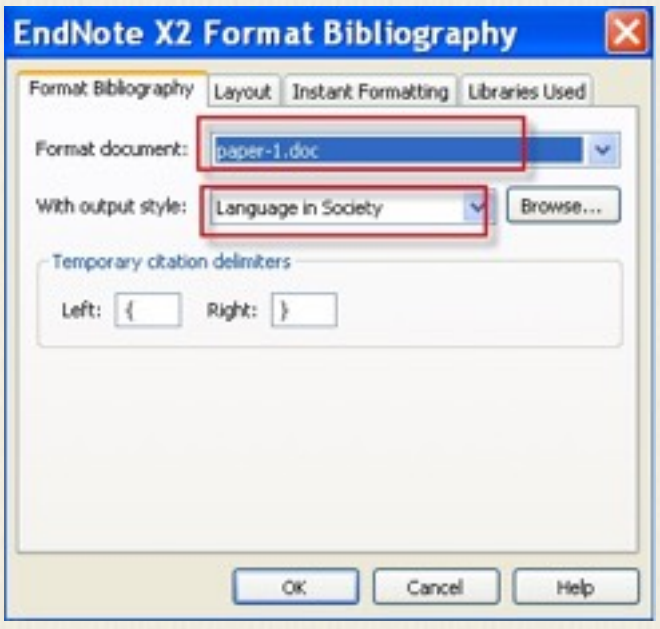
ผู้ที่ใช้ Office 2010 จะพบว่าเมื่อติดตั้งโปรแกรม EndNote แล้ว MS Word จะ hang หลังจากเปิดทำงาน ทั้งนี้เพราะ macro จากโปรแกรม EndNote วิธีแก้ให้เปิด Word โดยไม่เรียก macro โดยเรียกจาก command line “winword /a” เมื่อเข้าโปรแกรม Word 2010 แล้ว ไปที่ File – Options – Add-Ins แล้วเลือก Manage Com Add-Ins ในรายการ macro ที่มาจาก EndNote ให้ remove macro นั้นออกไป จากนั้นจึงจะเปิด Word 2010 ได้ตามปกติ

คำอธิบายต่อไปนี้ มาจากตัวอย่างการใช้กับ Word 2007

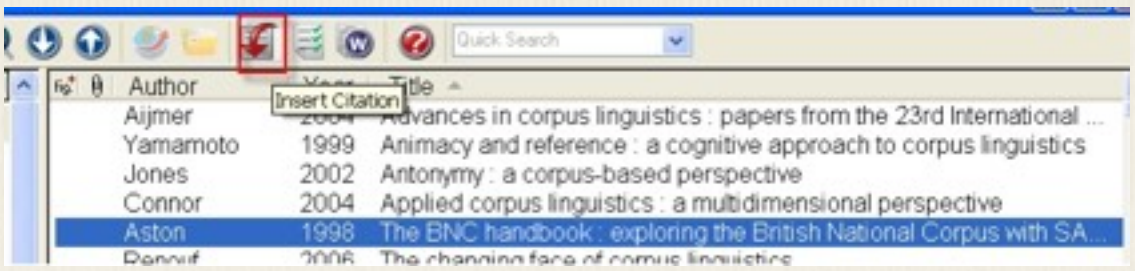
เลือก Format Bibliography



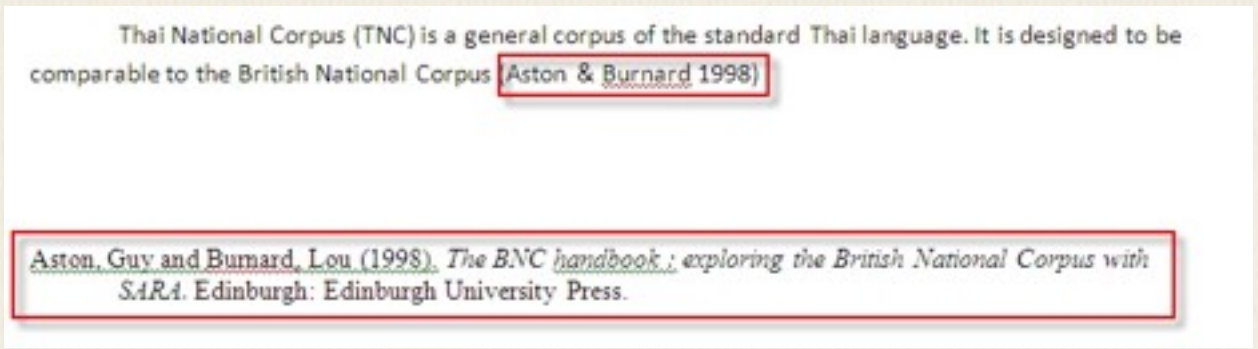

จะได้หน้าจอที่ให้เราเลือกไฟล์ Word ที่เราต้องการทำงาน และเลือกแบบบรรณานุกรมที่ต้องการ



ในระหว่างการพิมพ์งานเราสามารถ insert reference ไปได้เลย ทั้งตัว reference และรายการ reference จะถูกนำไปใส่ในไฟล์ Word โดยอัตโนมัติ โดยเลือกรายการที่ต้องการแล้วเลือกปุ่ม Insert Citation

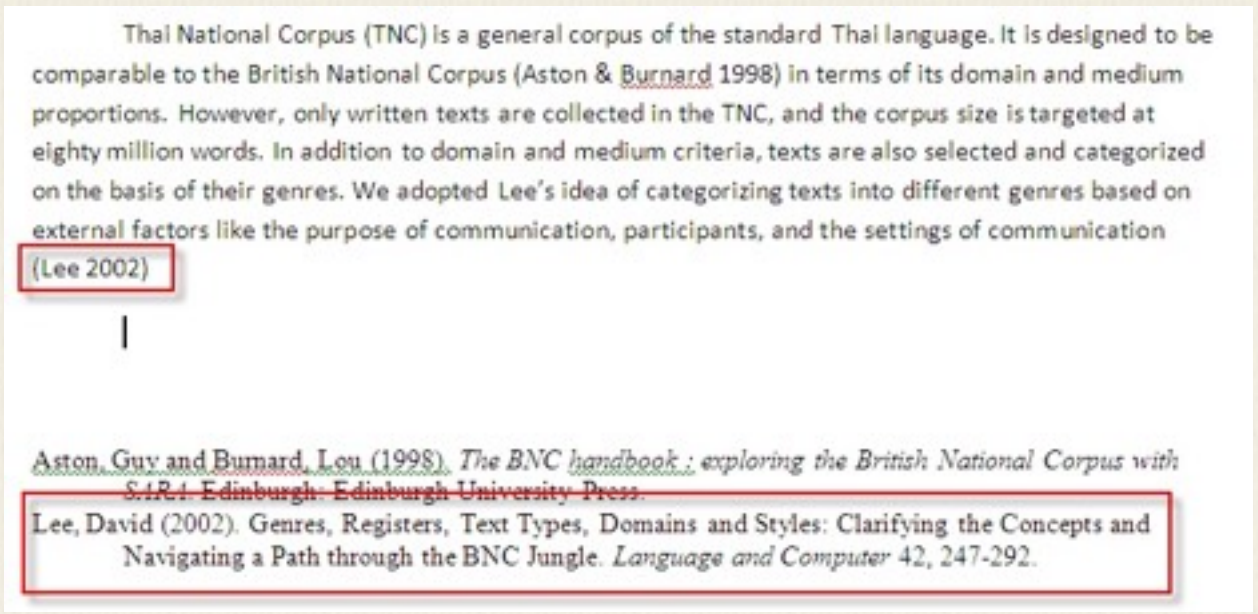



Citation และ reference ตาม format style ที่เราเลือกจะถูก paste ลงในไฟล์ Word ที่เราระบุไว้ใน Format Bibliography


Thai National Corpus (TNC) is a general corpus of the standard Thai language. It is designed to be comparable to the British National Corpus (Aston & Burnard 1998)

Aston, Guy and Burnard, Lou (1998). *The BNC handbook : exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

เมื่อพิมพ์งานต่อไปแล้วมีการอ้างอิงรายการเพิ่มก็ทำต่อได้เรื่อยๆ จนจบบทความ รายการอ้างอิงต่างๆ ก็จะถูกนำมาใส่ในบทความนั้น ตรงตาม format ที่เราได้เลือกไว้


Thai National Corpus (TNC) is a general corpus of the standard Thai language. It is designed to be comparable to the British National Corpus (Aston & Burnard 1998) in terms of its domain and medium proportions. However, only written texts are collected in the TNC, and the corpus size is targeted at eighty million words. In addition to domain and medium criteria, texts are also selected and categorized on the basis of their genres. We adopted Lee's idea of categorizing texts into different genres based on external factors like the purpose of communication, participants, and the settings of communication (Lee 2002)

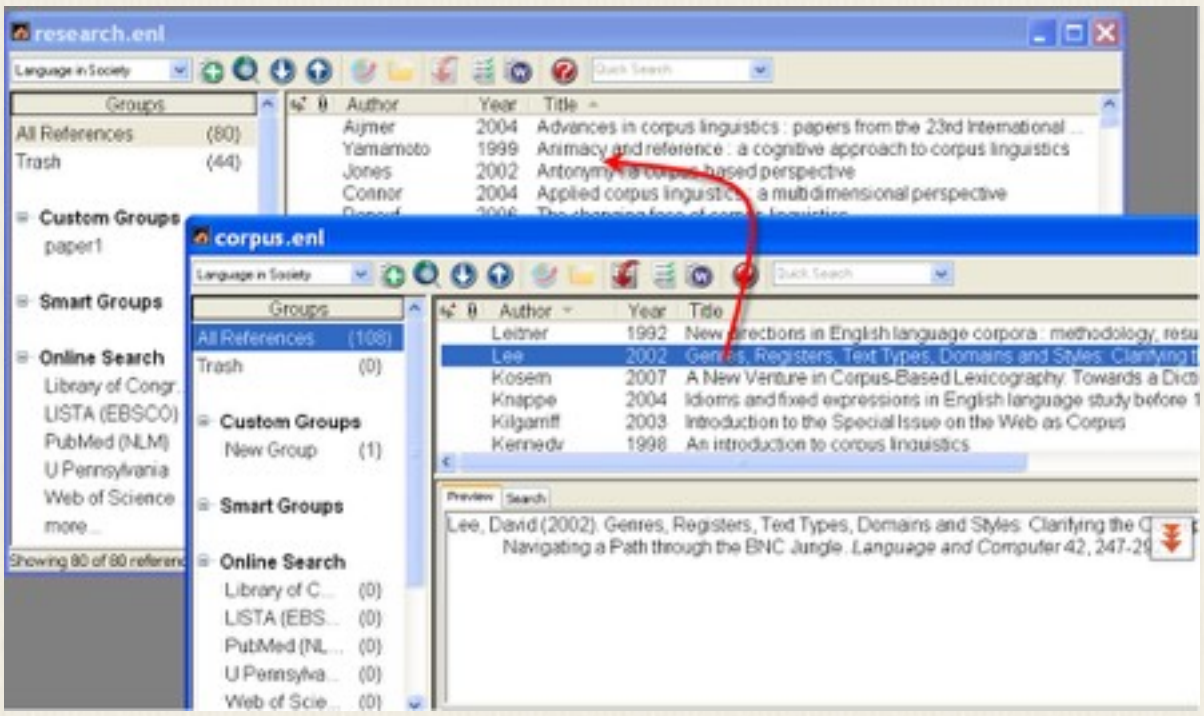
Aston, Guy and Burnard, Lou (1998). *The BNC handbook : exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Lee, David (2002). Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language and Computer* 42, 247-292.

# การย้ายข้อมูลระหว่างฐานข้อมูล

การย้ายข้อมูลระหว่างฐานข้อมูล

เราสามารถเปิดฐานข้อมูลมากกว่าหนึ่งฐานขึ้นมาใช้ได้ และสามารถ copy รายการจากฐานหนึ่งไปยังอีกฐานหนึ่งได้ ให้เลือกรายการที่ต้องการ แล้วกดลากไปยังอีกฐานหนึ่ง



CHAPTER 2

# โปรแกรม Zotero

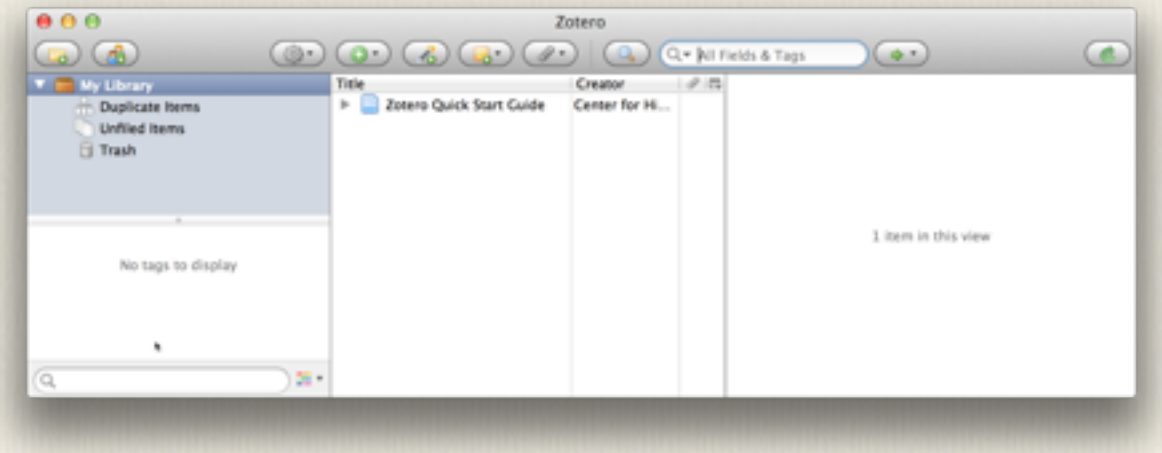
โปรแกรม Zotero เป็นโปรแกรมช่วยจัดเก็บข้อมูลบนเว็บที่พบ เดิมใช้กับ Firefox ทำให้สามารถคัดลอกข้อมูลบรรณานุกรมจากเว็บได้โดยตรง และสามารถย้ายข้อมูลไปเข้าใน EndNote ได้ ปัจจุบัน มีทั้งแบบที่เป็น plugin ของ Firefox และแบบที่เป็นโปรแกรมเอกเทศที่เชื่อมโยงกับ browser อื่น คือ Chrome, Safari หากเลือกแบบโปรแกรม add-on ผู้ใช้จะต้องมีโปรแกรม Firefox ก่อน ซึ่งเป็น web browser ตัวหนึ่ง เมื่อติดตั้ง Firefox แล้ว ให้ไปที่ http://www.zotero.org/ เพื่อโหลดโปรแกรม addon Zotero นี้

Zotero สามารถทำสิ่งต่างๆ เหล่านี้ได้ขณะที่ browse internet : ดึงเอา citation ออกมาโดยอัตโนมัติ เก็บไฟล์ pdf, web และรูปภาพ สร้าง library ของตัวเองโดยการเก็บข้อมูลและงานต่างๆ ที่พบระหว่างการใช้อินเทอร์เน็ต

# การติดตั้งโปรแกรม Zotero Standalone

## การติดตั้งโปรแกรม Zotero Standalone

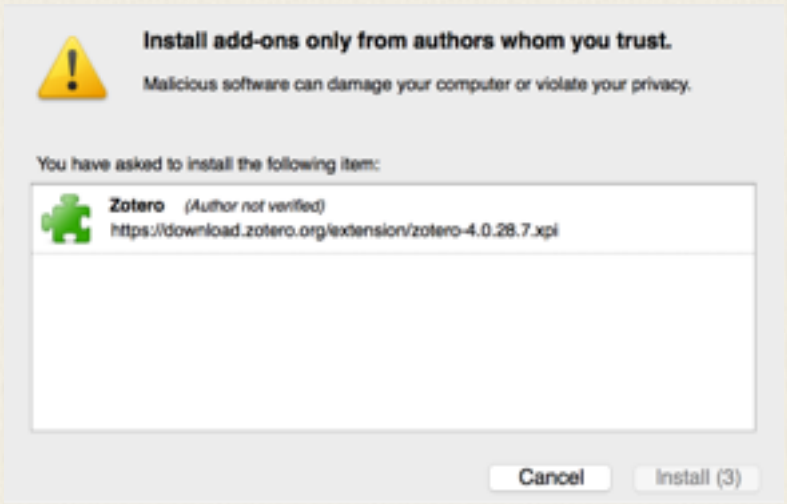
ในกรณีที่ไม่ต้องการใช้ Firefox แต่จะใช้ Chrome หรือ Safari ไปที่ http://www.zotero.org/ เลือก version standalone แล้วติดตั้งโปรแกรม exe (Windows) หรือ dmg (Mac OS X) เมื่อเรียกโปรแกรม Zotero ขึ้นมาจะเห็นหน้าโปรแกรมแบบนี้ จากนั้นให้ติดตั้ง Zotero Connectors ซึ่งก็คือ add on โปรแกรมสำหรับ browser ที่ต้องการใช้ แต่ในที่นี้ จะนำเสนอการใช้ Zotero on Firefox เพราะเป็นโปรแกรมหลักที่พัฒนาและมีความสะดวกในการใช้งานดึงข้อมูลต่างๆบน browser มาเก็บไว้



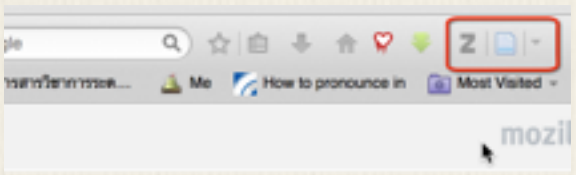
# การติดตั้งโปรแกรม addon Zotero

## การติดตั้งโปรแกรม ADDON ZOTERO

ไปที่ http://www.zotero.org/ เลือก version ใหม่สุด เมื่อกดแล้วจะได้หน้าจอถามให้ติดตั้งหรือไม่




เมื่อติดตั้งแล้วให้ restart โปรแกรม Firefox จะเห็นที่มุมขวามีไอค่อนของ Zotero อยู่

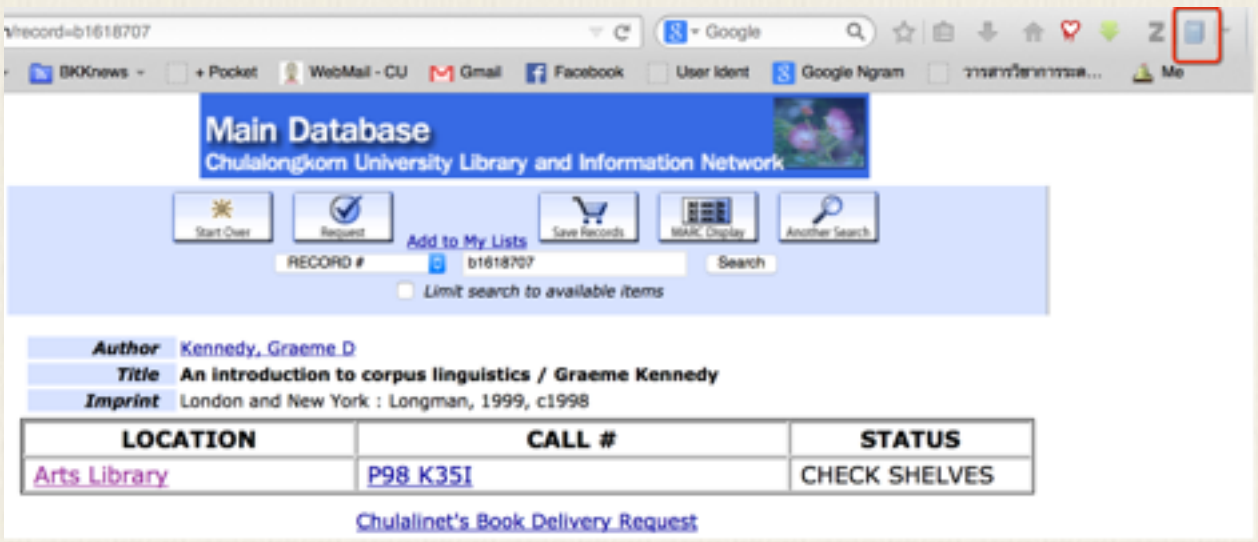


# การเก็บรายการข้อมูล

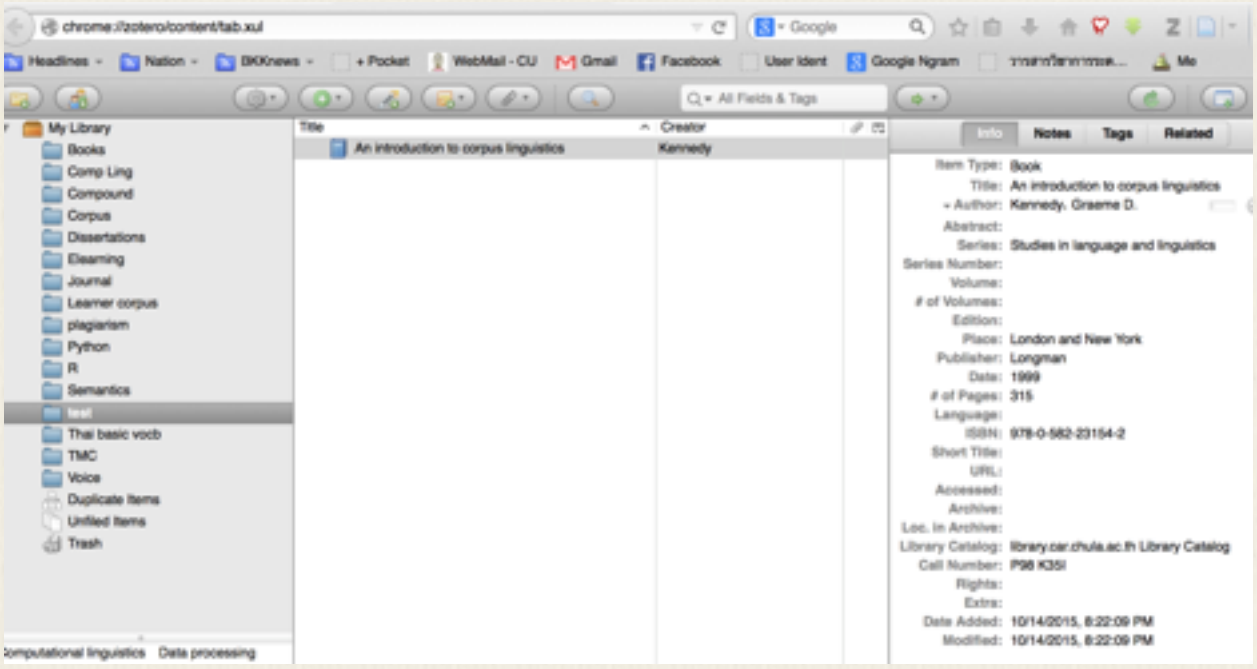
การเก็บรายการข้อมูล

เมื่อเข้าไปยัง web ที่มีรายการหนังสือ เช่น เว็บห้องสมุด เมื่อค้นรายการหนังสือที่ต้องการได้ จะสังเกตุว่าที่ address bar ด้านขวาจะมีไอค่อนรูปหนังสือ หมายความว่าสามารถนำข้อมูลบรรณานุกรมมาเก็บใน

Zotero ได้ ให้กดที่ไอค่อนนี้



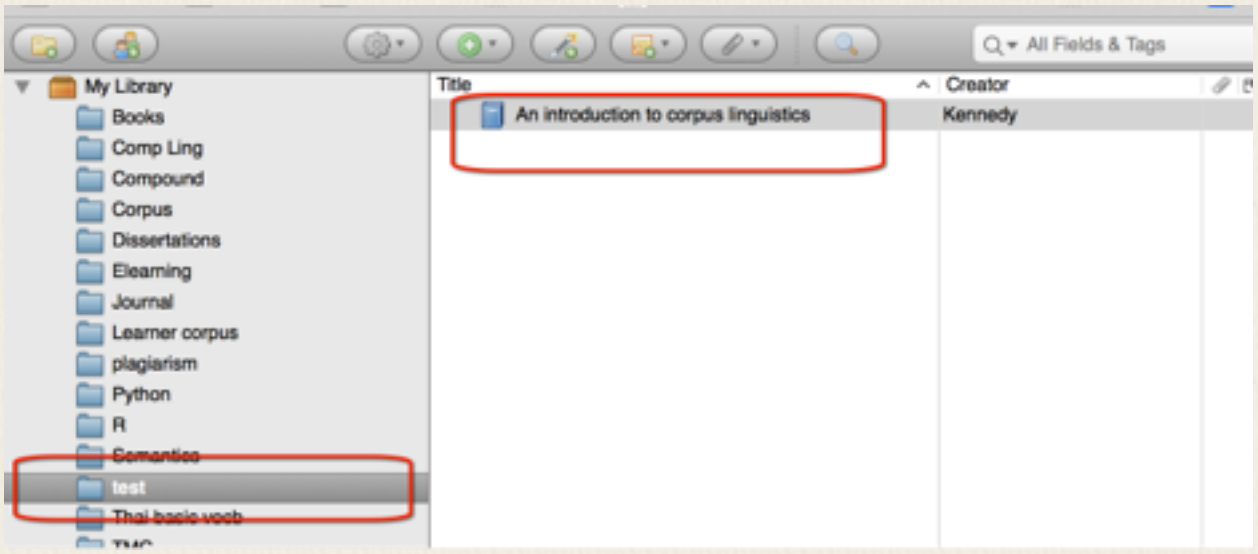
เมื่อกดที่ไอค่อน Zotero ตรงมุมขวาของ Firefox จะเห็นหน้าจอ Zotero ขึ้นมาในอีกแท็บ และรายการหนังสือนี้จะถูกเก็บใน Zotero แล้ว




เราควรสร้าง collection เฉพาะสำหรับงานวิจัยแต่ละเรื่อง ที่มุมซ้ายบนของ Zotero มีไอค่อน New Collection ให้ตั้งชื่อ collection ของงานตนเอง แล้วย้ายรายการข้อมลที่ต้องการไปยัง collection ที่สร้างขึ้น

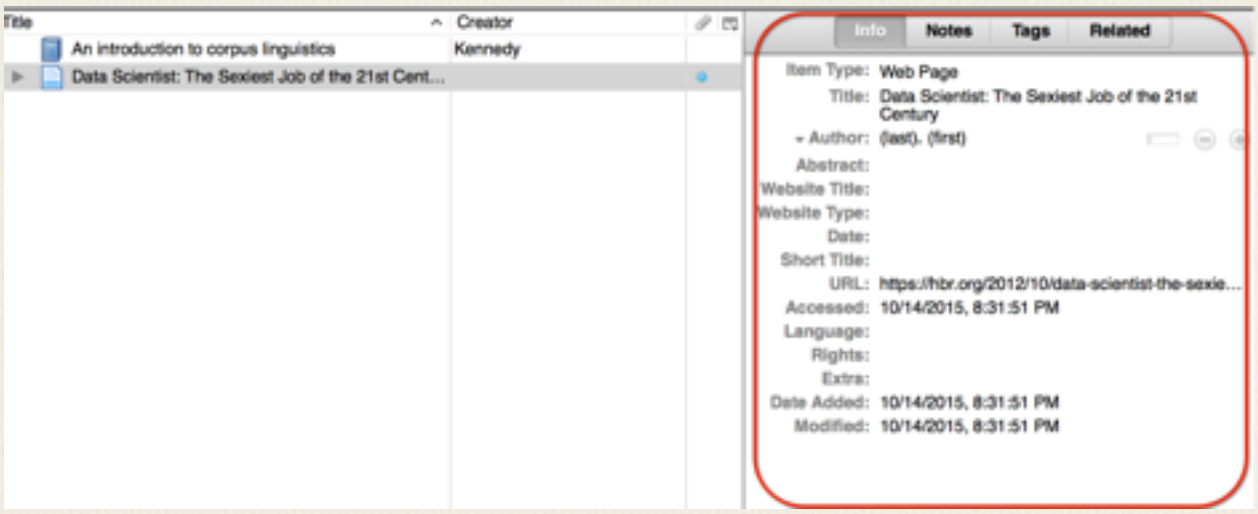
My Library จะเก็บรายการข้อมูลทั้งหมด item เดียวกันสามารถไปอยู่ในหลายๆ collection ได้ collection จึงไม่ได้เก็บข้อมูลรายการจริง แต่คล้ายๆ กับ playlist ใน media player ที่เราเลือกเฉพาะชื่อรายการมาจัดเป็น collection หนึ่งๆ แต่ตัวรายการจริงเก็บไว้อีกที่

เวลาที่ต้องการให้ข้อมูลไปอยู่ใน collection ไหน ก็ให้เปิด Zotero ค้างไว้ที่หน้า collection นั้น เมื่อมีการเก็บรายการข้อมูลใหม่ ชื่อรายการนั้นจะไปอยู่ใน collection ที่เราเลือกค้างไว้เอง



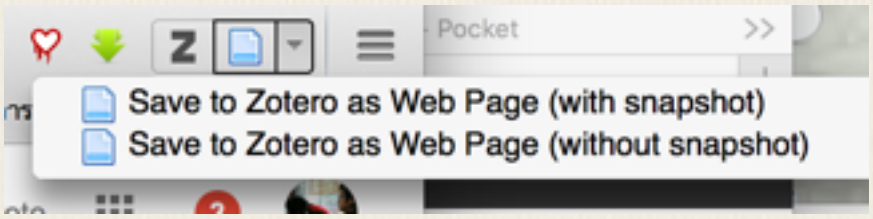

เว็บที่เกี่ยวข้องกับรายการหนังสือ วารสาร เช่น Amazon เว็บห้องสมุดต่างๆ เมื่อเราค้นเจอรายการที่ต้องการ โดยปกติเราจะเห็นไอค่อนรูปหนังสือให้ save รายการเก็บไว้ได้ จึงค่อนข้างสะดวกต่อการเก็บรายการอ้างอิงต่างๆ ไว้ใช้ภายหลัง

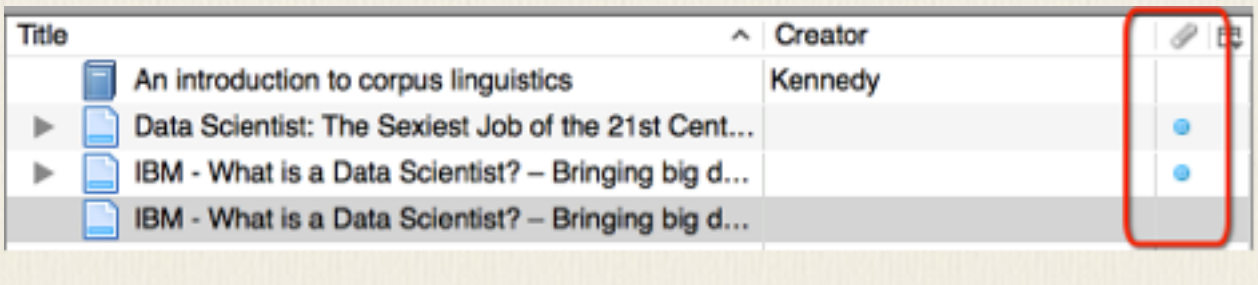
บางเว็บไซ้ต์ เราจะเห็นไอค่อนให้ save to Zotero เป็นรูปอื่นไม่ใช่รูปหนังสือ หมายความว่า เราสามารถเก็บข้อมูลเว็บนั้นลงใน Zotero ได้ ซึ่งจะได้ web address และข้อมูลอื่นที่เกี่ยวข้อง



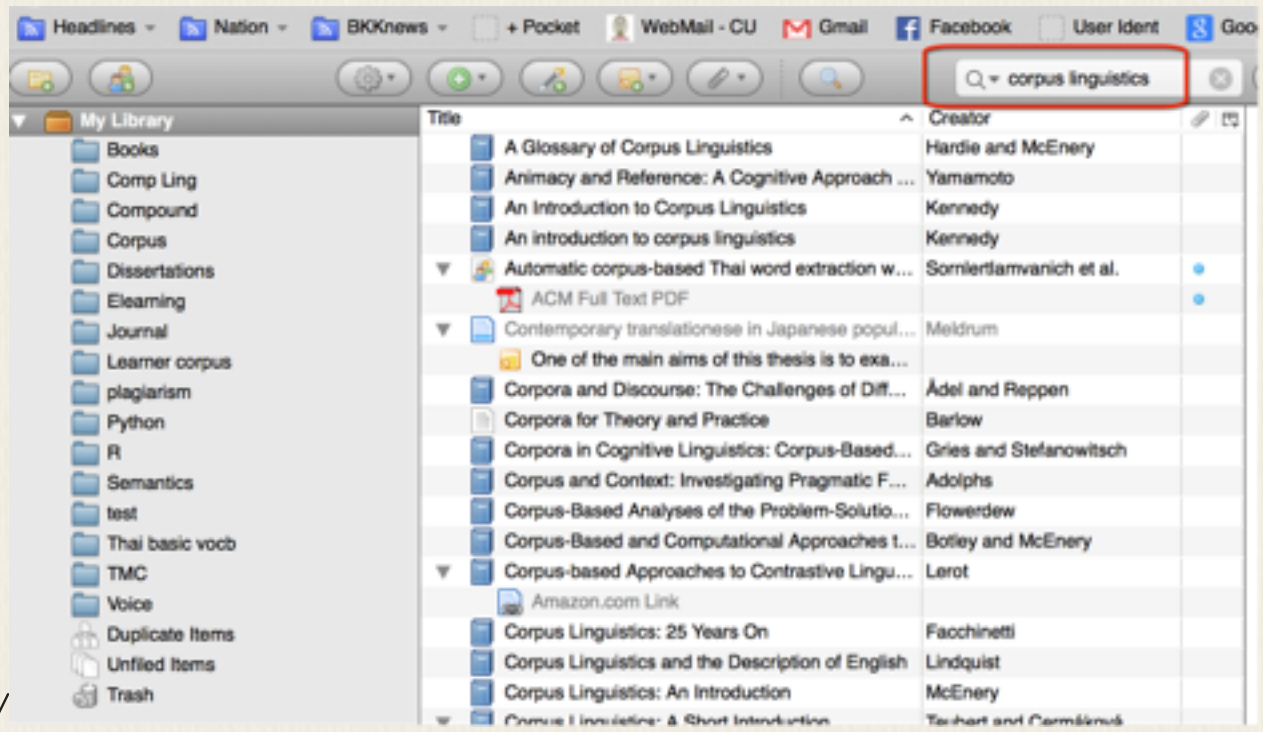

การลบรายการจาก My Library ให้เลือกรายการนั้นแล้วกดปุ่ม DEL ถ้าต้องการเพียงลบออกจาก collection (แต่ยังอยู่ใน My Library) ให้เลือกรายการใน collection นั้น แล้วกด CTRL-DEL

การ save item ลงใน Zotero สามารถเลือกได้ว่าจะ save เนื้อหาหน้าเว็บด้วยไหม (with snapshot) หากดูรายการที่ save ใน Zotero จะเห็นความต่างว่า แบบที่เก็บ snapshot จะมี attachment ติดอยู่ด้วย ข้อมูลที่เก็บแบบ snapshot ไว้จะสามารถมาเปิดดูภายหลังได้ แม้ว่าไม่ได้ต่ออินเทอร์เน็ตอยู่




ในโหมดที่ดู Sanpshot เราจะเห็นว่าข้อมูลนั้นเก็บที่เครื่องเรา ดูได้จาก address จะเป็น zotero://attachment/



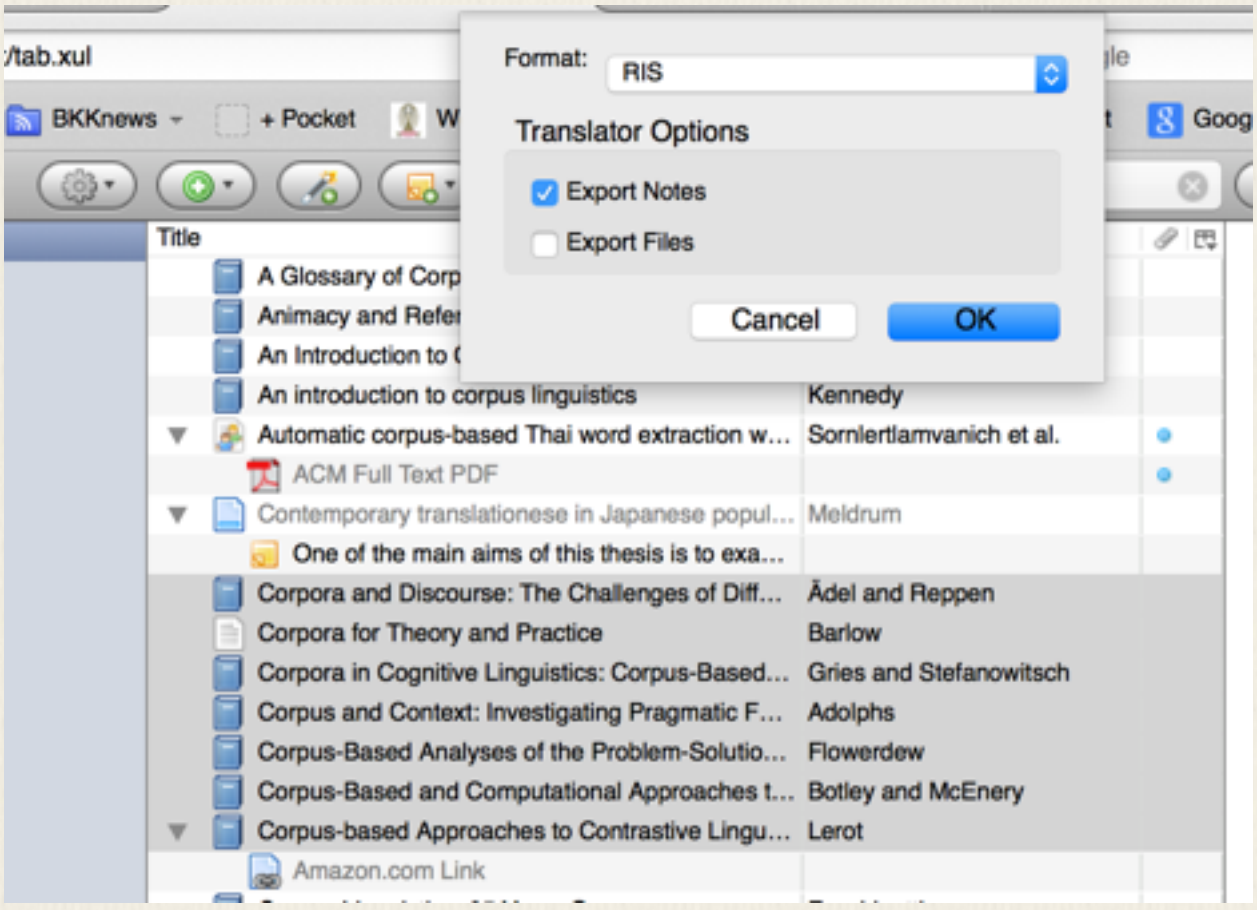

ตัวอย่างสาธิตการใช้งาน Zotero เพื่อเติมรายการข้อมูลต่างๆ ที่พบระหว่างการค้นผ่าน Firefox

# การย้ายข้อมูล Zotero ไป EndNote

## การย้ายข้อมูล Zotero ไป EndNote

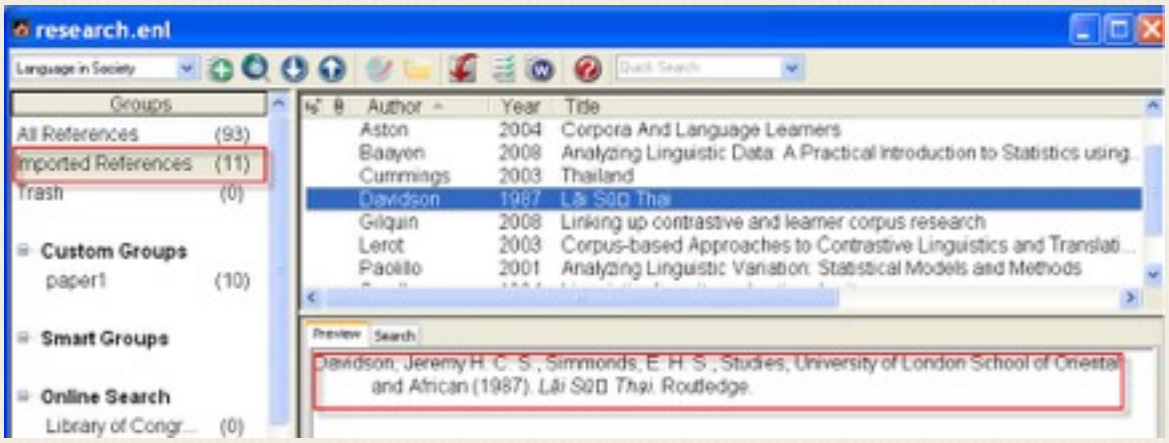
การย้ายข้อมูลบรรณานุกรมจาก Zotero ไปยัง EndNote สามารถทำได้โดยการ export เป็น RIS format แล้วไป import ด้วย EndNote

เลือก Action – Export Library แล้วเลือก format เป็น RIS




เมื่อ save ไฟล์ RIS แล้ว ให้เปิดไฟล์นั้น ข้อมูลต่างๆจะถูกนำเข้าในฐานข้อมูลของโปรแกรม End-Note ที่กำลังใช้งานอยู่ทันที

ข้อมูลที่ import มาจะแสดงแยกให้เห็น ข้อมูลหนังสือ หรือแม้แต่เว็บเพจ ก็จะถูกเก็บเข้าใน End-Note และนำไปใช้งานต่อได้


research.enl
Language in Society
Groups
All References (93)
Imported References (11)
Trash (0)
Custom Groups
paper1 (10)
Smart Groups
Online Search
Library of Congr... (0)
Author
Year
Title
Aston 2004 Corpora And Language Learners
Baayen 2008 Analyzing Linguistic Data: A Practical Introduction to Statistics using...
Cummings 2003 Thailand
Davidson 1987 Lâi Sɔ̄ɔ̄ Thai
Gilquin 2008 Linking up contrastive and learner corpus research
Lerot 2003 Corpus-based Approaches to Contrastive Linguistics and Translati...
Paolillo 2001 Analyzing Linguistic Variation: Statistical Models and Methods
Preview
Search
Davidson, Jeremy H. C. S.; Simmonds, E. H. S.; Studies, University of London School of Oriental and African (1987). Lâi Sɔ̄ɔ̄ Thai. Routledge.

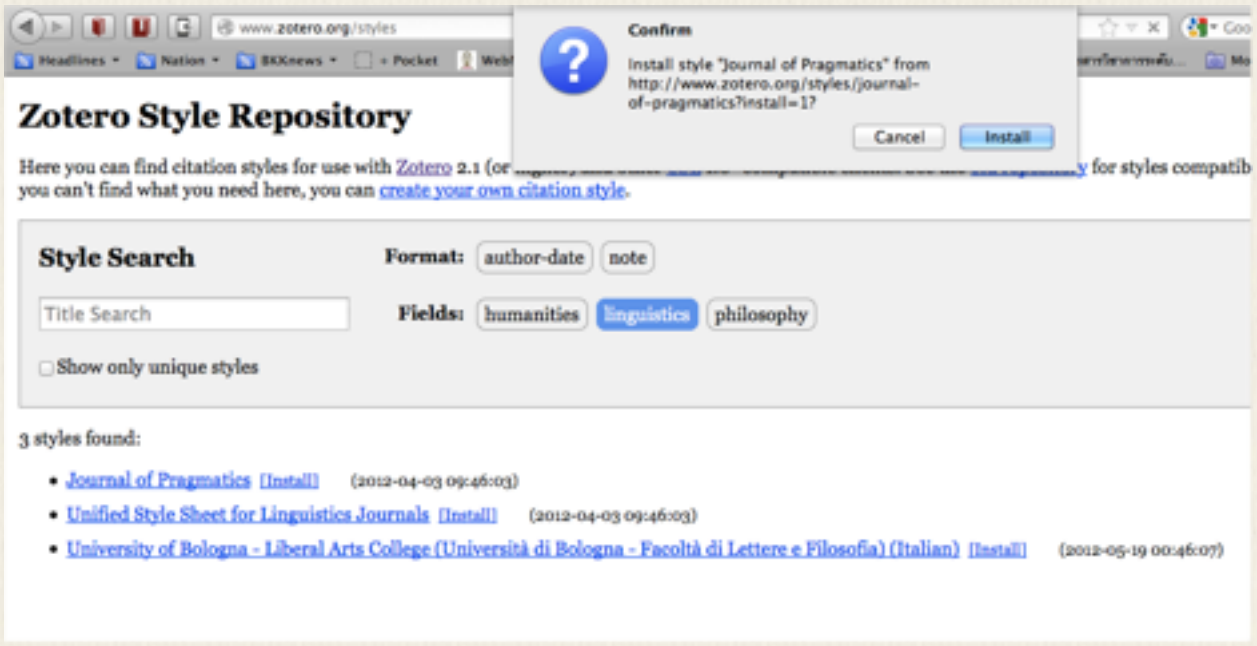
SECTION 5

# การใช้ Zotero ทำบรรณานุกรม

การใช้ ZOTERO ทำบรรณานุกรม

ความจริง ถ้าไม่มี EndNote จะใช้ Zotero เพื่อช่วยทำบรรณานุกรม และสร้างรายการตาม style ต่างๆ ก็ได้เช่นกัน สามารถติดตั้ง style ที่ไม่มีมาในตอนแรกได้โดยไปที่

http://www.zotero.org/styles แล้วเลือก install style ที่ต้องการใช้




โปรแกรม Zotero สามารถทำงานร่วมกับ MS Word ได้ สามารถดูวีดีโอสาธิตได้ที่

http://www.zotero.org/support/word_processor_integration แต่จะต้องติดตั้ง plug-in ของ

Word ก่อน ซึ่งสามารถ download ได้ที่

http://www.zotero.org/support/word_processor_plugin_installation

# การ backup และ sync Zotero

## การ BACKUP และ SYNC ZOTERO

ข้อมูลที่เก็บไว้ใน Zotero จะถูกเก็บโดย default ที่ต่อไปนี้

On a Mac:

/Users/<username>/Library/Application Support/Firefox/Profiles/<randomstring>/

zotero

On Windows 2000/XP:

C:\Documents and Settings\<username>\Application

Data\Mozilla\Firefox\Profiles\<randomstring>\zotero

On Windows Vista:

C:\Users\<User

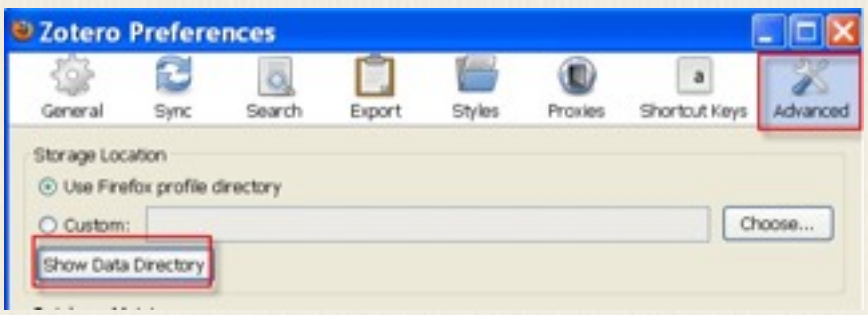
Name>\AppData\Roaming\Mozilla\Firefox\Profiles\<randomstring>\zotero

ถ้าต้องการ backup ข้อมูล Zotero ไว้ ให้ไปที่ folder ข้างบนแล้ว copy ทั้ง folder เก็บไว้

หากมีปัญหาภายหลังก็นำ folder ที่ copy ไว้นี้มาทับลงไปได้

สามารถเข้าไปที่ Preference ของ Zotero แล้วเรียกให้แสดง folder ที่เก็บข้อมูลหรือจะบอกให้นำ

ข้อมูลไปเก็บที่ใหม่ที่กำหนดก็ได้ โดยเลือกที่ Custom แล้วระบุ folder ที่ต้องการ



หากใช้ zotero ในหลายๆเครื่อง สามารถ sync ข้อมูลผ่านทาง zotero server ได้ โดยจะต้องมี account บน zotero ก่อน ให้เข้าไปที่ zotero preference เพื่อสร้าง account การ sync ข้อมูล สามารถกดไอค่อน sync ที่มุมขวาบนได้

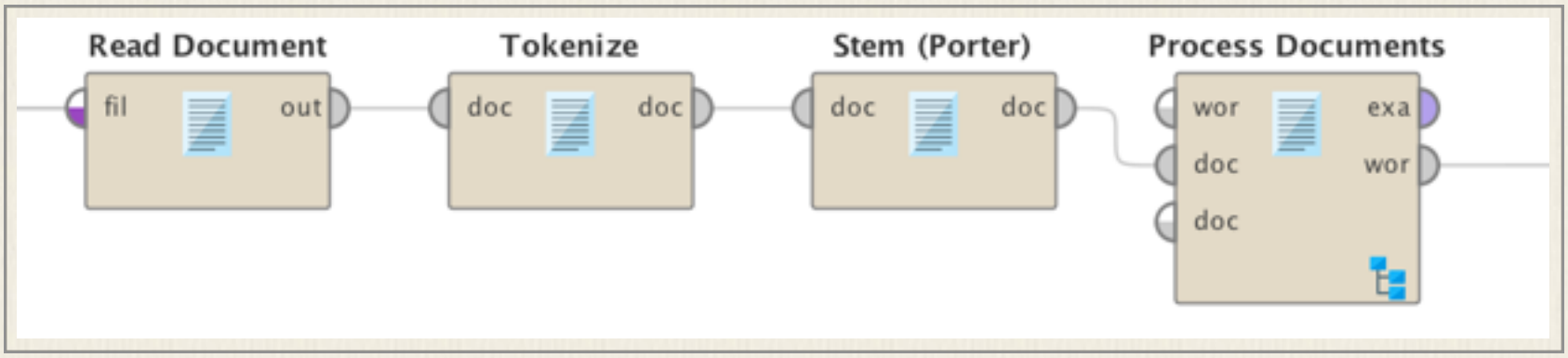
รายละเอียดการใช้งาน Zotero อ่านได้จาก http://www.zotero.org/support/

CHAPTER 3

# โปรแกรม RapidMiner

RapidMiner เป็นโปรแกรมช่วยทำ Text Mining โดยไม่ต้องอาศัยความรู้ในการเขียนโปรแกรม แต่ใช้โปรแกรมสำเร็จรูปและ plugin ต่างๆ ที่สร้างขึ้นมาสำหรับ RapidMiner ในการประมวลผลข้อความ RapidMiner มี license สำหรับการศึกษาซึ่งอนุญาตให้ใช้โปรแกรมเพื่อการศึกษาได้ สามารถโหลดโปรแกรม RapidMiner Studio ได้จาก https://rapidminer.com/

การใช้งานคือการนำ Operator มาประกอบกันเป็นลำดับขั้นตอนที่ควรเป็น และเชื่อมโยง out-put จาก operator หนึ่งไปเป็น input ของ operator ถัดไป ที่ต้องดูคือ ประเภทข้อมูลของ output ที่ออกมาจะต้องตรงกับประเภทของ input ใน operator ถัดไป เช่น doc, fil, wor etc




`แต่ละกล่องคือ operator หรือ process ที่ต้องการให้ดำเนินการกับข้อมูลที่ส่งเข้าไป และให้นำข้อมูลออกที่ได้ไปใช้ต่อในกระบวนการถัดไป Operator พื้นฐานจะมาพร้อมโปรแกรมแล้ว แต่หากต้องการติดตั้งเพิ่ม ให้ไปที่ Extensions - Market Place เพื่อค้นหา operator เพิ่มเติม ตัวที่ควรเพิ่มเติมเพื่อใช้ในการประมวลผลข้อความ คือ Text Processing, Text Analysis by AYLIEN, Web Mining

การเลือก Operator ต่าง ๆ สามารถทำได้โดยการพิมพ์คำค้นในแถบเครื่องมือ เมื่อพบ Operator ที่ต้องการแล้ว จึงลากเข้าไปที่หน้า Design และเชื่อมโยง input และ output ของแต่ละ operator ให้ถูกต้องตามลำดับที่ต้องการ

**การนำข้อมูลตัวบทเข้า**

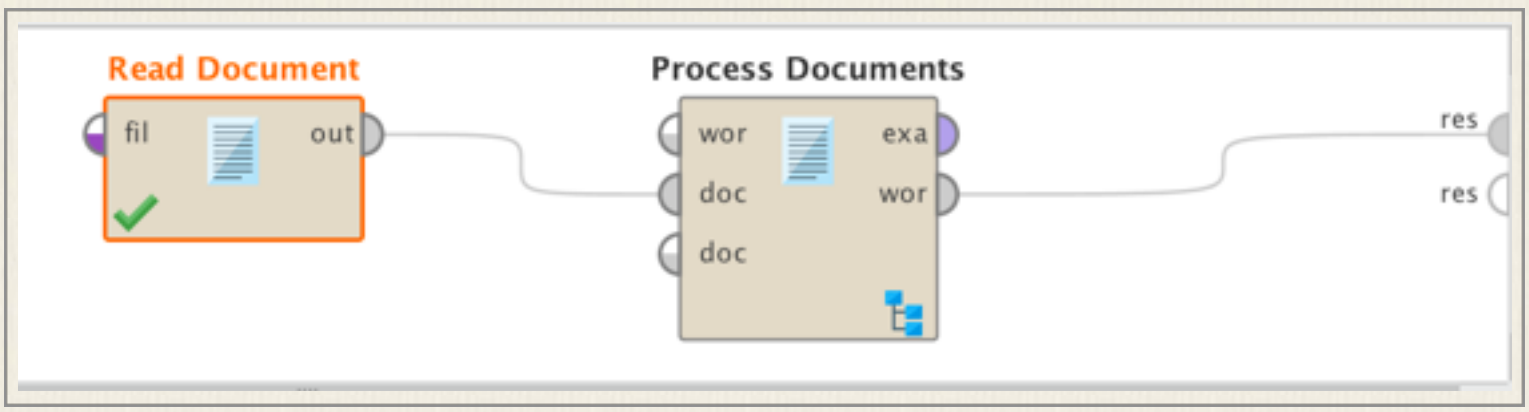
สามารถนำข้อมูลประเภทต่าง ๆ เข้าได้หลากหลายวิธี เช่น ใช้ “Create Document” แล้วค่อย copy-paste ตัวข้อความที่ต้องการเข้าไปใส่ หรือใช้ “Read Document” แล้วระบุชื่อไฟล์ที่ต้องการให้อ่านเข้ามา หรือถ้าต้องการอ่านไฟล์ทั้งหมดใน folder ให้ใช้ “Process Document from Files” แล้วระบุ folder และประเภทไฟล์ที่ต้องการให้อ่าน ตั้งชื่อ class สำหรับ folder นั้นๆ จากนั้นภายใน “Process Document from Files” จึงใส่กระบวนการย่อยต่าง ๆ ลงไป เริ่มตั้งแต่การทำ “Tokenize” เพื่อดึงคำออกมา การแปลงตัวอักษรเล็กใหญ่ “Transform Cases” การใช้ “Stem” แปลงเป็นรูปคำพื้นฐาน เป็นต้น

หากข้อมูลเข้าเป็นรูปแบบอื่น เช่น csv, excel, html, xml, ก็ให้ใช้ operator ที่ใช้สำหรับอ่านไฟล์เหล่านั้น เช่น “Read CSV”, “Read Excel”, “Read XML”, “Read URL”

**การสร้างรายการคำและความถี่**

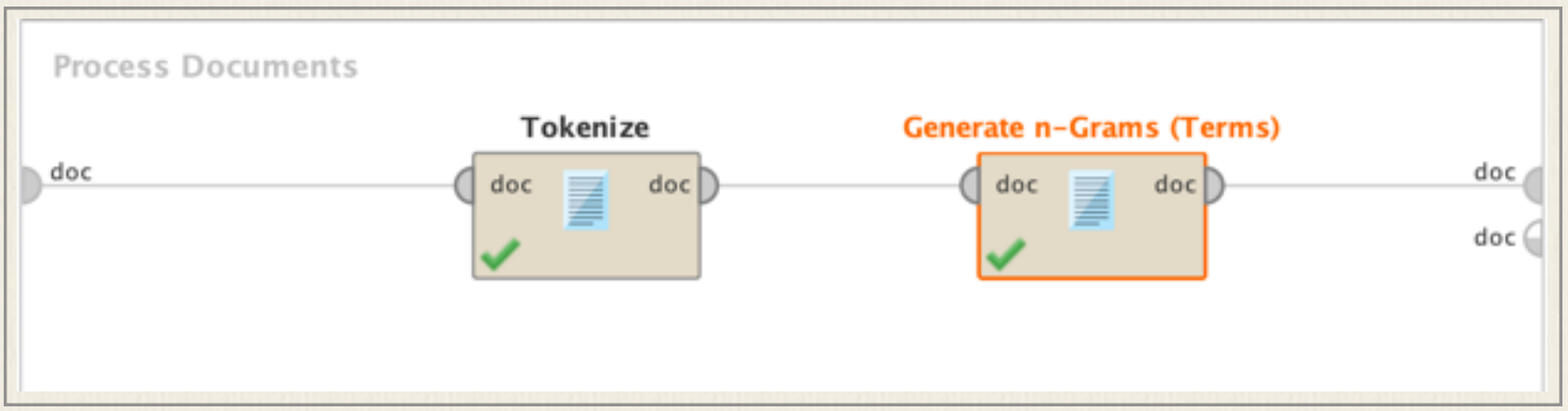
Operator หลักที่ใช้ คือ “Read Document” กับ “Process Document” ให้ค้นจากรายการเครื่องมือ แล้วลากกล่องนั้นมาที่หน้าจอออกแบบ ใน “Read Document” ให้เลือกไฟล์ที่ต้องการให้อ่านข้อมูล แล้วลาก output จาก “Read Document” ซึ่งเป็น doc ไปที่ input ของ “Process Document” ซึ่งเป็น doc แล้วเลือก output wor ซึ่งคือ wordlist ไปที่ res หรือ result ด้านขวาสุด ภายใน “Process Documents” เราจะใส่ operator “Tokenize” เพื่อแยกข้อความออกเป็น token ก่อน การสั่ง tokenize มีตัวเลือกตั้งแต่ tokenize ระดับตัวอักษร ระดับคำ ระดับประโยค หรือใช้ regular expression เป็นตัวแยกคำ เช่น ถ้าใช้ [\s\W]+ หมายถึงแยกคำโดยใช้ space

หรืออักขระอื่นที่ไม่ใช่ตัวอักษรหรือตัวเลข (เครื่องหมายต่าง ๆ) อย่างน้อยหนึ่งตัวเป็นตัวแยกคำ และอาจมี process อื่นๆ เช่น stem เพื่อแปลงคำเป็นรูปพื้นฐานเดียวกัน เมื่อสั่ง run process แล้ว หน้าจอแสดงผลจะได้ตารางรายการคำออกมา

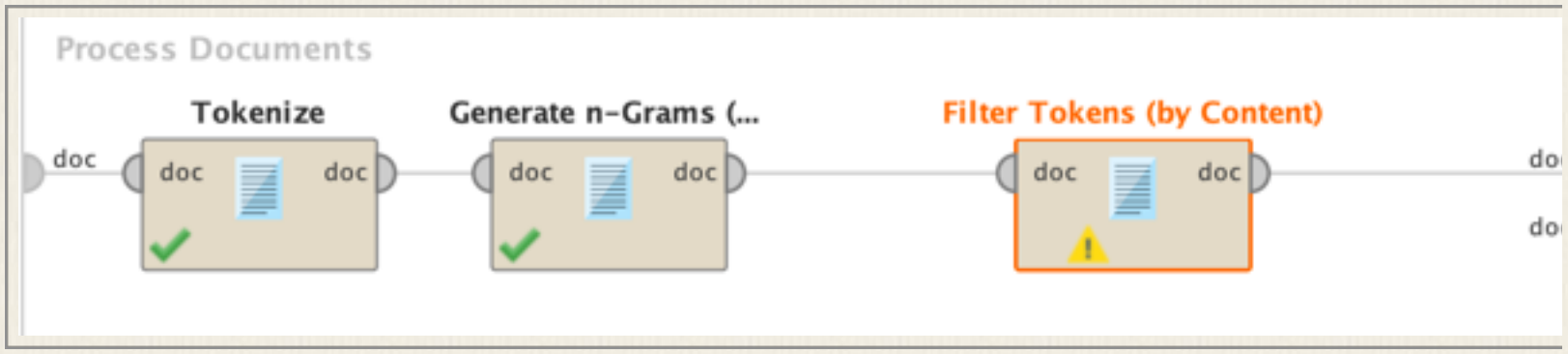



**การสร้างรายการ n-gram**

ทำเหมือนการสร้างรายการคำ คือมี “Read Documents” และ “Process Documents” ภายใน “Process Documents” หลังจาก “Tokenize” ให้ต่อด้วย “Generate n-Grams (Term)” โดยเลือกจำนวน n-gram ที่ต้องการ กระบวนการนี้จะสร้างข้อมูล n-gram ที่กำหนด รวมถึงข้อมูลย่อยของ n-gram ที่ต่ำลงไปจนถึงคำเดี่ยวด้วย เช่น ถ้าให้สร้าง 3-gram จะได้ข้อมูลรายการ 3-gram, 2-gram และคำเดี่ยว โดยมีเครื่องหมาย “_” ใช้เป็นตัวเชื่อมระหว่างคำใน n-gram

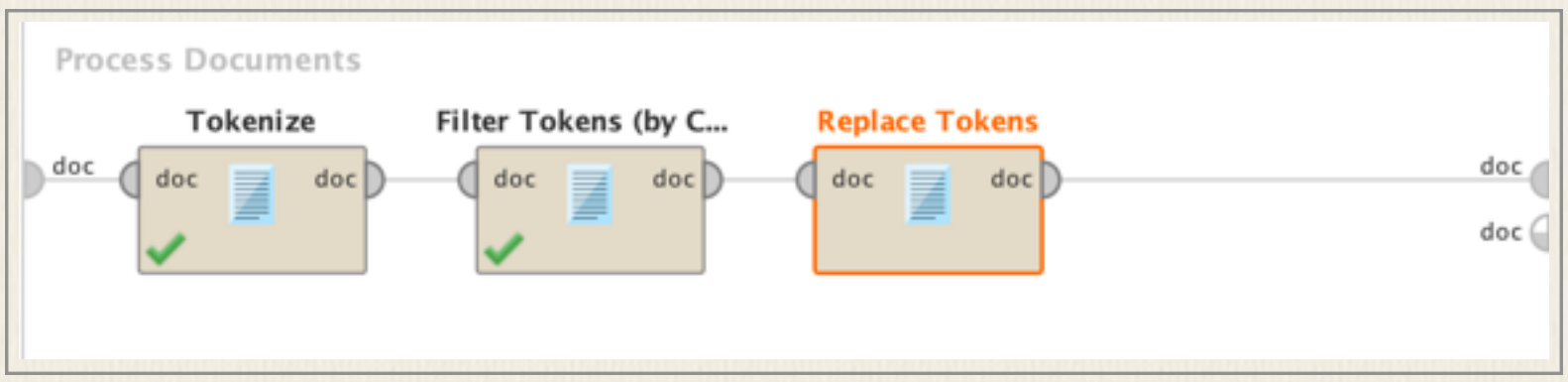



หากต้องการสกัดเฉพาะ n-gram ที่ต้องการ ก็จะต้องกำหนดตัวกรองหรือ filter เอง เช่น ถ้าต้องการ 3-gram ให้เติม operator “Filter Tokens (by Content)” ภายใน process นี้ให้กำหนดว่าดึงเฉพาะรายการ token ที่ “contain match” ตรงตามค้นแบบ “_.+_” ซึ่งเป็น regular expression ที่จะ match กับรายการที่มีเครื่องหมาย _ สองครั้ง




### การทำ simple concordance

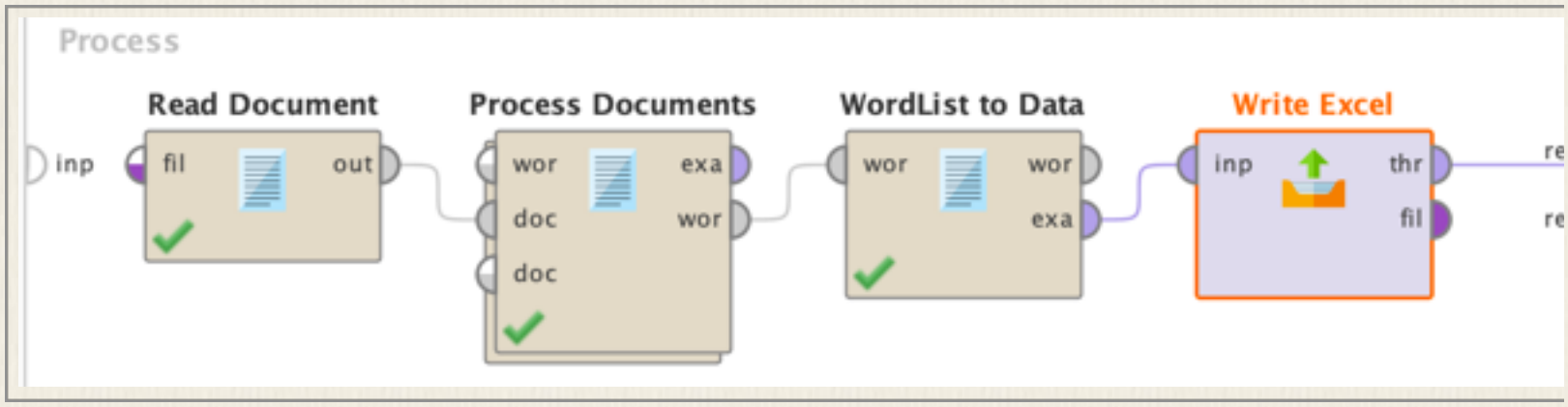
หากต้องการดึงตัวอย่างประโยคที่มีคำที่ต้องการจาก text สามาถทำได้โดยการใช้ ."Tokenize" โดยเลือก token แบบ "linguistic sentences" (ภาษาอังกฤษมีตัวเลือกนี้ให้ใช้) แล้วจึง "Filter Token (by Content)" โดยระบุคำที่ต้องการค้นเพื่อให้เหลือเฉพาะตัวอย่างประโยคที่มีคำนั้น หากต้องการนำผลที่ได้ไปเปิดใน Excel เพื่อใช้งานต่อ ก็อาจทำต่อโดยใช้ "Replace Tokens" และแทนที่คำค้นด้วยคำเดิมแต่มีเครื่องหมายเฉพาะ เช่น | อยู่หน้าหลังเพื่อให้ไปเปิดใน Excel เป็นสามคอลัมน์ได้ เช่น ถ้าต้องการค้นคำว่า digital ที่อักษรแรกเป็นตัวพิมพ์ใหญ่หรือเล็กก้ได้ ก็ใส่ replace what ด้วย "\s([d|D]igital)\s" และใส่ replace by ด้วย " \|$1\| "




### การเก็บข้อมูลลงไฟล์

หลังจากสร้างรายการคำหรือรายการ n-gram แล้ว หากต้องการเก็บผลที่ได้เป็นไฟล์ สามารถทำได้โดยการเรียก operator ต่อ คือ แปลงรายการคำเป็นข้อมูล ("WordList to Data") จกานั้นจึง

เรียก operator ที่เก็บข้อมูลลงไฟล์ในรูปแบบที่ต้องการ เช่น Excel, CSV (“Write Excel”, “Write CSV”) และระบุชื่อไฟล์ที่ต้องการให้ save ผลลัพธ์นั้น

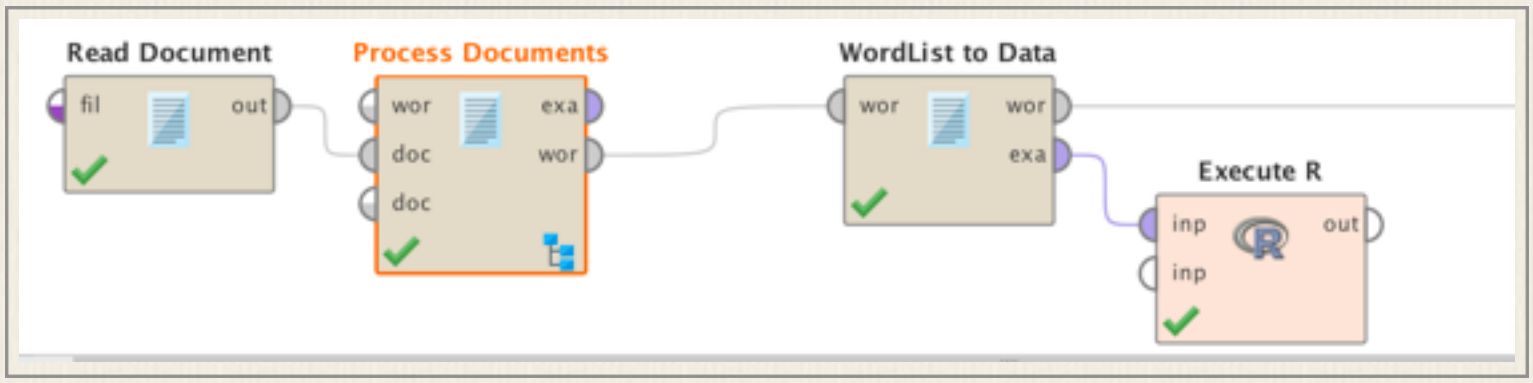



**การติดตั้ง R และ Python extension**

ใน RapidMiner เราสามารถติดตั้ง extension สำหรับใช้ run R หรือ Python scripts ได้ โดยค้นหา extension ชื่อ “R Scripting” และ “Python Scripting” หลังติดตั้งแล้ว จะได้ operator “Execute R” และ “Execute Python” สำหรับให้เขียนโปรแกรมเพิ่มเติมเองด้วยภาษา R หรือ Python ได้ โดยก่อนจะใช้งานได้ เราต้องไปบอกไว้ใน Preferences ของโปรแกรมก่อนว่าโปรแกรม R และ Python นั้นติดตั้งไว้ที่ไหนในเครื่อง

**การสร้าง Word Cloud ด้วย R scripts**

R มี library “wordcloud” ที่ช่วยสร้างภาพคำตามขนาดความถี่ที่พบในข้อมูล ตัวอย่างข้างล่างแสดงการเพิ่ม operator “Execute R” เข้ามาหลังจากสร้างรายการคำและความถี่แล้ว โดย “Read Document” จะอ่านข้อมูลจากไฟล์ที่กำหนด ภายใน “Process Document” มีกระบวนการย่อย คือ “Tokenize”,“Transform Cases” และ “Filter Stopwords (English)” ผลที่ได้เป็น word list (wor) ที่มีเฉพาะ content words ซึ่งต้องนำมาแปลงเป็น Data ก่อน แล้วส่งข้อมูลที่พบ (exa) ไปยัง operator “Execute R”



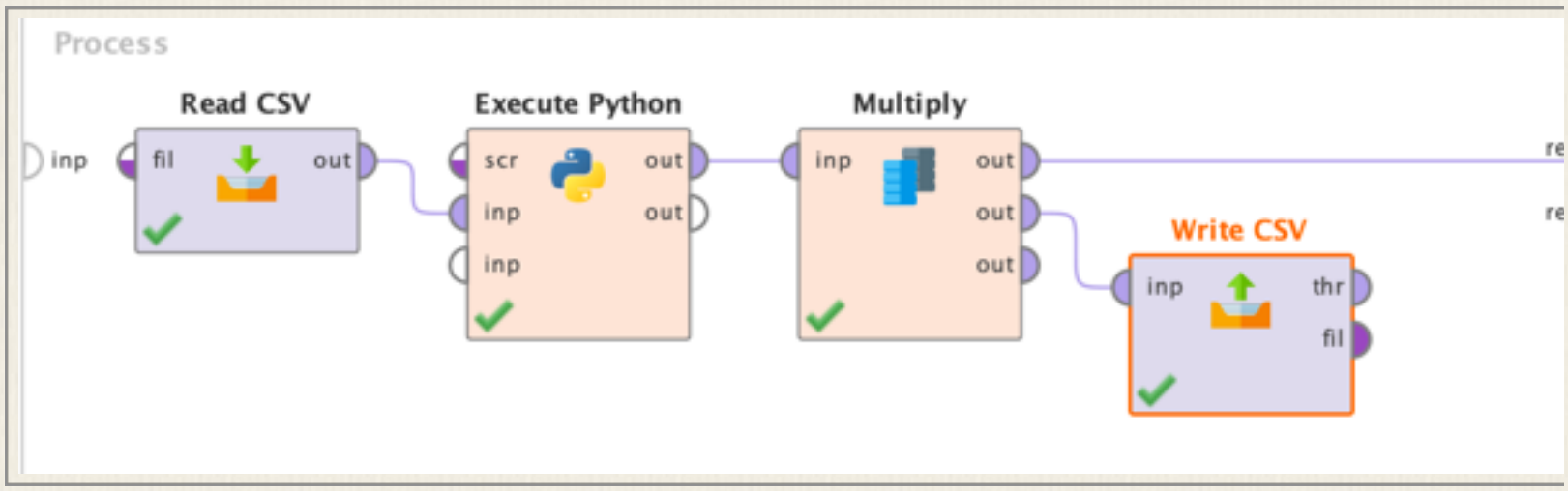

ภายในกล่องนั้น จึงเขียนโปรแกรม R ดังตัวอย่างนี้

```
library(wordcloud)
library(RColorBrewer)

png(filename="/Users/macbook/Downloads/test.png", width=800, height=600)

wordcloud::wordcloud(data$word, data$total, max.word=100, scale=c(10,2),
colors=brewer.pal(8, "Dark2"))

dev.off()
```

คำสั่ง png เป็นการกำหนดให้ภาพที่สร้างขึ้นถูกเก็บเป็นไฟล์ตามที่ตั้งชื่อ เนื่องจาก RapidMiner ไม่สามารถแสดงผลออกาหน้าจอได้ จึงต้องให้เก็บเป็นไฟล์ png ส่วนคำสั่ง dev.off บอกให้จบการนำ output ลงในไฟล์ ส่วนคำสั่ง wordcloud เป็นคำสั่งสำหรับสร้างภาพคำตามขนาดความถี่ที่พบ รายละเอียดการใช้งานจริง ให้ค้นคว้าต่อในเรื่องการเขียนโปรแกรมด้วยภาษา R ภาพที่ได้จากการใช้คำสั่งนี้จะเก็บเป็นไฟล์ test.png และมีรายละเอียด ดังนี้

### การใช้ Python module

RapidMiner มี operator Execute Python ที่ช่วยให้เราสามารถเขียนโปรแกรมภาษา Python ประกอบการใช้งาน RapidMiner ได้ ตัวอย่างข้างล่างเป็นการใช้โมดูลเพื่อทำ pos tagging ข้อมูลภาษาไทย โดยข้อมูลภาษาไทยเตรียมไว้เป็นไฟล์ csv มีหนึ่งคอลัมน์ชื่อ “content” แต่ละแถวเป็นข้อความที่ต้องการตัดคำและใส่ POS




ภายในกล่อง Execute Python เราต้องใส่ code ที่ต้องการใช้ ในที่นี้จะใช้ TLTK python module การใช้ Python ใน RapidMiner ข้อมูลนำเข้าจะมาเป็น data frame ที่ชื่อ data และอยู่ใน rm_main เสมอ การเขียนโปรแกรมจึงเป็นการเขียน code เติมไปเพื่อ process ข้อมูลใน data-frame “data” นี้ ซึ่งในที่นี้มีคอลัมน์เดียวชื่อ “content” ส่วน data2 เป็น dataframe ที่สร้างขึ้นโดยนำข้อมูลจาก data มาใช้และเติมคอลัมน์ “postag”สำหรับเรียมเก็บข้อมูลที่ได้จากการทำ pos tagging ในตัวอย่างจะเห็นการเรียนใช้ loop ดึงข้อมูลจาก data2 มาทีละแถวและเรียกใช้ tltk.nlp.pos_tag กับข้อความที่อยู่ใน “content” เมื่อได้ output จาก pos_tag ที่เป็น list ของ (word, pos) จึงนำมาต่อกันเป็น string เก็บไว้ใน “out” จากนั้นจึงนำผลลัพธ์ที่ได้นี้ไปเติมในคอลัมน์ “postag” ที่เตรียมไว้

ผลลัพธ์ที่ได้สามารถ save ออกมาเป็น csv ไฟล์ หรือจะนำไปใช้เพื่อประมวลผลต่อใน RapidMiner ก็ได้

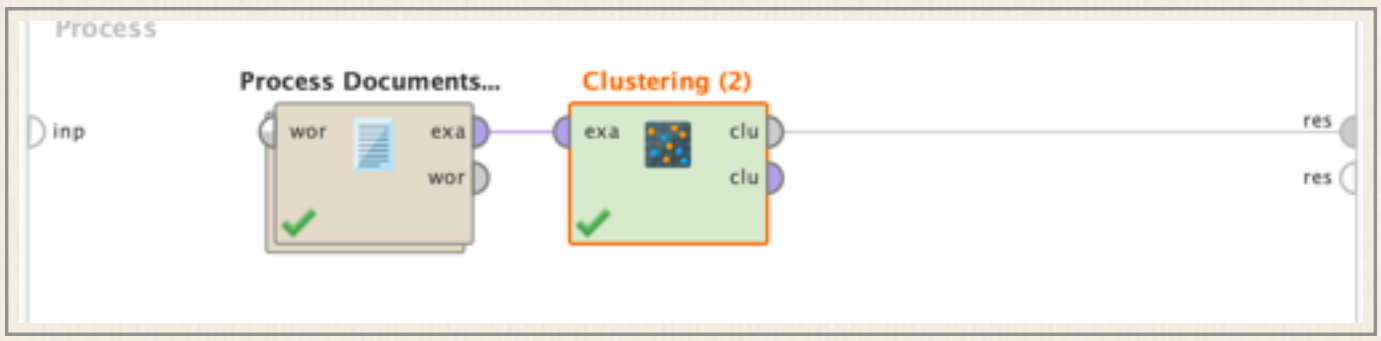
```python
import pandas as pd
import tltk

# rm_main is a mandatory function,
# the number of arguments has to be the number of input ports (can be none)
def rm_main(data):

    data2 = pd.DataFrame(data)
    data2['postag'] = ""

    for idx, row in data.iterrows():
        intxt = row['content']
        intxt = intxt.replace('"','')
#        out = tltk.nlp.word_segment(intxt)
        lstx = tltk.nlp.pos_tag(intxt)
        out = ''
        for lst in lstx:
            for (w, pos) in lst:
                out += w+'/'+pos+'|'

        data2.loc[idx].postag = out
    return data2

```

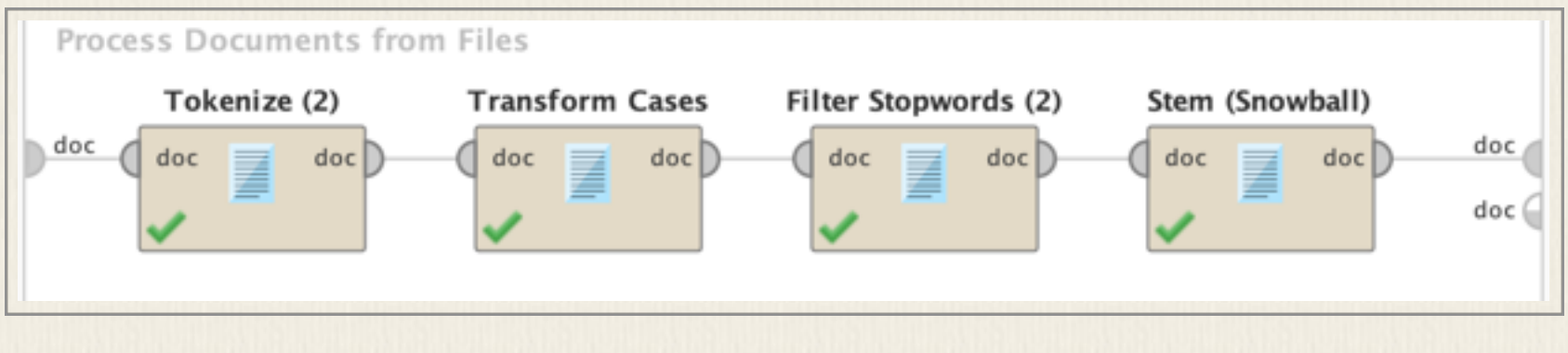
```
"content","postag"
"ลดน้ำหนักวิธีใหม่","ลด/VERB|น้ำหนัก/NOUN|วิธี/NOUN|ใหม่/ADJ|<s/>/PUNCT|"
"คนจำนวนมากต้องการลดน้ำหนักโดยไม่เข้าใจศาสตร์ในเรื่องนี้อย่างลึกซึ้ง",
"คน/NOUN|จำนวน/NOUN|มาก/ADJ|ต้องการ/VERB|ลด/VERB|น้ำหนัก/NOUN|โดย/SCONJ
ไม่/PART|เข้าใจ/VERB|ศาสตร์/NOUN|ใน/ADP|เรื่อง/NOUN|นี้/DET|อย่าง/PART|
ลึกซึ้ง/ADV|<s/>/PUNCT|"
"จึงทำให้ไม่ได้ผล และแถมต้องปวดใจไม่ได้บริโภคสิ่งที่ตนเองปรารถนาอีกด้วย",
"จึง/SCONJ|ทำให้/VERB|ไม่ได้/VERB|ผล/NOUN|<s/>/PUNCT|และ/CCONJ|แถม/CCONJ
ต้อง/AUX|ปวดใจ/VERB|ไม่ได้/AUX|บริโภค/VERB|สิ่ง/NOUN|ที่/SCONJ|ตนเอง/NOUN|
ปรารถนา/VERB|อีก/ADV|ด้วย/ADV|<s/>/PUNCT|"
```

**การจัดกลุ่มข้อมูล text (clustering)**

RapidMiner มี operator จำนวนหนึ่งที่ช่วยในการทำ clustering หรือจัดกลุ่มข้อมูลที่มีความคล้ายคลึงกัน ตัวอย่างข้างล่างเป็นการใช้ operator “K-mean” สำหรับทำ Clustering” ซึ่งเป็นวิธีการจัดกลุ่มที่นิยมใช้กันวิธีหนึ่ง โดยเราต้องกำหนดจำหนวนกลุ่มที่ต้องการ และกำหนดจำนวนรอบที่เครื่องทดลองการขัดกลุ่ม ผลที่ได้จะเป็นกลุ่มตามจำนวนที่กำหนดให้ โดยมีเอกสารที่คล้ายกันอยู่ในกลุ่มนั้น ๆ และมีตารางแสดงรายการคำที่บอกถึงน้ำหนักความเกี่ยวข้องกับแต่ละ

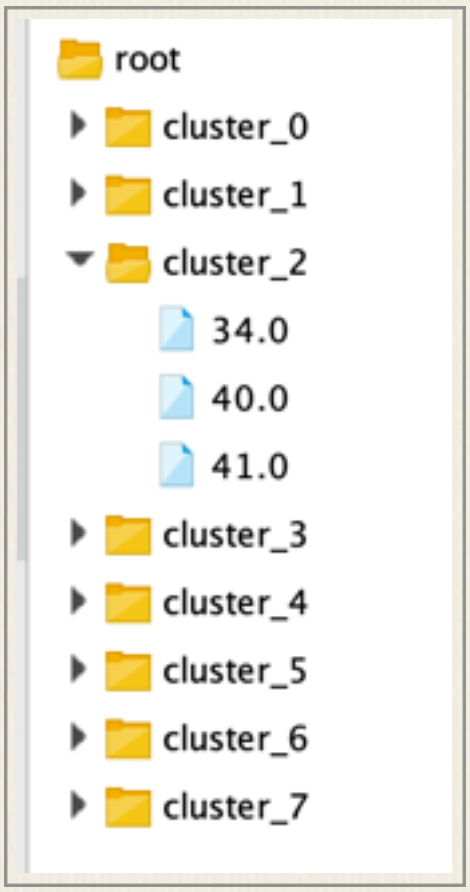
cluster ตัวอย่างที่เห็นเป็นการ process documents ซึ่งแต่ละ document เป็นคำปราศรัยการรับตำแหน่งของประธานาธิบดีแต่ละคน กำหนดให้จัดกลุ่มเป็น 8 กลุ่ม คำปราศรัยแต่ละอันจะถูกจัดเข้ากลุ่มตามความคล้ายคลึงของเนื้อหา (ภายใน “Process Documents” นอกจากการ “Tokenize” คำ แปลงเป็นตัวพิมพ์เล็ก “Transform Cases” ตัดคำไวยากรณ์ออก “Filter Stop-words (English)” และแปลงเป็นรากคำ “Stem (Snowball”)



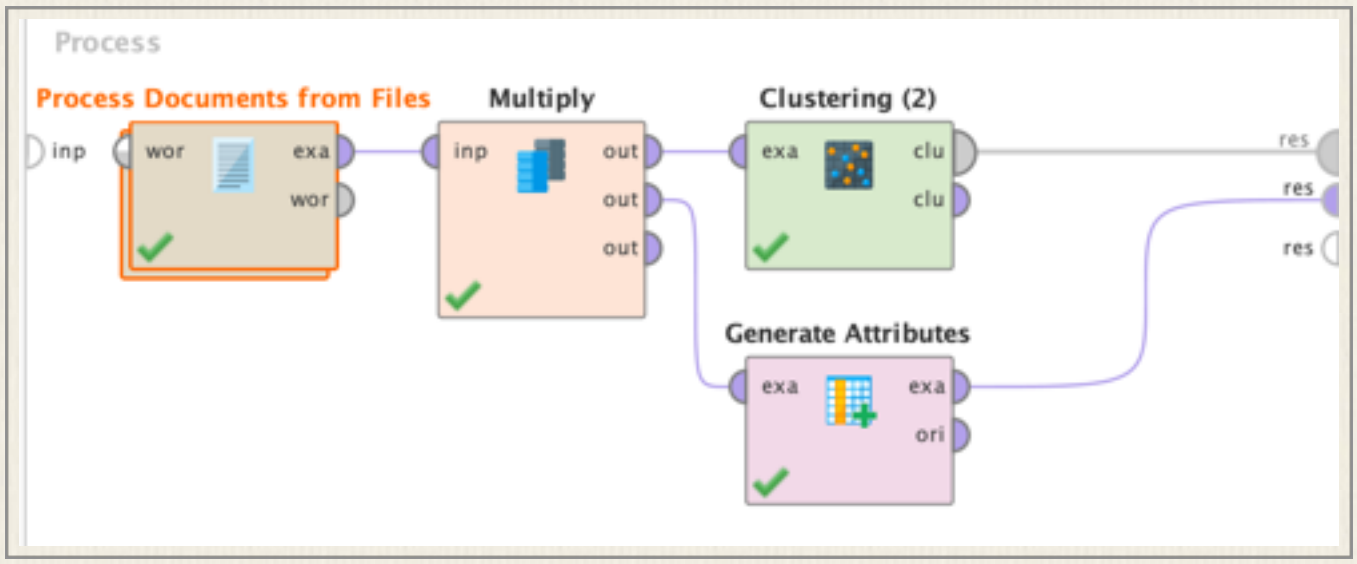

## Cluster Model

```
Cluster 0: 3 items
Cluster 1: 6 items
Cluster 2: 3 items
Cluster 3: 17 items
Cluster 4: 8 items
Cluster 5: 4 items
Cluster 6: 18 items
Cluster 7: 1 items
Total number of items: 60
```



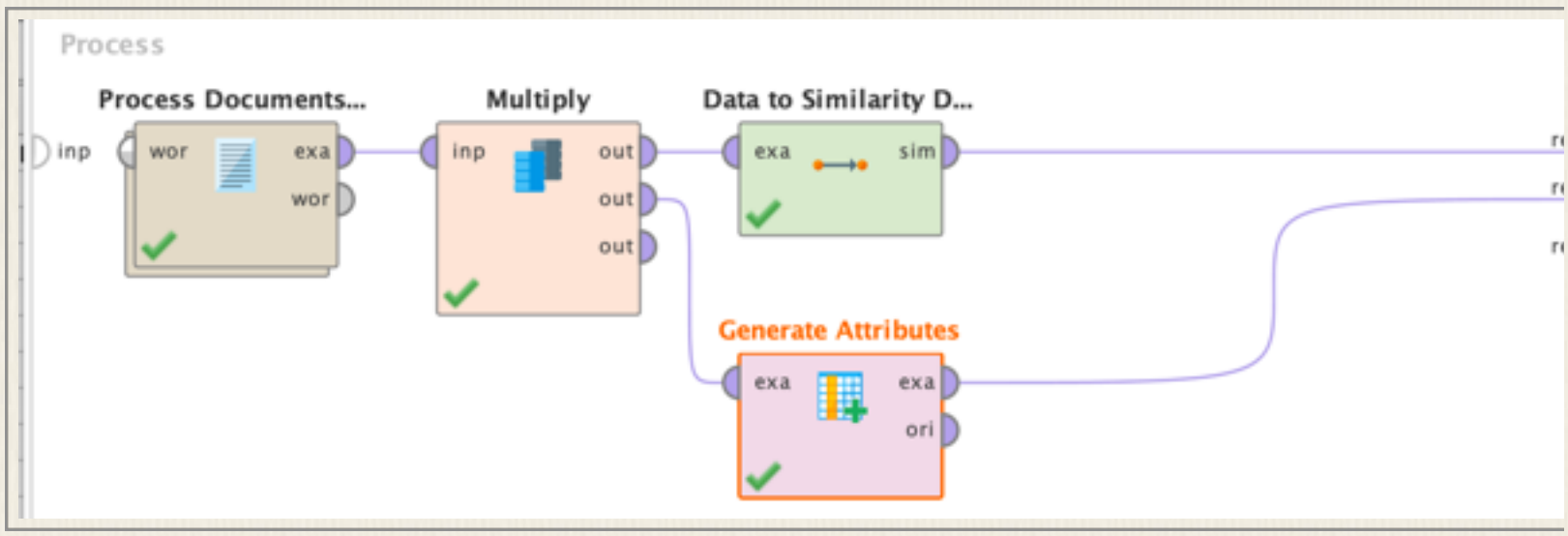
| Attribute | cluster_0 | cluster_1 | cluster_2 | cluster_3 | cluster_4 | cluster_5 | cluster_6 | cluster_7 |
|---|---|---|---|---|---|---|---|---|
| abandon | 0.010 | 0.015 | 0.008 | 0.005 | 0.007 | 0.015 | 0.007 | 0 |
| abat | 0 | 0 | 0 | 0.004 | 0 | 0 | 0 | 0 |
| abdic | 0 | 0 | 0 | 0 | 0 | 0.017 | 0 | 0 |
| abey | 0 | 0 | 0 | 0.006 | 0 | 0 | 0 | 0 |
| abhor | 0 | 0 | 0 | 0.004 | 0 | 0 | 0 | 0 |
| abid | 0.011 | 0.011 | 0.009 | 0.011 | 0 | 0.009 | 0.003 | 0 |
| abil | 0 | 0.003 | 0.006 | 0.011 | 0 | 0.021 | 0.014 | 0 |
| abject | 0 | 0 | 0 | 0.004 | 0 | 0 | 0 | 0 |
| abl | 0 | 0.011 | 0.009 | 0.011 | 0.004 | 0 | 0.012 | 0 |
| abli | 0 | 0 | 0 | 0 | 0 | 0 | 0.003 | 0 |
| abnorm | 0 | 0 | 0 | 0.005 | 0 | 0 | 0 | 0 |
| abod | 0 | 0 | 0 | 0 | 0 | 0 | 0.003 | 0 |
| abolish | 0.019 | 0 | 0 | 0.001 | 0.005 | 0 | 0.003 | 0 |

ในกรณีที่ถ้าต้องการรู้ว่า หมายเลขไฟล์แต่ละอันคืออะไร  อาจทำได้โดยการดึง attribute File-name ออกมาด้วยโดยใช้ “Generate Attributes”  และกำหนดให้ดึง  “metadata_file” ออกมาด้วยเป็นผลอีกตารางหนึ่ง  และเนื่องจากต้องนำข้อมูลจาก “Process Documents from files” ไปใช้สองที่  จึงเติม “Multiply” เพื่อสร้างสองเส้นทางออกมา



**การหาค่าความคล้ายกันของข้อความ (text similarity)**

RapidMiner มี operator ที่ช่วยให้เราเปรียบเทียบข้อความและดูความคล้ายกันของเอกสาร คือ “Data to Similarity Data” ตัวอย่างข้างล่างแสดงการอ่านข้อมูลคำปราศรัยใน Inauguration corpus โดยใช้ “Process Documents from Files” ภายในมีการทำ “Tokennize”, “Transform cases”, “Filter Stopwords (English)”, และ “Stem (Snowball)” จากนั้นนำผลของแต่ละ speech เปรียบเทียบความคล้ายกันด้วยวิธีการวัดแบบต่าง ๆ เช่น “CosineSimilarity”

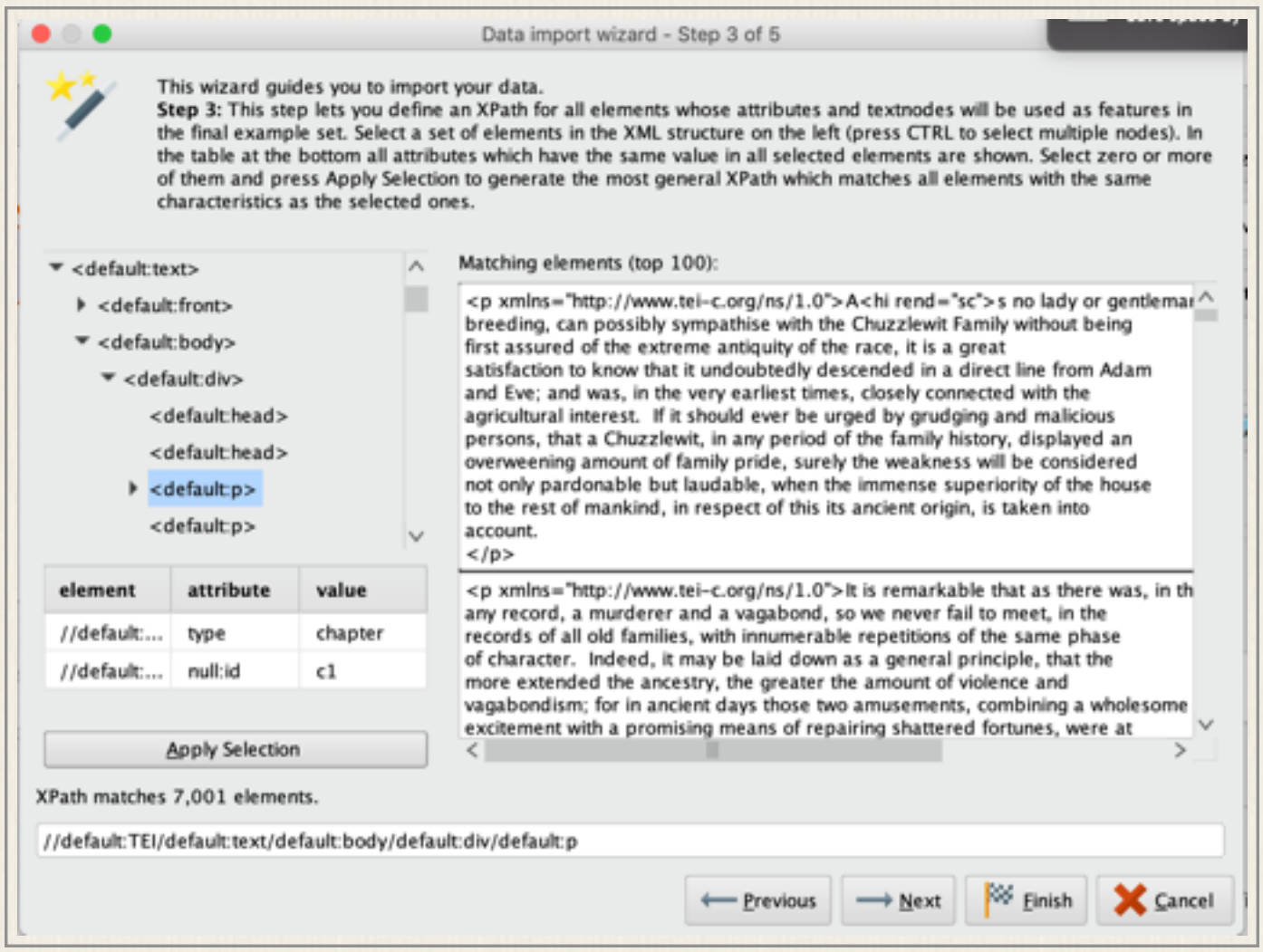


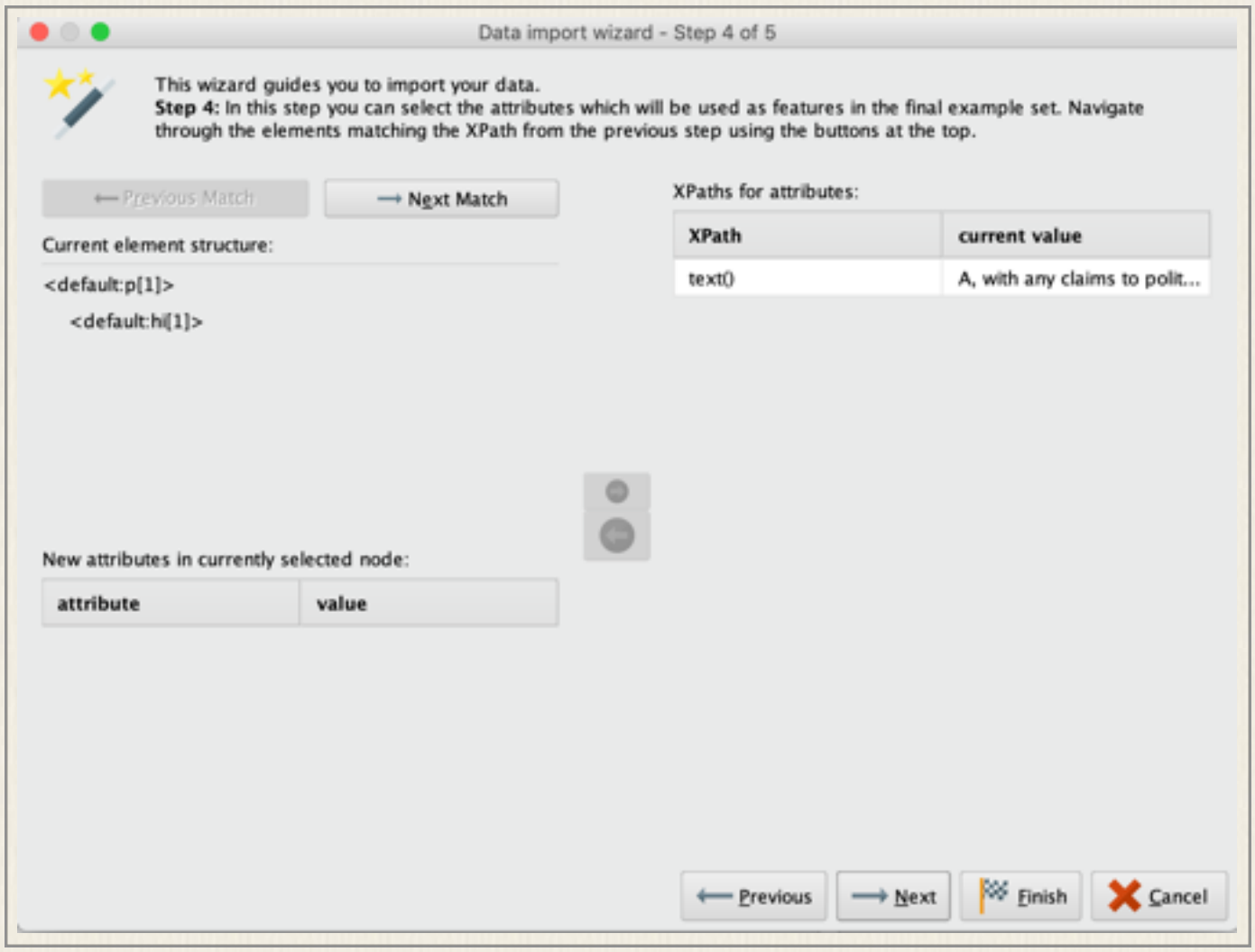

ExampleSet (3364 examples, 0 special attributes, 3 regular attrib... Fi

| Row No. | FIRST_ID | SECOND_ID | SIMILARITY |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 1 | 2 | 0.052 |
| 3 | 1 | 3 | 0.145 |
| 4 | 1 | 4 | 0.117 |
| 5 | 1 | 5 | 0.111 |
| 6 | 1 | 6 | 0.141 |
| 7 | 1 | 7 | 0.081 |
| 8 | 1 | 8 | 0.161 |
| 9 | 1 | 9 | 0.137 |
| 10 | 1 | 10 | 0.149 |

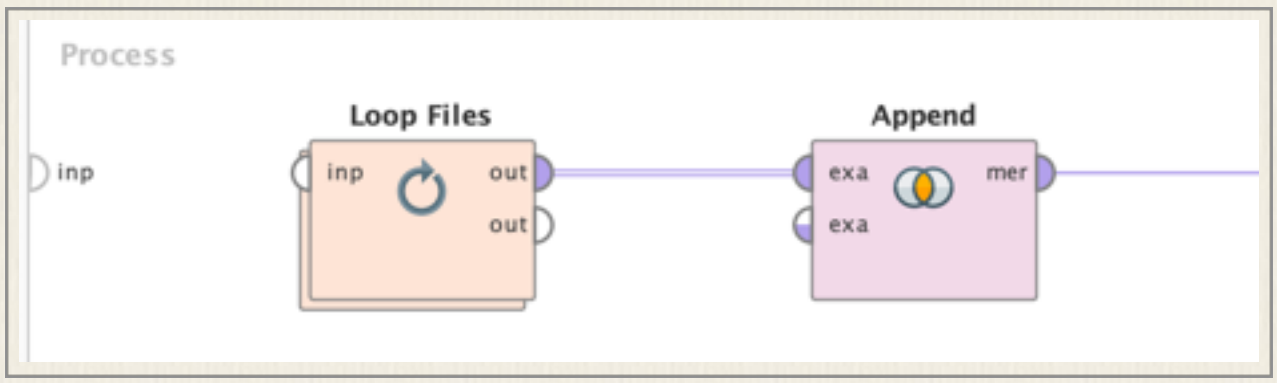
## การอ่านข้อมูล XML

ให้ใช้ operator “Read XML” และระบุไฟล์ xml ที่ค้องการอ่าน จากนั้นจะต้องบอกว่าให้ดูแท็กชื่ออะไรในการสกัดข้อมูลออกมาโดยการระบุ xpath วิธีที่สะดวกสุดคือใช้ Import Configuration Wizard ในการค้นหาแท็ก จนกระทั่งเห็นข้อมูล text ที่ต้องการปรากฏ ตัวอย่างข้างล่างเป็นไฟล์ xml ที่ได้มาจาก Oxford Text Archive ซึ่งข้อมูลที่ต้องการคือตัวข้อความในหนังสือ อยู่ภายใต้แท็ก <p> และใช้ function text() เพื่อดึงตัวข้อความออกมา รูปข้างล่างแสดงการตั้งค่าและตัวอย่างข้อมูลที่สกัดออกมาได้จากการใช้ Read XML






หากต้องการอ่านหลาย ๆ ไฟล์ที่มีโครงสร้างแบบเดียวกันนี้ ก็ทำได้โดยการใช้ operator “Loop Files” ตามด้วย “Append” เพื่อนำผมที่ได้แต่ละไฟล์ไปต่อกันเป็นข้อมูลตารางเดียว แล้วให้ “Read XML” อยู่ภายใน “Loop Files”

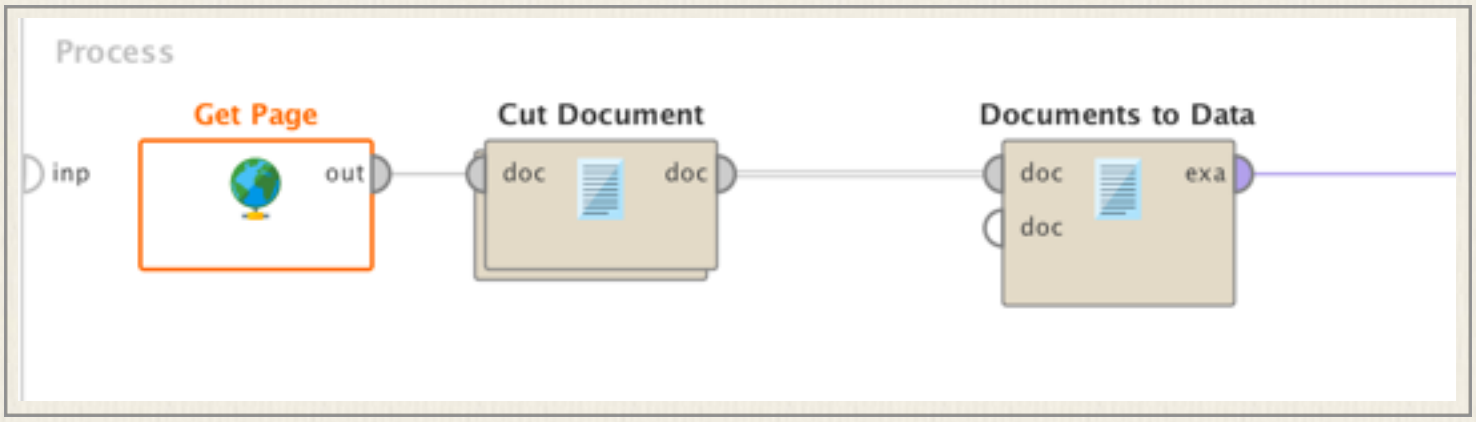


ExampleSet (16954 examples, 0 special attributes, 1 regular att... Filter (16,954 / 16,954 examples): al

| Row No. | text() |
|---|---|
| 1 | S to the of this direful Scourge of Mankind call'd the I find the moſt learn... |
| 2 | I muſt here indeed freely own that I am much pleaſed with the and of the learned ... |
| 3 | But it may not be improper to add ſomething under this from that great and (while... |
| 4 | What the Learned Dr. advances, about the Beginning of 4. of his Diſcourſe, in theſe... |
| 5 | In 7. the briefly mentions that moſt terrible which carried over the greateſt Part of... |
| 6 | Beſides what I have quoted from Mr. I ſhall entertain the Reader with a more Circu... |
| 7 | If that Part of the Story be Fact concerning the Worms, Snakes, which ar... |
| 8 | This moſt fearful is alſo taken Notice of by many other relates that at it carried off ... |
| 9 | Many other deſtructive before and after this are mentioned by Hiſtorians, and ſeveral |
| 10 | In 11. towards the End, the juſt mentions the Opinion of ſome who have thought th... |
| 11 | This Notion ſeems to be much a-kin to that which was advanced by (that famous i... |
| 12 | What the Learned ſuggeſts in Page 22, 23, againſt the old but unhappy Cuſtom of ... |

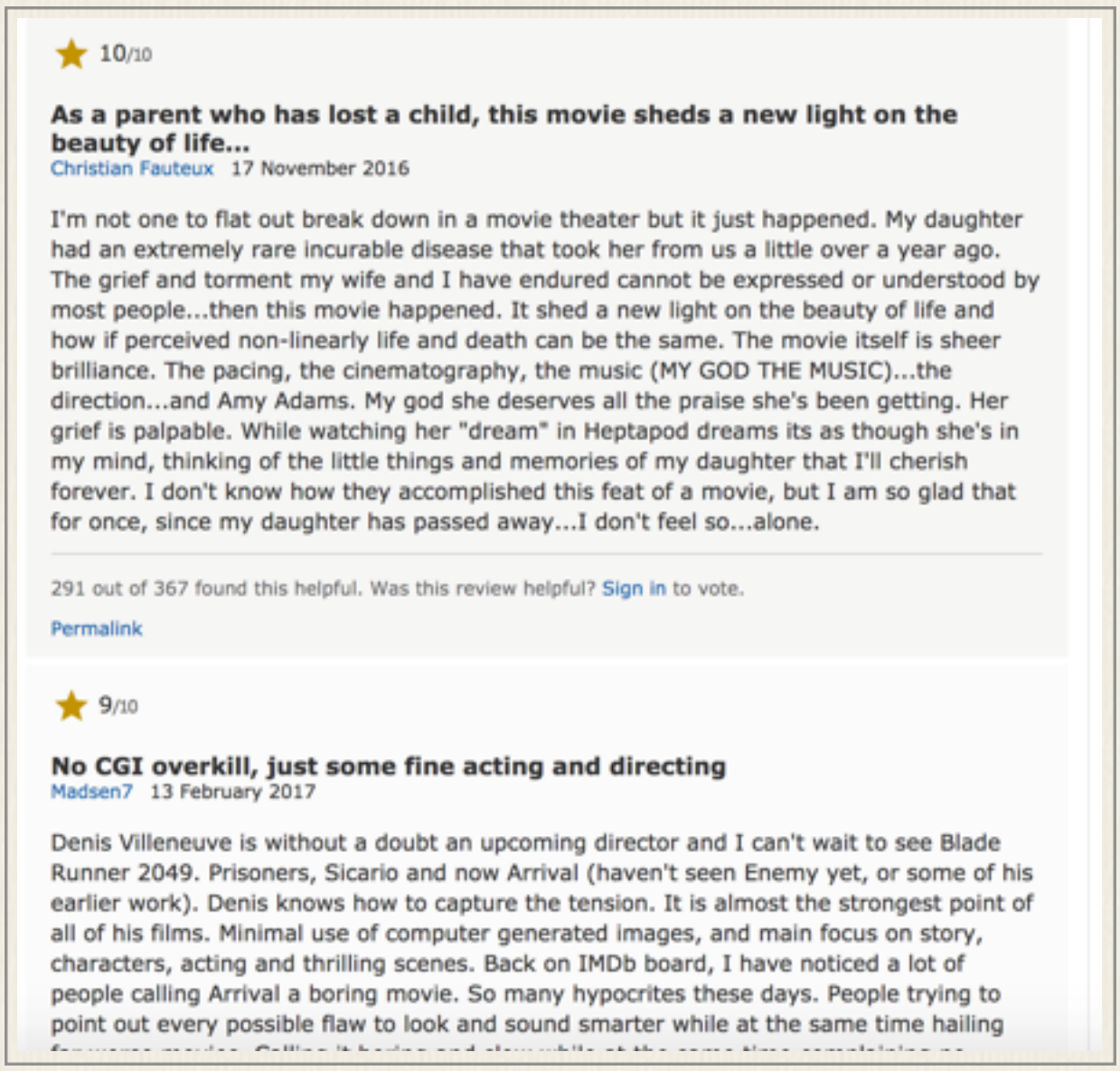
## การสกัดข้อมูลจากเว็บ

Operator “Crawl Web” ใช้ในการดึงข้อมูลจากเว็บ โดยระบุ url และระดับความลึกของการกดเข้าไปดูข้อมูล ไฟล์ที่ได้จากการ crawl สามารถเก็บใน folder ที่กำหนดใน “output dir” ได้ สามารถเลือกเก็บข้อมูลเป็น html เพื่อคงไฟล์เดิมไว้หรือให้เก็บมาเป็น text ไฟลีเลยก็ได้ ในการ crawl ข้อมูลเราสามารถกำหนดกฎได้ เช่น ให้ตรวจสอบ url ว่าควรมีคำใด เช่น .+review.+ คือมีคำ review อยู่ใน url หรือตรวจเนื้อความว่าควรมีคำหรือข้อความใด เป็นต้น เมื่อเก็บข้อมูลลงในเครื่องแล้วจึงมา process ภายหลังได้โดยใช้ “Process Documents from Files” ได้ แล้วกำหนดให้ใช้ “Extract Content” สำหรับข้อมูล html ออกมา




เริ่มด้วยการใช้ Get Page ระบุ url ที่ต้องการไปสกัดข้อมูล เช่น

https://www.imdb.com/title/tt2543164/reviews?spoiler=hide&sort=helpfulnessScore&dir=desc&ratingFilter=0



10/10

**As a parent who has lost a child, this movie sheds a new light on the beauty of life...**

Christian Fauteux 17 November 2016

I'm not one to flat out break down in a movie theater but it just happened. My daughter had an extremely rare incurable disease that took her from us a little over a year ago. The grief and torment my wife and I have endured cannot be expressed or understood by most people...then this movie happened. It shed a new light on the beauty of life and how if perceived non-linearly life and death can be the same. The movie itself is sheer brilliance. The pacing, the cinematography, the music (MY GOD THE MUSIC)...the direction...and Amy Adams. My god she deserves all the praise she's been getting. Her grief is palpable. While watching her "dream" in Heptapod dreams its as though she's in my mind, thinking of the little things and memories of my daughter that I'll cherish forever. I don't know how they accomplished this feat of a movie, but I am so glad that for once, since my daughter has passed away...I don't feel so...alone.

291 out of 367 found this helpful. Was this review helpful? Sign in to vote.

Permalink

9/10

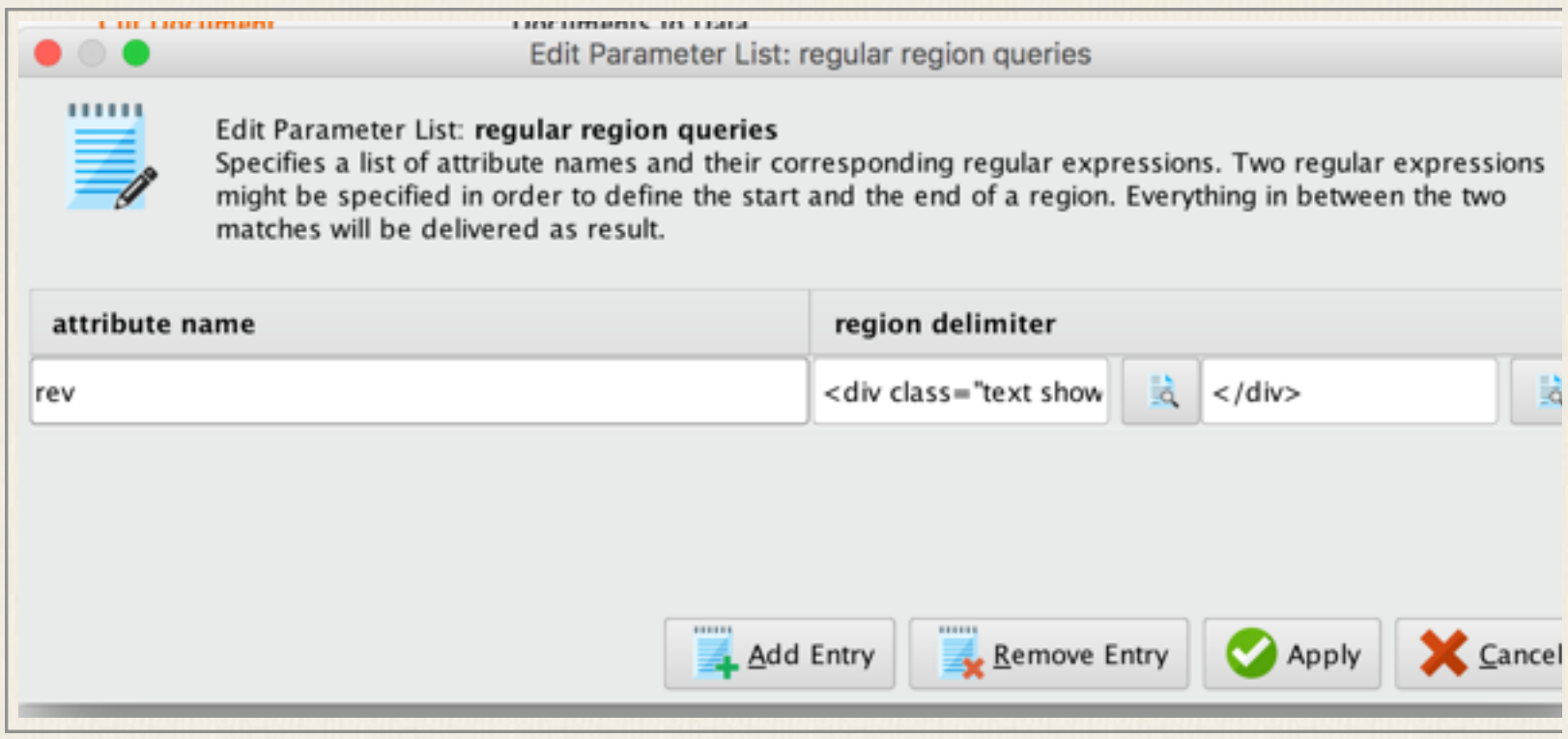
**No CGI overkill, just some fine acting and directing**

Madsen7 13 February 2017

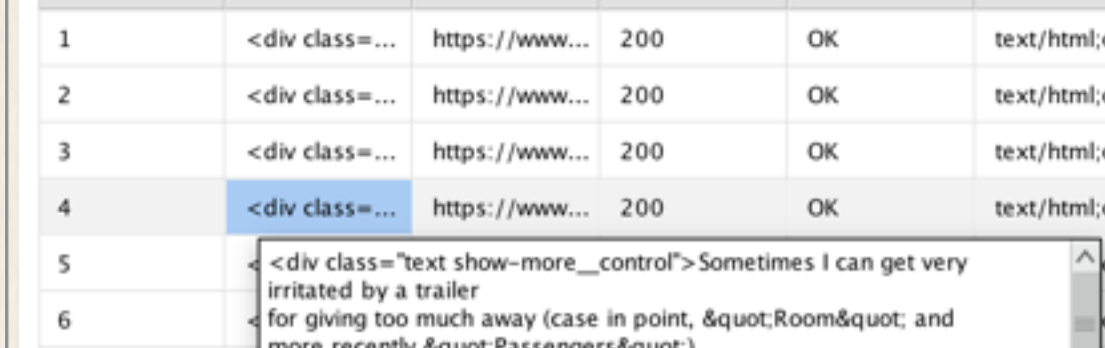
Denis Villeneuve is without a doubt an upcoming director and I can't wait to see Blade Runner 2049. Prisoners, Sicario and now Arrival (haven't seen Enemy yet, or some of his earlier work). Denis knows how to capture the tension. It is almost the strongest point of all of his films. Minimal use of computer generated images, and main focus on story, characters, acting and thrilling scenes. Back on IMDb board, I have noticed a lot of people calling Arrival a boring movie. So many hypocrites these days. People trying to point out every possible flaw to look and sound smarter while at the same time hailing

เราต้องการสกัดตัวบทวิจารณ์ที่เห็นของแต่ละคนมาเก็บรวบรวมเป็นคลังข้อมูล Get Page จะดึงข้อมูลในหน้าเว็บนั้นออกมาเป็น document จากนั้นจึงใช้ operator Cut Document ทำหน้าที่สกัดเฉพาะส่วนของข้อความที่ต้องการ โดยสามารถระบุวิธีการสกัดได้ ในตัวอย่างนี้นี้จะใช้ regular

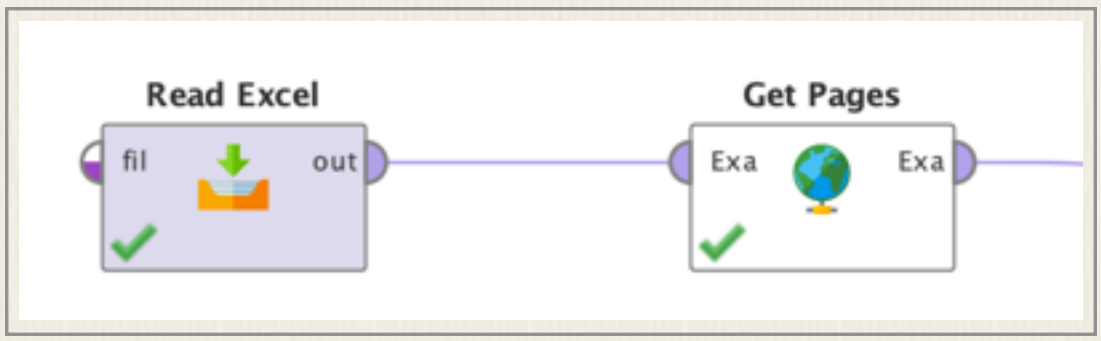
region โดยระบุ tag เริ่มต้นที่ต้องการและ tag ปิด  ในตัวอย่างนี้ บทวิจารณ์แต่ละคนจะอยู่ภายในแท็ก  ... <.div>  หากเราต้องการข้อมูลส่วนอื่นก็สามารถเติม field อื่น ๆ เพิ่มได้




จากนั้นจึงส่งผลที่สกัดเป็น document ได้ส่งต่อ operator Document to Data เพื่อแปลงข้อมูลที่ได้เป็น example set (exa)  ตัวอย่างผลที่สกัดได้เป็นตารางข้อมูลที่มี field “rev” ตามที่ต้องการ



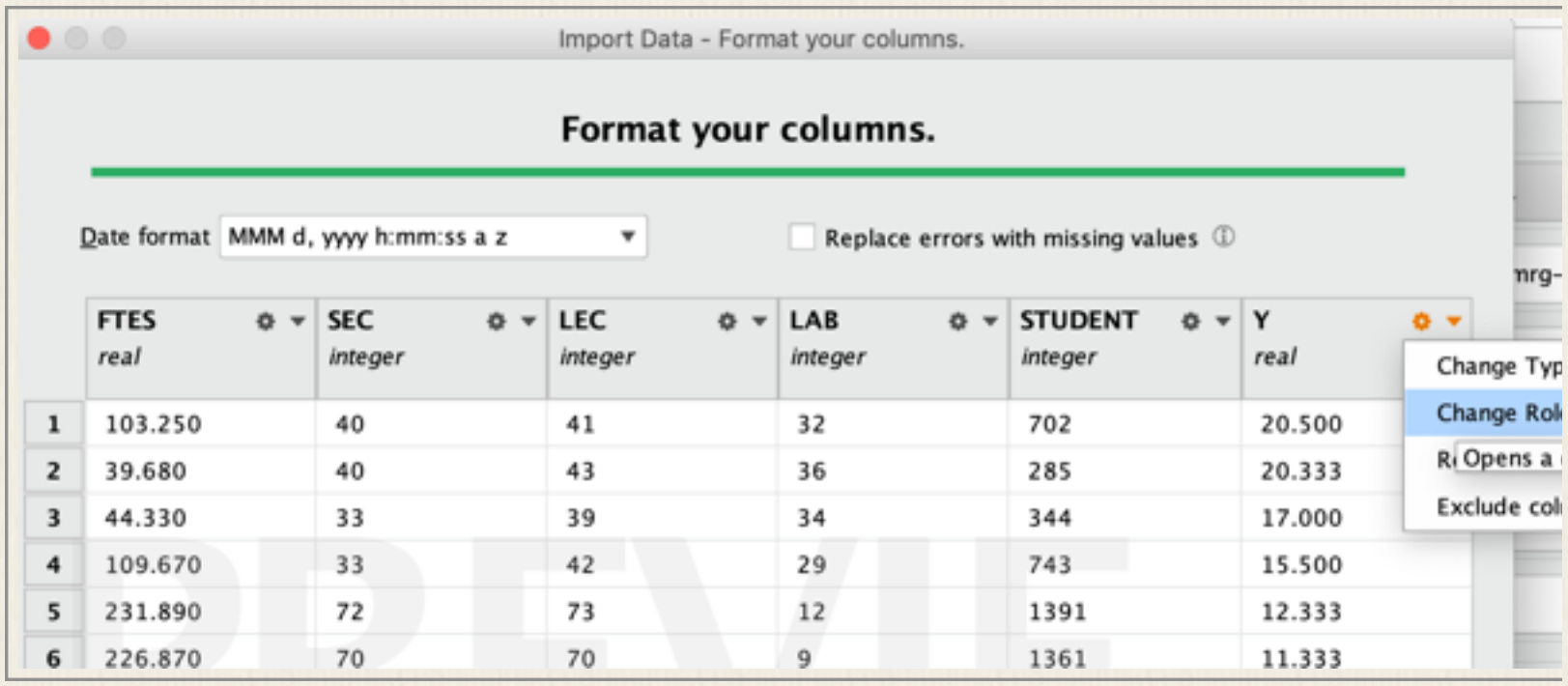

หากต้องการดึงข้อมูลจากเว็บหลาย ๆ page กสามารถใช้ operator Get Pages ได้โดยเอารายการ url ทั้งหมดไปเก็บไว้ใน Excel ไฟล์ก่อน ใช้ operator Read Excel เพื่ออ่านข้อมูลมาให้กับ operator Get Pages

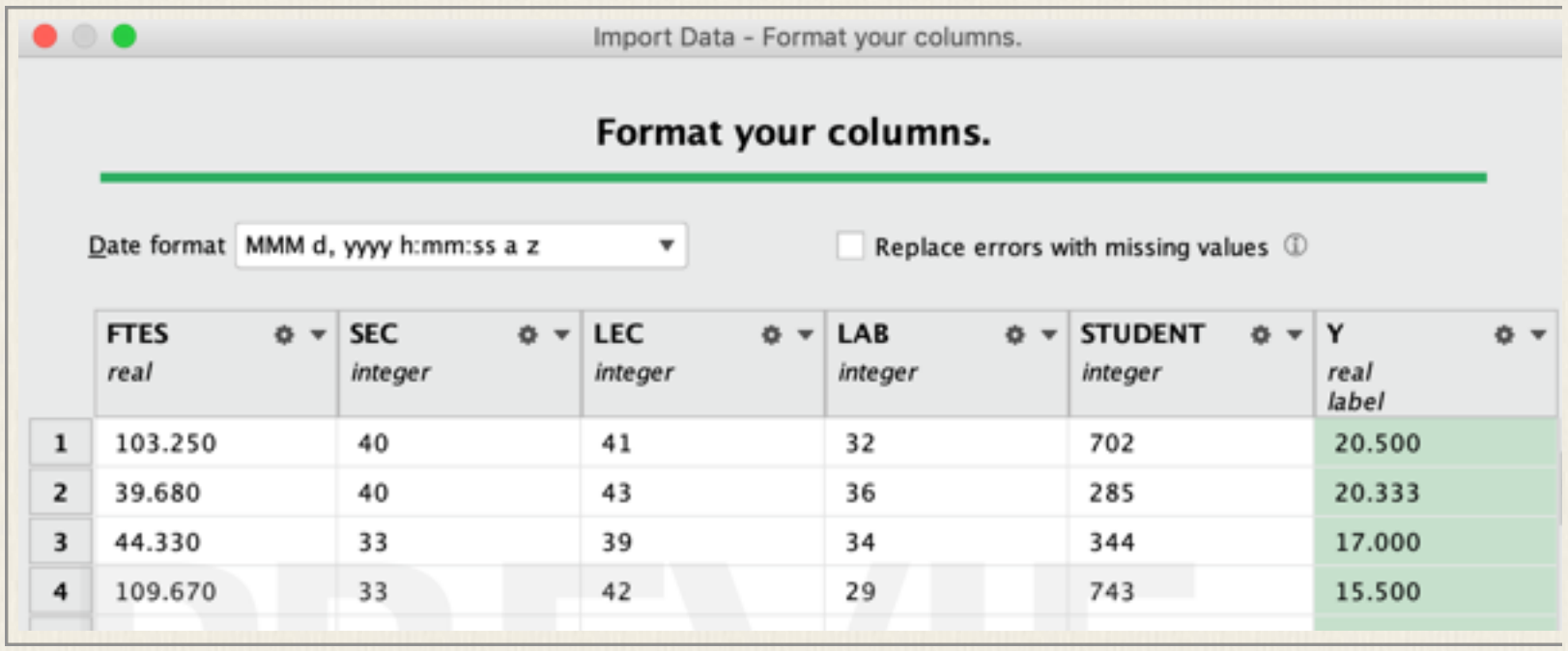



**การสร้าง model สำหรับ multiple regression**

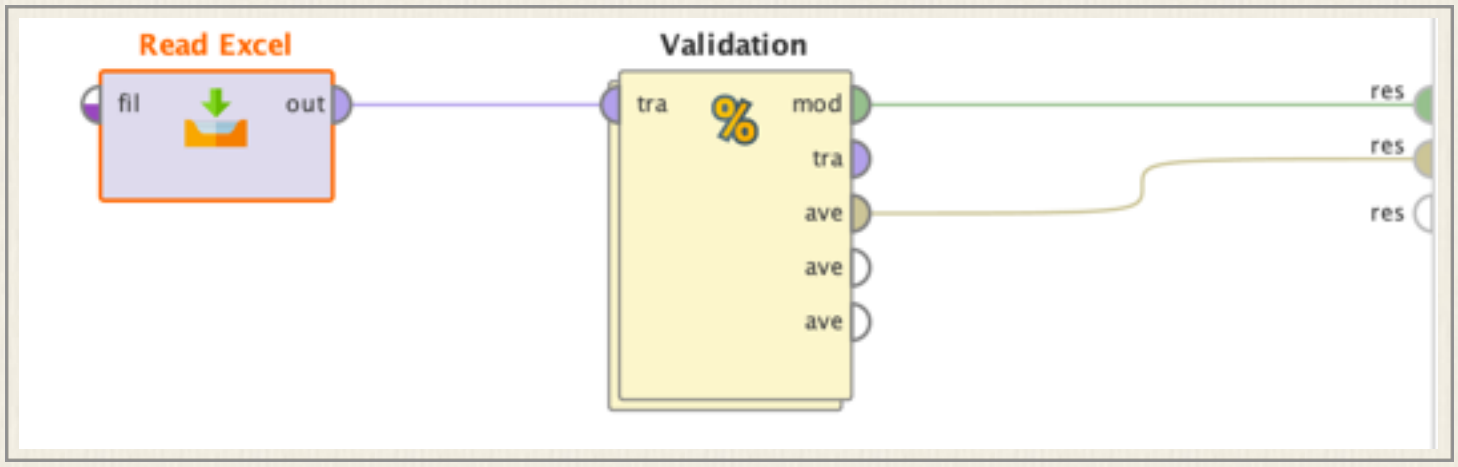
multiple regression model เป็นการสร้างโมเดลสำหรับทำนายค่าตัวแปรตามโดยที่ตัวแปรค้นที่เกี่ยวข้องอาจมีหลาย ๆ ตัวแปร ตัวอย่างต่อไปนี้ ข้อมูลเก็บไว้อยู่ใน Excel คอลัมน์ที่เป็นตัวแปรตามคือ Y ที่เหลือเป็นตัวแปรที่อาจเกี่ยวข้องกับโมเดลได้ .ให้ใช้ import wizard เพื่อตรวจสอบข้อมูล

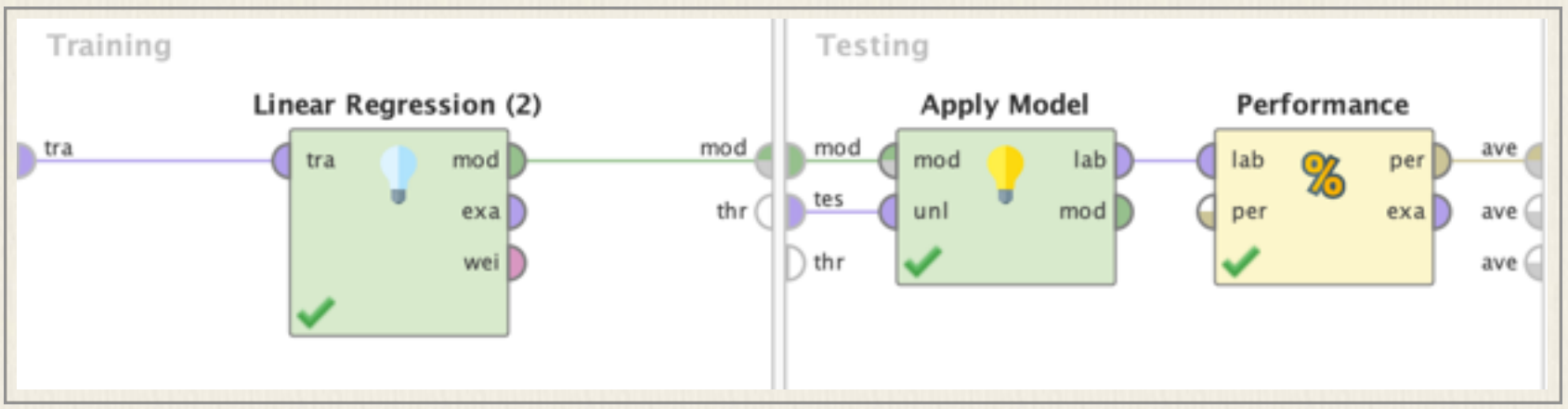
ว่านำเข้าได้ถูกประเภท และจะได้กำหนด role ของคอลัมน์สุดท้ายซึ่งเป็นข้อมูลตัวแปรตามให้เป็น label ตามตัวอย่างในรูป ซึ่งเมื่อเปลี่ยน role แล้วจะเห็นเป็นแถบสีเขียวขึ้น




หลังจากนั้น จึงเลือก operator Validation ซึ่งไว้สำหรับสร้างและทดสอบโมเดล ภายใน Validation จะแยกเป็นสองส่วน คือ training กับ testing เลือก split ratio ถ้าเป็น 0.7 หมายความว่า 70% ใช้ training 30% ใช้ testing




ภายใน Training ให้นำ Linear Regression มาใช้ และนำ Apply Model กับ Performance มาใช้เพื่อทดสอบโมเดลที่สร้างและรายงานความถูกต้อง ผลที่ได้เป็นค่าสัมประสิทธิ์ของตัวแปรต่างๆ ที่จะเกี่ยวข้อง

| Attribute | Coefficient | Std. Error | Std. Coef... | Tolerance | t-Stat | p-Value | Code |
|---|---|---|---|---|---|---|---|
| SEC | 0.195 | 0.016 | 0.798 | 0.454 | 11.952 | 0 | **** |
| LEC | -0.051 | 0.010 | -0.222 | 0.491 | -5.232 | 0.000 | **** |
| LAB | 0.430 | 0.014 | 0.659 | 0.736 | 31.116 | 0 | **** |
| STUDENT | -0.001 | 0.001 | -0.129 | 0.588 | -2.559 | 0.012 | ** |
| (Intercept) | -0.675 | 0.169 | ? | ? | -3.984 | 0.000 | **** |

## LinearRegression

```
  0.195 * SEC
- 0.051 * LEC
+ 0.430 * LAB
- 0.001 * STUDENT
- 0.675
```

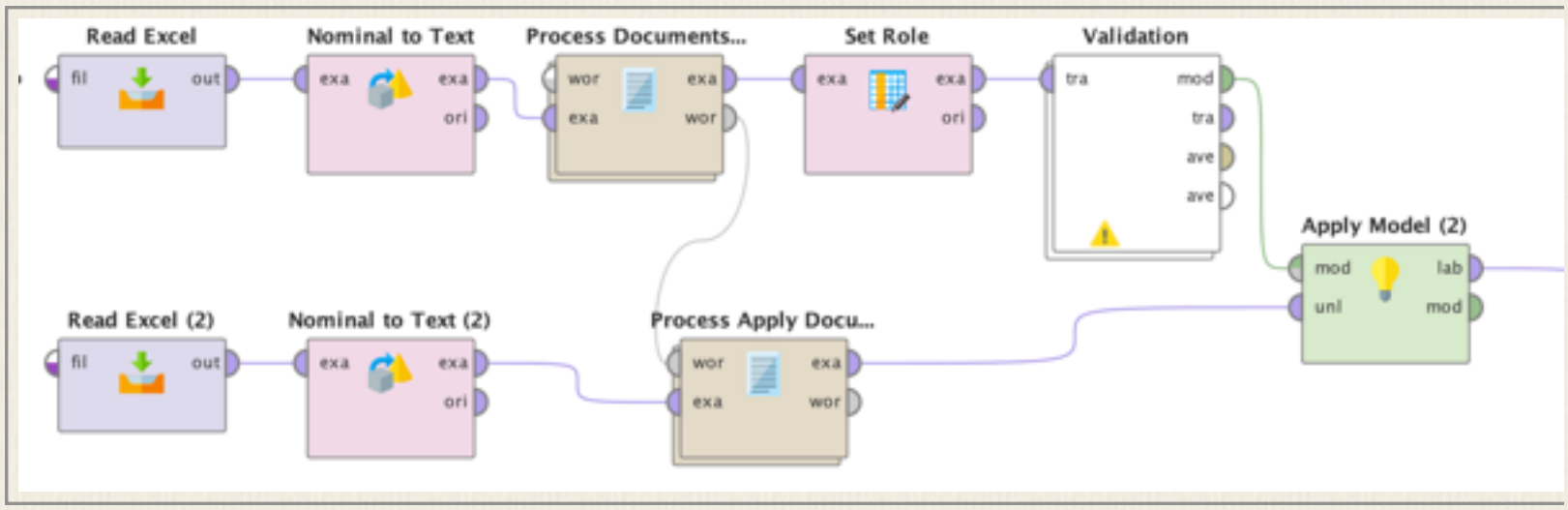
## root_mean_squared_error

```
root_mean_squared_error: 0.819 +/- 0.000
```

ในกรณีที่ต้องการใช้โมเดลแบบอื่นและทดสอบ เช่น decision tree ก็สามารถทำลักษะแบบเดียวกันนี้ได้ อย่างไรก็ดี ตัวโมเดลนี้โปรแกรม RapidMiner สร้างให้โดยขึ้นกับข้อมูลที่นำเข้า ส่วนผลที่ได้ว่าโมเดลบอกถึงความสัมพันธ์ระหว่างตัวแปรต่าง ๆ จริงหรือไม่ เป็นเรื่องที่ผู้วิจัยต้อง verify ข้อมูลต่าง ๆ ก่อนว่าเข้าเงื่อนไขที่จะใช้โมเดลนั้น ๆ ได้หรือไม่

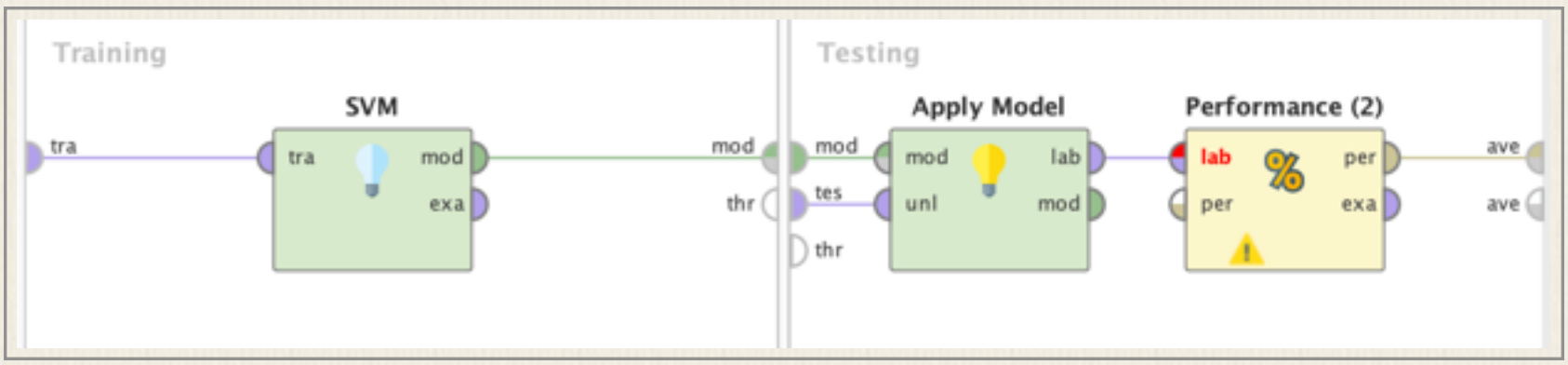
**การสร้าง model แบบง่ายสำหรับ sentiment analysis**

sentiment analysis เป็นการวิเคราะห์ข้อความเพื่อจัดประเภทว่าข้อความนั้นว่าบอกลักษณะอารมณ์แบบใด วิธีแบบง่าย ๆ ที่ทำกันคือแยกเป็นสองขั้วอารมณ์คืออารมณ์บวกหรือลบ โดยมีข้อมูลชุดหนึ่งเป็นข้อมูลที่มีการวิเคราะห์ให้ label แล้วว่าข้อความนั้น ๆ บอกอารมณ์บวกหรือลบ และใช้เป็นข้อมูลฝึกสอนเพื่อสร้างโมเดลสำหรับไว้ตัดสินข้อความอื่น ๆ ว่าบอกอารมณ์ไปในทางใด ตัวอย่างข้างล่างเป็นตัวอย่างจาก forum ของผู้ใช้ RapidMiner ที่อธิบายการสร้าง SVM model สำหรับ sentiment analysis (https://community.rapidminer.com/discussion/31827/sentiment-analysis-as-a-supervised-learning-problem)




ตัวอย่างข้อมูลในไฟล์ sentiment_training เป็น Excel ไฟล์มีสองคอลัมน์ คือ Text กับ Sentiment เป็นข้อมูลที่กำกับขั้วอารมณ์ไว้แล้ว operator ต่าง ๆ ตามรูปทำหน้าที่ตามลำดับ คือ อ่านไฟล์ Excel เข้ามา ใช้ “Nominal to Text” เพื่อระบุคอลัมน์เดียวให้เป็น text ก่อนจะส่งไป “Process Documents from Data” ซึ่งทำ “Tokenize” คำออกมาและสร้าง word vector โดยใช้

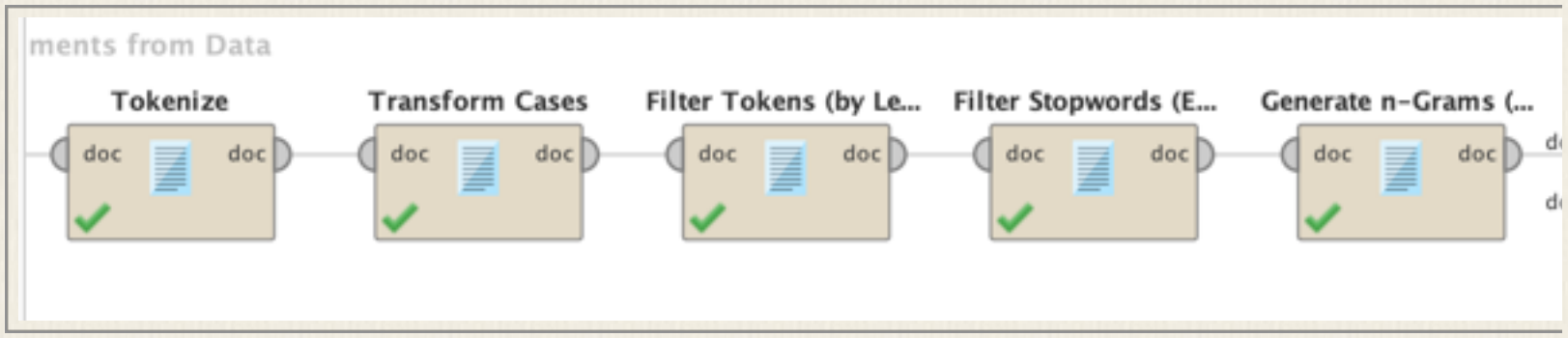
TF-IDF ก่อนจะส่งให้ “Set Role” เพื่อกำหนดว่า Sentiment เป็นข้อมูลที่เป็น label แล้วใช้ Validation เพื่อทำ training และ testing จากข้อมูลโดยในที่นี้เลือกใช้ SVM model และรายงานผลมาเป็น %accuracy (หากต้องการใช้ model อื่น เช่น Naive Bayes หรือ Decision Tress ก็เปลี่ยน operator จาก SVM เป็นตัวที่ต้องการใช้แทน)




ในการใช้งานกับข้อมูลอื่นภายหลัง คือ ไฟล์ Excel “sentiment_actual” ซึ่งมีสองคอลัมน์คือ Text กับ Sentiment แต่คอลัมน์หลังยังไม่มีการกำกับขั้วอารมณ์ไว้ ขั้นตอนจะเหมือนกับตอน train คืออ่านไฟล์ Excel มา บอกให้รู้ว่าคอลัมน์ “Text” เป็นส่วนของข้อความ ก่อนจะส่งไป “Process Documents from Data” ซึ่งรับข้อมูลใหม่นี้กับรายการคำจาก training data มาสร้างเป็น word vector โดยใช้ TF-IDF จากนั้นส่งไปที่ input “unl” (unlabelled) และใช้ model ที่ได้จากการ trainning มาคำนวณน้ำหนักขั้วบวกหรือลบออกมา ได้อะไรมากกว่าก็ทำนายว่าเป็นอารมณ์นั้น ตามผลลัพธ์ที่แสดง จะเห็น prediction ของ text นั้นว่าเป็น positive หรือ negative และจะเห็นค่า confidence (negative) กับ confidence (positive) ที่คำนวณออกมาได้ของแต่ละ text

จะเห็นได้ว่าวิธีการนี้เป็นวิธีการแบบง่าย ๆ คือมองที่คำเป็นหลักในการสร้างโมเดลมาทำนายต่อ ไม่มีการใช้ข้อมูลทางภาษาระดับอื่นมาใช้ด้วย เป็นวิธีการมองภาษาแบบที่เรียกว่าเป็นถุงคำ (bag of words) ลำดับการปรากฏของคำในประโยคไม่ได้ถูกนำมาใช้ด้วย หากต้องการทำให้ดีขึ้นอาจใช้ n-gram ของคำมาเป็น token ในการสร้าง vector แทนที่จะใช้แค่คำเดียวก็ได้ กรณีนี้ทำได้โดยการปรับ Process Documents from Data ให้เป็นดังรูป คือ มีการ “Generate n-gram” มาใช้แทนคำ

| prediction(... | confidence(... | confidence(... | text |
|---|---|---|---|
| positive | 0.486 | 0.514 | event events people waited twenty y... |
| positive | 0.473 | 0.527 | accepting oscar producer year s pic... |
| positive | 0.330 | 0.670 | people dislike french films lack clos... |
| positive | 0.451 | 0.549 | synopsis committed asylum marquis... |
| positive | 0.494 | 0.506 | documentary twin hughes brothers ... |
| positive | 0.474 | 0.526 | mimi leder known stunning work dir... |
| positive | 0.458 | 0.542 | edward burns tackles third picture l... |
| positive | 0.377 | 0.623 | plot young man loves heavy metal m... |
| positive | 0.355 | 0.645 | warren beatty returns screens funni... |
| negative | 0.502 | 0.498 | tom dicillo directs superficial comed... |
| positive | 0.377 | 0.623 | bleak look boston underworld oper... |



CHAPTER 4

# โปรแกรมเพื่องานภาษาศาสตร์

ส่วนนี้จะกล่าวถึงโปรแกรมต่างๆ ที่สามารถนำมาใช้เพื่องานวิจัยทางภาษาศาสตร์ได้ ได้แก่

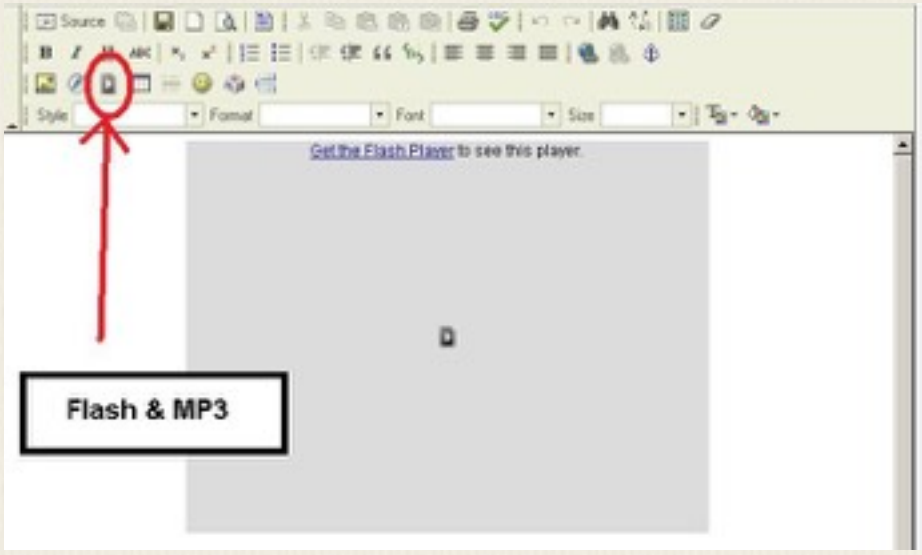
- การทำแบบสอบถามออนไลน์
- การเขียนต้นไม้ด้วยโปรแกรม TreeForm
- โปรแกรมกำกับข้อมูล POS/Semantic Tagger
- การวิเคราะห์เสียงด้วยโปรแกรม Praat
- การสร้าง Wordcloud จากข้อมูลภาษา
- การสร้างภาพความสัมพันธ์เครือข่าย
- การสกัดข้อมูลจากเว็บ
- การทำ Dimension Reduction ของข้อความ
- การทำ Topic Modeling จากคลังข้อมูล
- การกำกับข้อมูลด้วยโปรแกรม Corpus Annotator

# การทำแบบสอบถามผ่านคอมพิวเตอร์

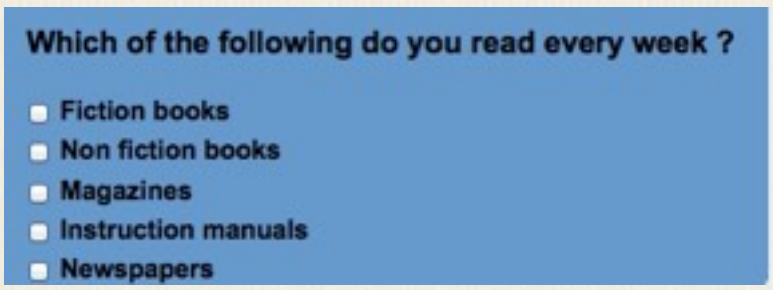
การทำแบบสอบถามผ่านคอมพิวเตอร์

มีเครื่องมือและเว็บไซ้ต์จำนวนมากที่เปิดให้บริการสร้างแบบสอบถามและจัดส่ง link ทางอีเมล์ให้ทำแบบสอบถาม และสรุปผลให้ ส่วนมากเป็นบริการที่คิดเงิน มีให้ใช้ฟรีได้บ้างแต่มักจะมีข้อจำกัด เช่น จำนวนข้อที่ถามไม่เกิน 20 ข้อ จำนวนคนที่ตอบแบบสอบถามไม่เกิน 200 คน ตัวอย่างเช่น http://www.survs.com/ ใช้บริการฟรีได้ไม่เกิน 200 response แต่บางที่ยอมให้ใช้ได้เป็นปริมาณมาก เช่น http://www.kwiksurveys.com/ แต่จะคิดค่าบริการจากการให้คำปรึกษาแทน Kwik surveys http://www.kwiksurveys.com/

สามารถแสดงคำถามเป็น text หรือเป็นเสียง mp3 หรือ flash วีดีโอได้




สามารถเลือกสร้างรูปแบบคำตอบได้หลายแบบ ตัวอย่างเช่น



**Please rate the following aspects of our service**

| | Excellent | Good | Neutral | Fair | Poor | N/A |
|---|---|---|---|---|---|---|
| Room Service | ○ | ○ | ○ | ○ | ○ | ○ |
| Restaurant | ○ | ○ | ○ | ○ | ○ | ○ |
| Laundry service | ○ | ○ | ○ | ○ | ○ | ○ |
| Booking process | ○ | ○ | ○ | ○ | ○ | ○ |

**Please Enter Your Contact Details**

First Name

Last Name

Telephone Number

Email Address

**Rank the following items of importance when going to the beach:**

**Scale:**

- 1 Must Have
- 5 Do Not Need.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Ice Cream | ○ | ○ | ○ | ○ | ○ |
| Sun Cream | ○ | ○ | ○ | ○ | ○ |
| Deck Chair | ○ | ○ | ○ | ○ | ○ |
| Towel | ○ | ○ | ○ | ○ | ○ |
| Sun Hat | ○ | ○ | ○ | ○ | ○ |

Google doc เป็นอีกวิธีหนึ่งที่สามารถใช้สร้างแบบสอบถาม online ได้ฟรี เพียงแค่มี Google account แล้วสร้าง form ขึ้นมาใน Google doc รายละเอียดสามารถอ่านได้จาก tutorial ต่อไปนี้

https://support.google.com/docs/answer/87809?hl=en

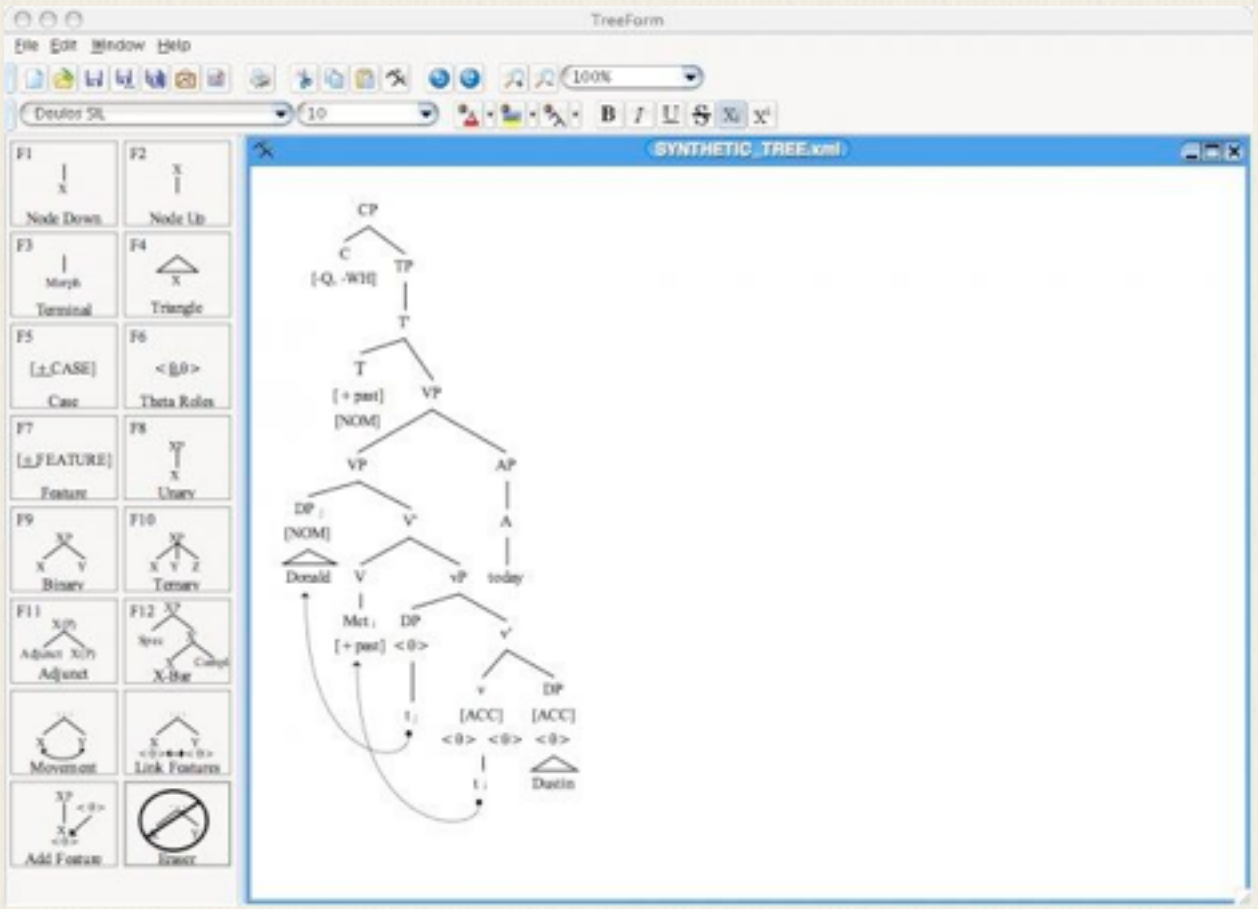
http://teacherlink.ed.usu.edu/tlresources/training2/Google/GoogleForms.pdf

*https://www.surveymonkey.com/ เป็นอีกหนึ่งทางเลือกที่สามารถสร้าง online questionnaire ได้ สามารถกำหนดให้หนึ่ง device มีสิทธิการทำหนึ่งครั้ง Google doc ความจริงก็ทำได้ แต่ใช้วิธี*

*การจำกัดการใช้ด้วยการเช็ค account  ผู้ที่จะทำแบบสอบถามที่กำหนดแบบนี้จึงต้องมี google account เพื่อ login*

# โปรแกรม TreeForm Syntax Tree Drawing

โปรแกรม TreeForm Syntax Tree Drawing เป็นโปรแกรม open source สามารถโหลดได้จาก http://sourceforge.net/projects/treeform/ด้วยภาษาจาวาจึงใช้กับเครื่อง windows หรือ Mac ก็ได้ เหมาะสำหรับงานที่จำเป็นต้องเขียนโครงสร้างต้นไม้ โปรแกรมจะช่วยอำนวยความสะดวกในการสร้าง node และ branch ต่างๆ ดังภาพตัวอย่างจากโปรแกรมนี้



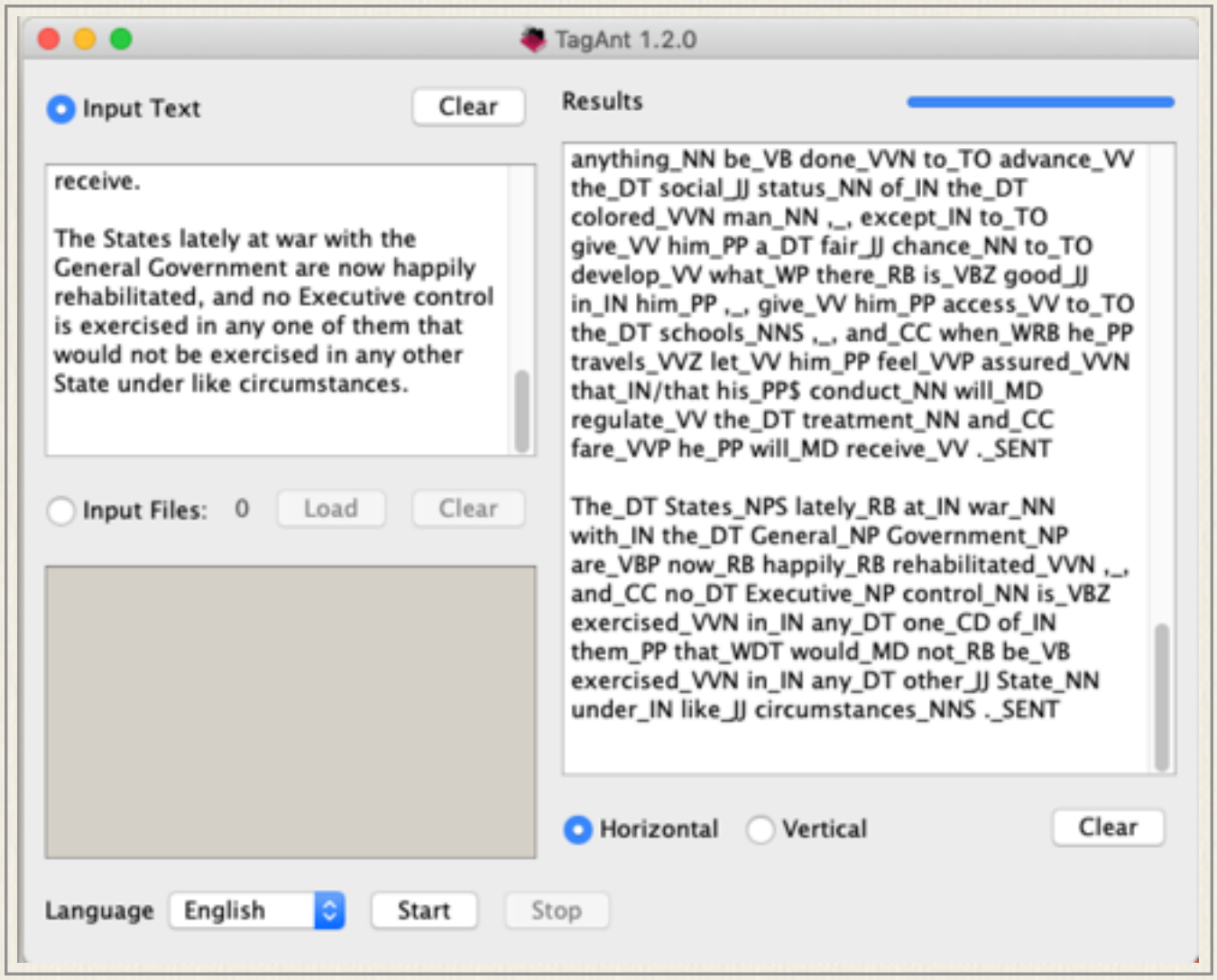
# โปรแกรม Tagger

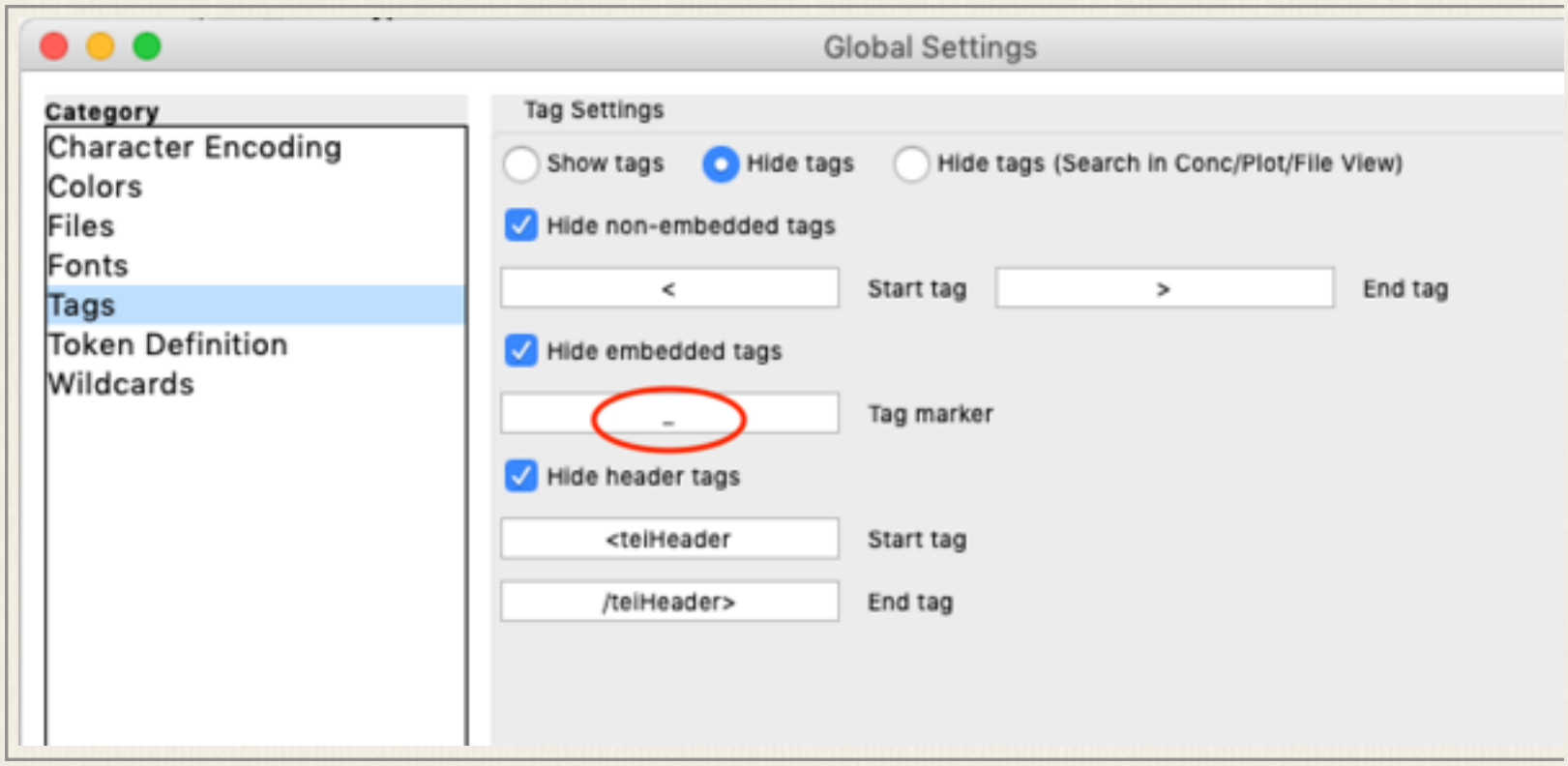
## โปรแกรม POS Tagger

โปรแกรมสำหรับ tag ข้อมูลโดยอัตโนมัติ ตัวพื้นฐานเป็น POS tagger ที่แท็กข้อมูลหมวดคำ โปรแกรม TagAnt เป็นโปรแกรม POS Tagger ที่ให้ใช้ได้ฟรีถ้าไม่ได้ใช้ในเชิงพาณิชย์ โดย Anthony ได้นำโปรแกรม Tree Tagger ซึ่งใช้ Penn Treebank Tag set สำหรับภาษาอังกฤษ (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/) มาใช้โดยทำ GUI ครอบเพื่อให้เรียกใช้งานโปรแกรม tagger ได้ง่ายขึ้น โปรแกรม TagAnt สามารถเลือก tag POS ของภาษาเหล่านี้ได้ German, English, French, Italian, Dutch, Spanish ได้ตัวโปรแกรม Tree Tagger เอง มีโมดูลให้ใช้กับภาษาต่อไปนี้ได้ด้วย คือ Danish, Bulgarian, Russian, Portuguese, Galician, Greek, Chinese, Swahili, Slovak, Slovenian, Latin, Estonian, Polish, Romanian, Czech

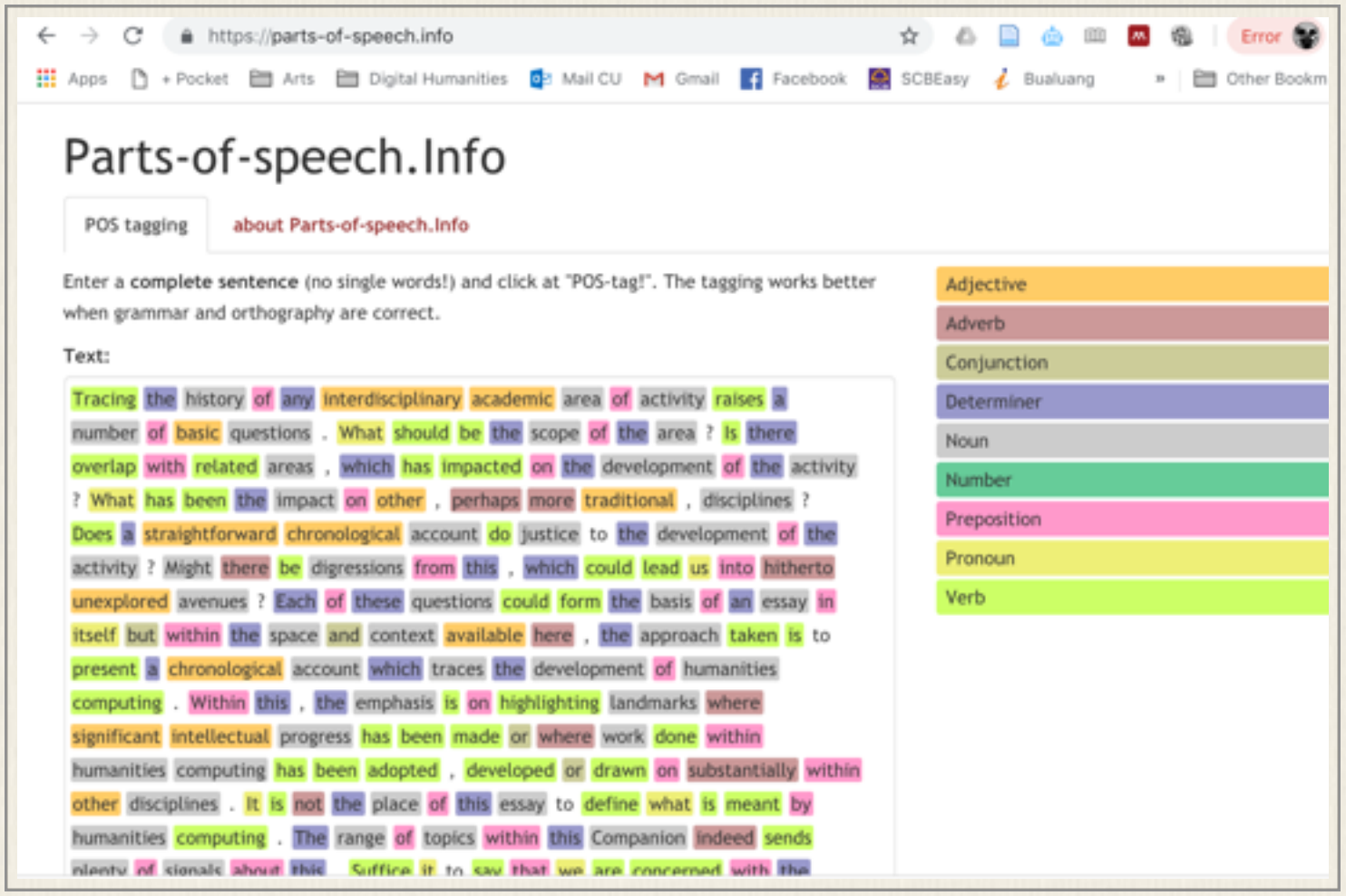
การใช้งานโปรแกรมทำได้โดยการ copy text เข้ามา เลือกภาษาที่ต้องการและสั่ง start ให้แท็กข้อมูล ผลลัพธ์ที่ได้มีสองแบบ แบบที่สามารถนำไปใช้กับโปรแกรม AntConc ต่อได้คือแบบแนวนอนที่ POS จะถูกแท็กตามหลังคำด้วยเครื่องหมาย _ ตามที่เห็นในหน้าจอผล จากนั้นจึง copy ผลที่ได้ไปเก็บเป็นไฟล์สำหรับใช้งานต่อไป

ข้อมูลที่แท็กด้วย TagAnt สามารถนำไปใช้ร่วมกับโปรแกรม AntConc ซึ่งเป็นโปรแกรม concordance ได้ทันที โดยไป set ที่ Global Setting ของ AntConc ว่าให้ show หรือ hide tags ค่า default กำหนด embedded tags ด้วยเครื่องหมาย _ จึงทำให้ AntConc สามารถใช้กับข้อมูลแท็กแบบนี้หรือแท็กตามแบบมาตรฐาน TEI ได้


TagAnt 1.2.0
Input Text
Clear
receive.
The States lately at war with the General Government are now happily rehabilitated, and no Executive control is exercised in any one of them that would not be exercised in any other State under like circumstances.
Input Files: 0
Load
Clear
Language
English
Start
Stop
Results
anything_NN be_VB done_VVN to_TO advance_VV the_DT social_JJ status_NN of_IN the_DT colored_VVN man_NN ,_, except_IN to_TO give_VV him_PP a_DT fair_JJ chance_NN to_TO develop_VV what_WP there_RB is_VBZ good_JJ in_IN him_PP ,_, give_VV him_PP access_VV to_TO the_DT schools_NNS ,_, and_CC when_WRB he_PP travels_VVZ let_VV him_PP feel_VVP assured_VVN that_IN/that his_PP$ conduct_NN will_MD regulate_VV the_DT treatment_NN and_CC fare_VVP he_PP will_MD receive_VV ._SENT
The_DT States_NPS lately_RB at_IN war_NN with_IN the_DT General_NP Government_NP are_VBP now_RB happily_RB rehabilitated_VVN ,_, and_CC no_DT Executive_NP control_NN is_VBZ exercised_VVN in_IN any_DT one_CD of_IN them_PP that_WDT would_MD not_RB be_VB exercised_VVN in_IN any_DT other_JJ State_NN under_IN like_JJ circumstances_NNS ._SENT
Horizontal
Vertical
Clear


Global Settings
Category
Character Encoding
Colors
Files
Fonts
Tags
Token Definition
Wildcards
Tag Settings
Show tags
Hide tags
Hide tags (Search in Conc/Plot/File View)
Hide non-embedded tags
<
Start tag
>
End tag
Hide embedded tags
_
Tag marker
Hide header tags
<teiHeader
Start tag
/teiHeader>
End tag

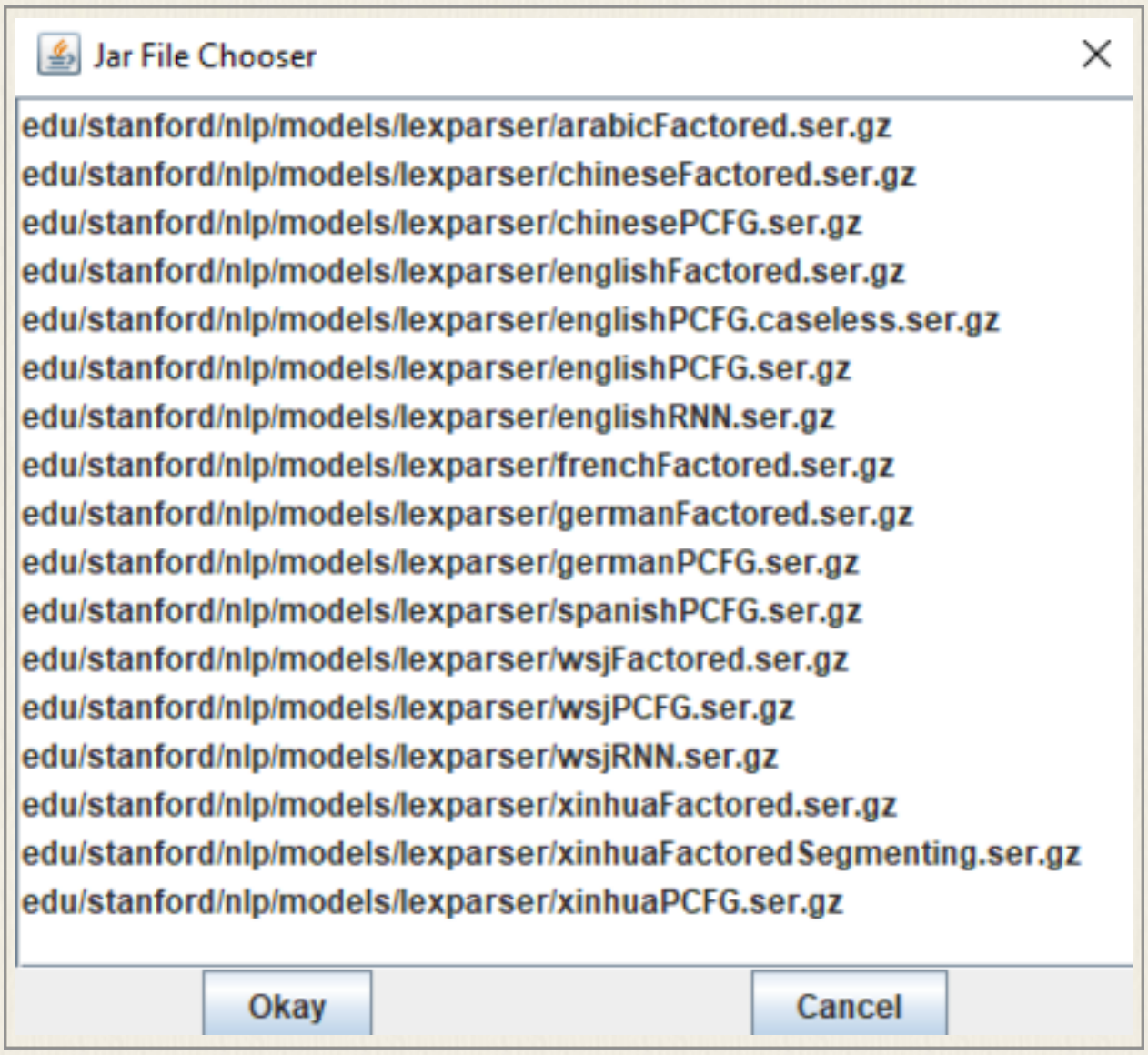
โปรแกรม POS Tagger ที่เป็นที่รู้จักแพร่หลายอีกตัวคือ CLAW Tagger ซึ่งเป็น POS Tagger สำหรับภาษาอังกฤษ เป็นบริการมีค่าใช้จ่าย แต่เปิดให้ทดลองใช้กับข้อมูลไม่เกิน 100,000 คำได้ที่ http://ucrel.lancs.ac.uk/claws/trial.html Tag set ที่ใช้จะเป็นอีกชุดหนึ่งต่างจาก Penn Tree-bank POS นอกจาก CLAW tagger ก็สามารถใช้ตัวอื่น เช่น Stanford POS Tagger เพียงแต่วิธีใช้จะต้องดาวน์โหลดโปรแกรมไปและเรียนใช้ผ่านคำสั่งที่กำหนดเอาเอง POS Tagger แต่ละตัวก็อาจใช้ POS tagset ที่แตกต่างกันได้ เวลาใช้งานจึงต้องดู POS Guideline ประกอบและเลือกตามความเหมาะสม บางเว็บไซต์อาจมีบริการให้แท็กข้อมูลแแนไลน์ได้ เช่น ที่ parts-of-speech.info ให้แท็กข้อมูลและให้ POS ออกมาเป็นแบบที่ผู้ใช้ทั่วไปรู้จัก ดังตัวอย่าง

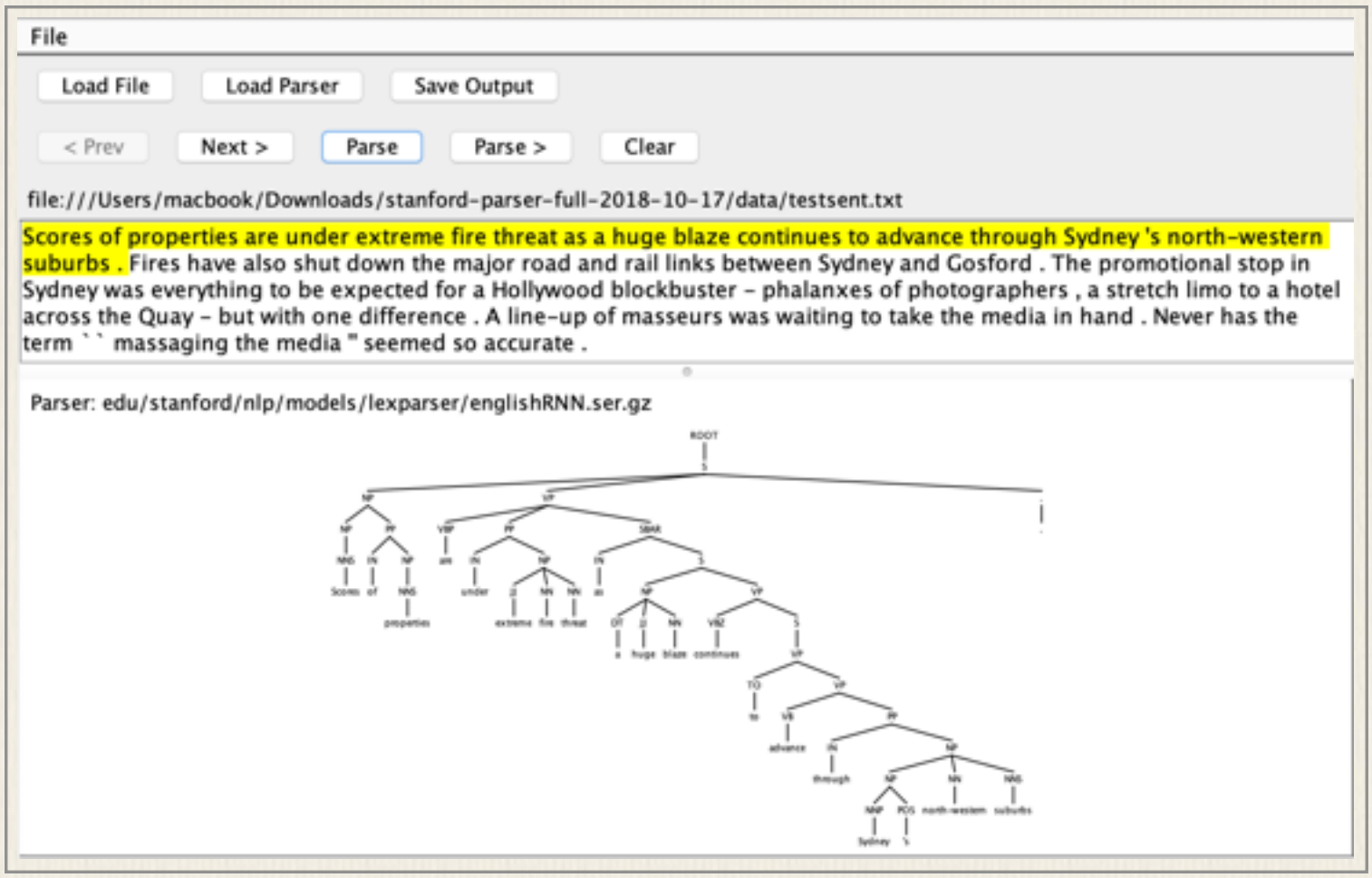



แต่หากคนที่สามารถเขียนโปรแกรมได้เอง จะพบว่ามี package ภาษาต่าง ๆ ให้เลือกใช้เองได้ เช่น ในภาษา Python จะมี package NLTK สำหรับงานประมวลผลภาษาอังกฤษเรื่องต่าง ๆ ถ้าต้องการใช้ POS Tagging ภาษาไทยก็สามารถเลือกใช้ TLTK ที่มีฟังก์ชั่น pos_tag ให้ใช้ได้

**โปรแกรม Parser**

โปรแกรม parser เป็นโปรแกรมที่นอกจากจะวิเคราะห์ POS Tag ในข้อมูลแล้วยังวิเคราะห์ความสัมพันธ์ทางวากยสัมพันธ์เพื่อสร้างโครงสร้างต้นไม้แบบต่าง ๆ อาจจะให้ผลของมา Phrase Structure Tree หรือ Dependency structure ก็ได้ ตัวอย่างโปรแกรมที่มีให้ใช้ คือ Stanford Statistical Parser (https://nlp.stanford.edu/software/lex-parser.shtml) ซึ่งสามารถดาวน์โหลดโปรแกรมและโมเดลสำหรับการ parse มาใช้ได้ ถ้าใช้บน Windows หลังจาก unzip ไฟล์ออกมาแล้วจะเห็นไฟล์ lexparser-gui.bat ให้ run โปรแกรมจากไฟล์นี้ ถ้าเป็น Mac ให้ run จากไฟล์ lexparser-gui.command จากนั้นทดลองเลือกตัวอย่างข้อมูลจากไฟล์ testsent.txt แล้วเลือก parser ที่ต้องการใช้โดยเลือกที่ไฟล์โมเดลก่อน เช่น stanford-parser-3.9.2-models แล้วค่อยระบุโมเดล parser ที่ต้องการใช้ โปรแกรม gui ช่วยให้เรียกโมเดลและเห็นต้นไม้ได้ชัด แต่ไม่สามารถ save output ได้ จะต้องเรียกใช้ผ่าน command line






เช่น  หากต้องการใช้ dependency parser ผ่านคำสั่ง ดังตัวอย่างนี้

```
java -Xmx2g -cp "*" edu.stanford.nlp.parser.nndep.DependencyParser \
-model edu/stanford/nlp/models/parser/nndep/english_UD.gz \
-textFile data/english-onesent.txt -outFile data/english-onesent.txt.out
```

ผลลัพธ์ที่ได้จากการ parse ประโยค “The quick brown fox jumped over the lazy dog.” จะเก็บไว้ที่ไฟล์ english-onesent.txt.out ตามที่ระบุในคำสั่งข้างบน  ผลลัพธ์จะแสดง head-dependent และ relation ระหว่างคำ  ดังตัวอย่างนี้

```
det(fox-4, The-1)
amod(fox-4, quick-2)
amod(fox-4, brown-3)
nsubj(jumped-5, fox-4)
root(ROOT-0, jumped-5)
```

case(dog-9, over-6)

det(dog-9, the-7)

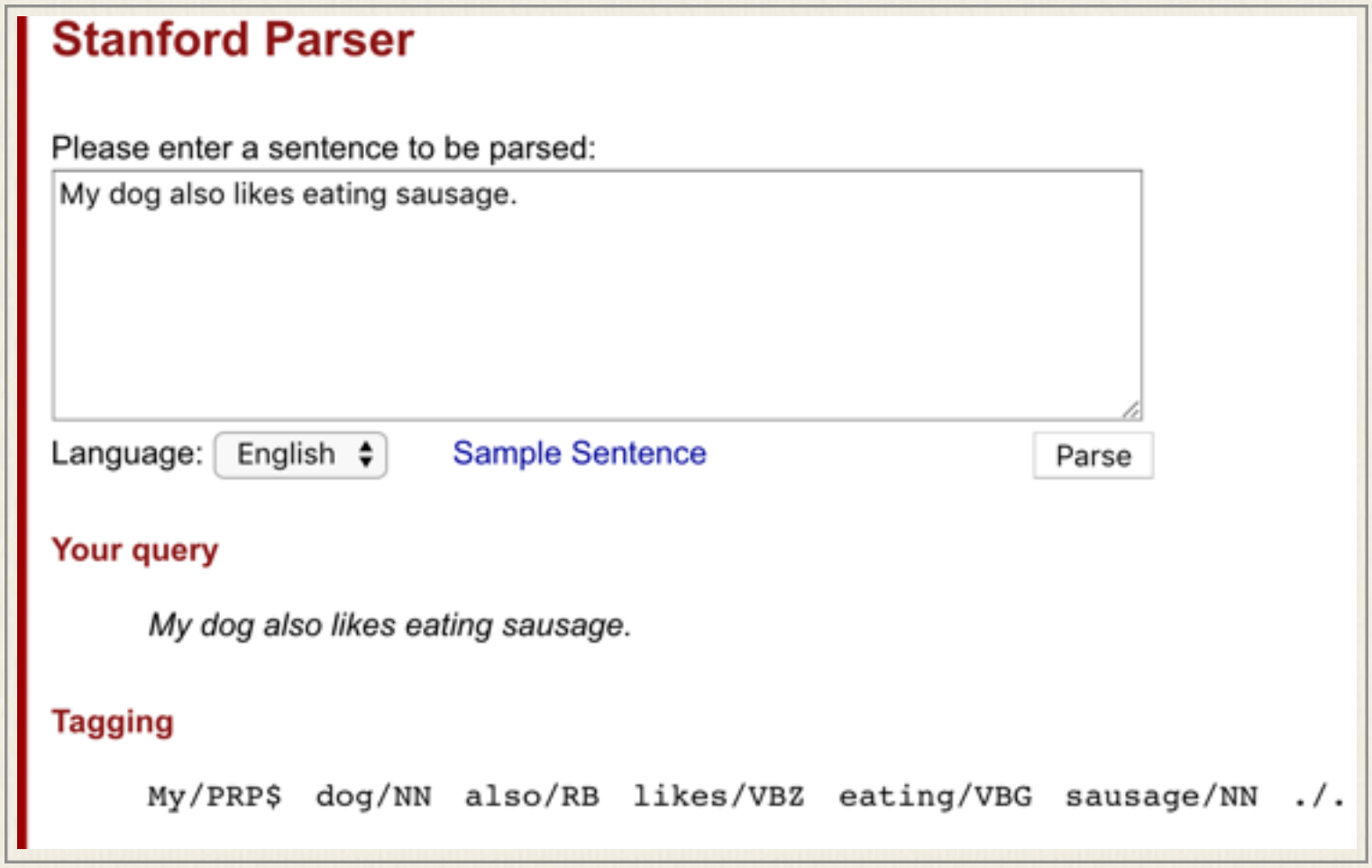
amod(dog-9, lazy-8)

nmod(jumped-5, dog-9)

punct(jumped-5, .-10)

แตกต่างจากผลที่เป็น PS tree (ROOT (S (NP (DT The) (JJ quick) (JJ brown) (NN fox)) (VP (VBD jumped) (PP (IN over) (NP (DT the) (JJ lazy) (NN dog)))) (. .)))

นอกจากโปรแกรมที่ให้ดาวน์โหลดมาใช้ Stanford Parser ยังมีเวอร์ชั่นที่ให้ทดลองใช้ออนไลน์ที่ http://nlp.stanford.edu:8080/parser/ ที่จะแสดงผลทั้ง POS Tagger, PS Tree, และ dependency tree สะดวกสำหรับการดูผลการวิเคราะห์ประโยคที่ต้องการได้ทันที



**Parse**

```
(ROOT
  (S
    (NP (PRP$ My) (NN dog))
    (ADVP (RB also))
    (VP (VBZ likes)
      (S
        (VP (VBG eating)
          (NP (NN sausage)))))
    (. .)))
```

**Universal dependencies**

```
nmod:poss(dog-2, My-1)
nsubj(likes-4, dog-2)
advmod(likes-4, also-3)
root(ROOT-0, likes-4)
xcomp(likes-4, eating-5)
dobj(eating-5, sausage-6)
```

## โปรแกรม Semantic Tagger

โปรแกรม semantic tagger จะแท็กข้อมูลคำโดยระบุว่าคำนั้นอยู่ในกลุ่มความหมายอะไร โปรแกรม semantic tagger ของ UCREL Semantic Analysis System (USAS) อาศัยการจัดกลุ่มคำในพจนานุกรม Longman ที่จัดกลุ่มคำเป็น 21 กลุ่มความหมายใหญ่ (discourse field) แล้วจึงแตกออกมาเป็น 232 ความหมายย่อย มีบริการออนไลน์ให้ใช้ครั้งละไม่เกิน 100,000 คำ ที่ http://ucrel.lancs.ac.uk/usas/tagger.html USAS และ Claw tagger เป็นบริการที่ใช้งานผ่าน Wmatrix ซึ่งเป็นโปรแกรมสำหรับทำงานกับคลังข้อมูลภาษา แต่ต้องสมัครสมาชิกและมีค่าใช้บริการ

# UCREL Semantic Analysis System (USAS)

USAS Home Page | English tagger | Dutch tagger | Chinese tagger | Italian tagger | Portuguese tagger | Spanish tagger | GUI download

## English Semantic Tagger

The UCREL semantic analysis system is a framework for undertaking the automatic semantic analysis of text. The framework has been designed and used across a number of research projects and this page collects together various pointers to those projects and publications produced since 1990.

The semantic tagset used by USAS was originally loosely based on Tom McArthur's Longman Lexicon of Contemporary English (McArthur, 1981). It has a multi-tier structure with 21 major discourse fields (shown here on the right), subdivided, and with the possibility of further fine-grained subdivision in certain cases. We have written an introduction to the USAS category system (PDF file) with examples of prototypical words and multi-word units in each semantic field.

| | | | |
|---|---|---|---|
| **A** general and abstract terms | **B** the body and the individual | **C** arts and crafts | **E** emotion |
| **F** food and farming | **G** government and public | **H** architecture, housing and the home | **I** money and commerce in industry |
| **K** entertainment, sports and games | **L** life and living things | **M** movement, location, travel and transport | **N** numbers and measurement |
| **O** substances, materials, objects and equipment | **P** education | **Q** language and communication | **S** social actions, states and processes |
| **T** Time | **W** world and environment | **X** psychological actions, states and processes | **Y** science and technology |
| **Z** names and grammar | | | |

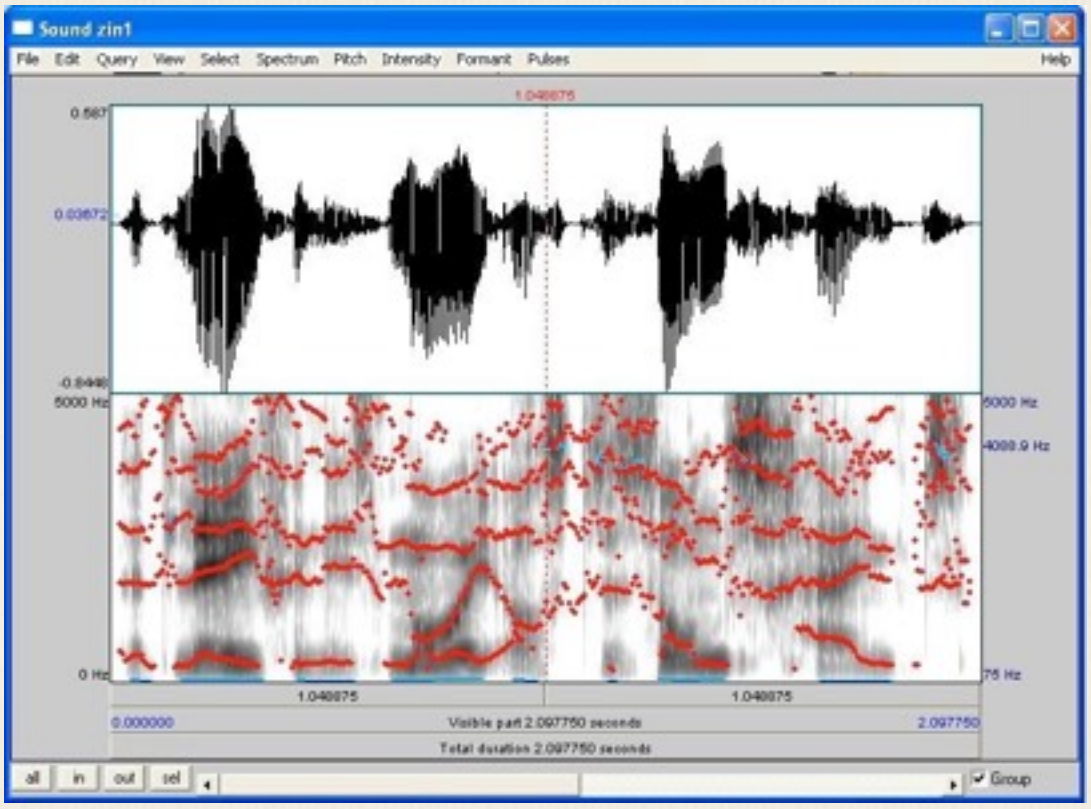
**USAS online** If you wish to view your semantically tagged text in a different style, press the back button on your browser, change the options and re-submit the form.

296 words tagged

```
Tracing_A10+ the_Z5 history_T1.1.1 of_Z5 any_N5.1+ interdisciplinary_P1
academic_P1 area_M7 of_Z5 activity_A1.1.1 raises_M2 a_N5+[i1.3.1
number_N5+[i1.3.2 of_N5+[i1.3.3 basic_A6.2+ questions_Q2.2 ._PUNC
What_Z8 should_S6+ be_A3+ the_Z5 scope_A1.7- of_Z5 the_Z5 area_M7 ?_PUNC
Is_A3+ there_Z5 overlap_N5.2+ with_Z5 related_A2.2 areas_M7 ,_PUNC which_Z8
has_Z5 impacted_A1.1.1 on_Z5 the_Z5 development_A2.1+ of_Z5 the_Z5
activity_A1.1.1 ?_PUNC
What_Z8 has_Z5 been_A3+ the_Z5 impact_A2.2 on_Z5 other_A6.1- ,_PUNC perhaps_A7
more_A13.3 traditional_S1.1.1 ,_PUNC disciplines_G2.1 ?_PUNC
Does_A1.1.1 a_Z5 straightforward_A12+ chronological_T1/N4 account_I1
do_A1.2+[i2.3.1 justice_A1.2+[i2.3.2 to_A1.2+[i2.3.3 the_Z5 development_A2.1+
of_Z5 the_Z5 activity_A1.1.1 ?_PUNC
Might_A7+ there_Z5 be_A3+ digressions_A6.1- from_Z5 this_Z8 ,_PUNC which_Z8
could_A7+ lead_S7.1+ us_Z8 into_Z5 hitherto_T1.1.1 unexplored_A10-
avenues_M3/H3 ?_PUNC
Each_N5.1+ of_Z5 these_Z5 questions_Q2.2 could_A7+ form_T2+ the_Z5 basis_A2.2
of_Z5 an_Z5 essay_Q1.2/P1 in_Z5 itself_Z8 but_Z5 within_Z5 the_Z5 space_N3.6
and_Z5 context_O4.1/A3+ available_A3+ here_M6 ,_PUNC the_Z5 approach_X4.2
taken_A9+ is_A3+ to_Z5 present_A9- a_Z5 chronological_T1/N4 account_I1
which_Z8 traces_A10+ the_Z5 development_A2.1+ of_Z5 humanities_P1 computing_Y2
._PUNC
Within_Z5 this_Z8 ,_PUNC the_Z5 emphasis_A11.1+ is_A3+ on_Z5
highlighting_X5.1+ landmarks_M7/A11.2+ where_M6 significant_A11.1+
```

# โปรแกรม Praat

โปรแกรม Praat เป็นโปรแกรมที่นิยมใช้วยงานวิจัยทางด้านเสียง สามารถโหลดมาใช้ได้ฟรีจาก http://www.fon.hum.uva.nl/praat/ มีทั้งที่ทำงานบน windows และ Mac สามารถใช้ในงาน speech analysis, speech synthesis, ทำ label and segmentation, manipulate speech เป็นต้น ผู้สนใจสามารถอ่านคำแนะนำการใช้งานโปรแกรม Praat เพื่องานทางภาษาศาสตร์ได้ที่ http://savethevowels.org/praat/ หรืออ่านคู่มือเบื้องต้นได้ที่ http://www.stanford.edu/dept/linguistics/corpora/material/PRAAT_workshop_manual_v421.pdf และ http://savethevowels.org/praat/UsingPraatforLinguisticResearchLatest.pdf



*สำหรับคนที่ใช้ภาษา Python จะมี library Parselmouth ที่สามารถใช้งาน Praat ได้เลย รายละเอียดสามารถศึกษาได้จาก https://parselmouth.readthedocs.io/*

# โปรแกรมสร้างการทดลองภาษาศาสตร์จิตวิทยา

โปรแกรมสร้างการทดลองภาษาศาสตร์จิตวิทยามีหลายโปรแกรมที่สามารถใช้ได้

- DMDX (http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm)

  เป็นฟรีโปรแกรมรุ่นแรกๆ จึงทำงานบน DOS หรือ windows XP นำเสนอคำถามที่เป็นภาพและเสียง แล้วเก็บ response กับ reaction time จากคำตอบผู้ทดลอง ผู้สนใจสามารถดู Tutorial ได้ในเว็บของ Matt Davis (http://www.mrc-cbu.cam.ac.uk/people/matt.davis/dmdx.html)
- E Prime (https://www.pstnet.com/eprime.cfm)

  เป็นโปรแกรมจำหน่ายสำหรับทำการทดลองทางภาษาศาสตร์จิตวิทยาที่ได้รับความนิยมเพราะสะดวกในการใช้งาน
- Psyscope X (http://psy.ck.sissa.it/)

  เป็นโปรแกรมฟรีทำงานบน OS X นำเสนอคำถามที่เป็นภาพและเสียง แล้วเก็บ response กับ reaction time จากคำตอบผู้ทดลองได้ ปัจจุบันกำลังปรับปรุงให้มี GUI ที่ง่ายขึ้นต่อการออกแบบการทดลอง แต่ไม่สามารถใช้งานบน Windows
- PXLab (http://irtel.uni-mannheim.de/pxlab/index.html)

  เป็นอีกโปรแกรมสำหรับงานทดลองทางจิตวิทยา เขียนด้วยภาษา Java เพื่อทำการทดลองรูปแบบต่างๆ ได้

- Psychopy (http://psychopy.org/)

  เป็นโปรแกรมฟรีเขียนด้วยภาษา Python สามารถใช้ได้ทั้งกับ Windows และ Max OS X มี GUI ที่ช่วยในการออกแบบการทดลองและกำหนดค่าต่างๆ ที่ต้องการใช้ในการทดลอง เป็นตัวเลือกที่

**Build your first PsychoPy experiment (Stroop task)  https://youtu.be/VV6qhuQgsiI**

สามารถใช้แทน E Prime ได้  มี Tutorial video ที่เห็นจะช่วยให้เข้าใจการใช้ Psychopy ในการออกแบบและสร้างการทดลองง่าย ๆที่แสดงข้อความและรอรับคำตอบจากผู้ทดลอง

# การสร้าง wordcloud

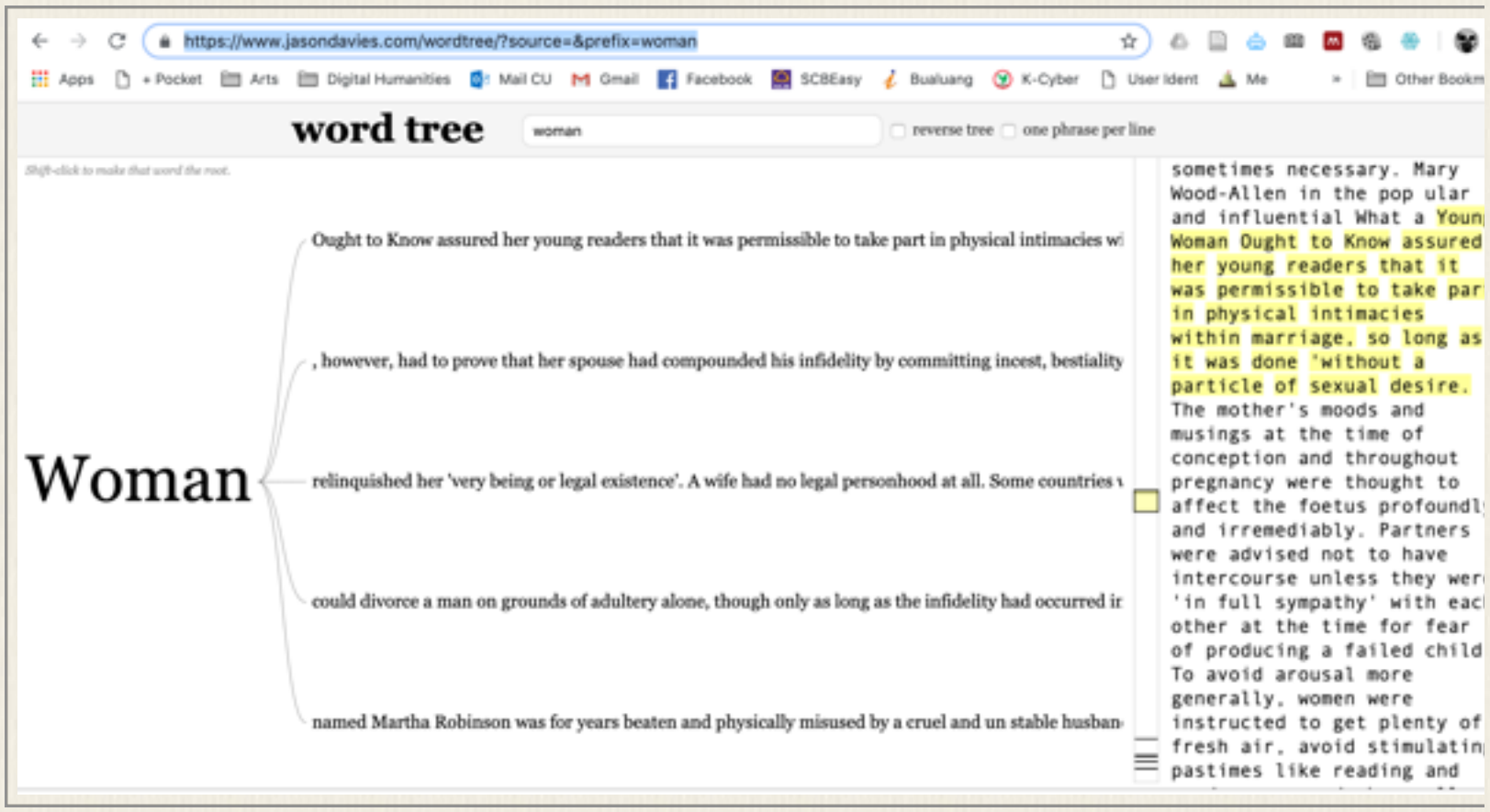
wordcloud เป็นรูปแบบของการแสดงคำที่ปรากฏในข้อมูลตามความถี่ที่พบ แต่แทนที่จะแสดงเป็นตัวเลขก็แสดงเป็นขนาดของตัวอักษรแทน คำที่ปรากฏใหญ่กว่าจะมีความถี่การปรากฏมากกว่า เราสามารถสร้าง wordcloud ได้ง่าย ๆ โดยใช้บริการที่มีในอินเทอร์เน็ต เช่น https://www.wordclouds.com/ สามารถ upload ไฟล์ขึ้นไปและแสดงผลเป็น wordcloud พร้อมปรับแต่งรูปแบบต่าง ๆ ได้

SECTION 7

# wordtree

เป็นการสร้าง tree จากตัวบท โดยการระบุคำที่ต้องการค้น โปรแกรมจะดูคำนั้นว่ามีอะไรตามท้ายได้บ้างไล่ไปเรื่อยๆ จนจบประโยค สามารถใช้งานโดยการ upload text ไปที่เว็บ

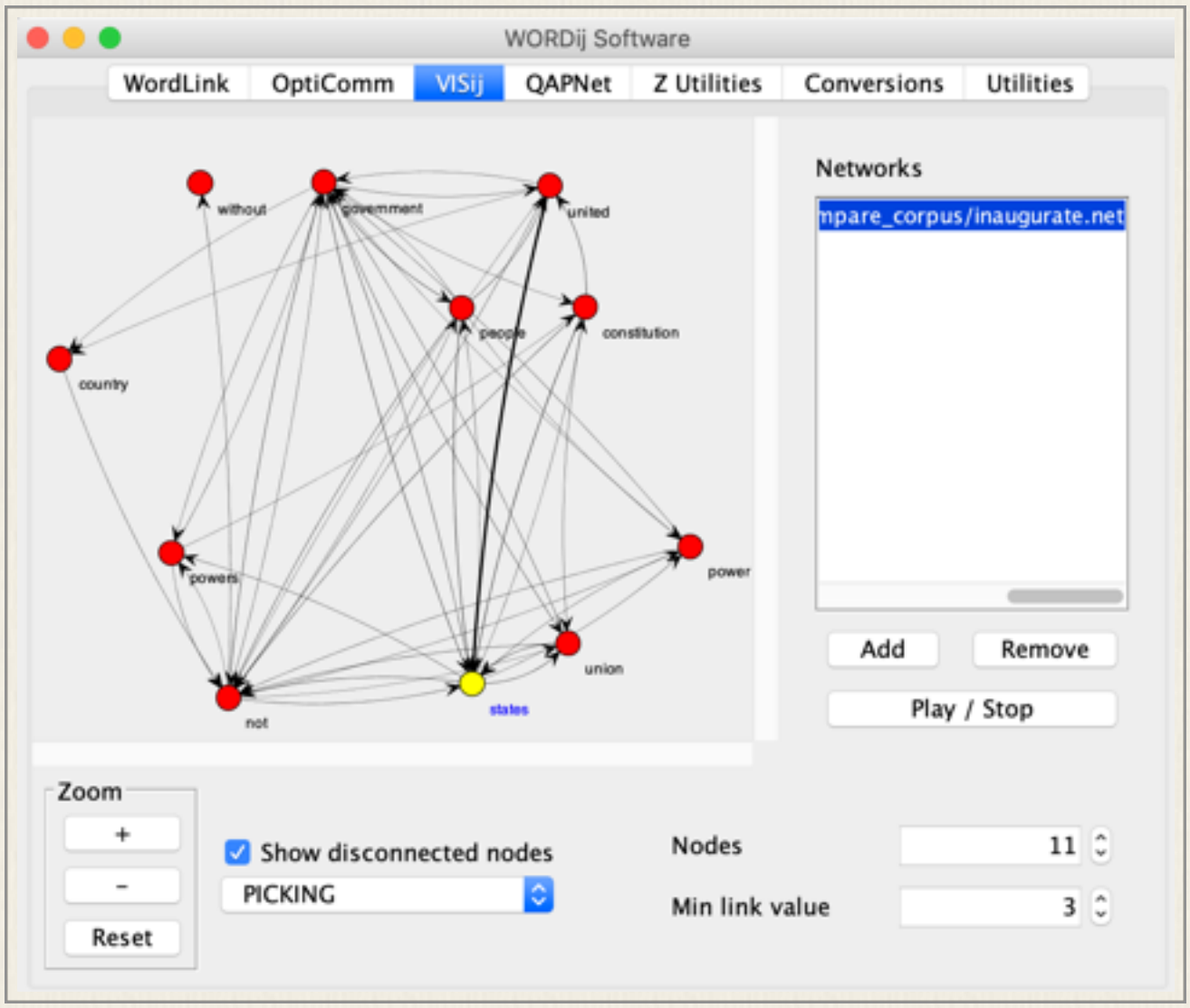
https://www.jasondavies.com/wordtree/



# โปรแกรมสร้าง network

## WORDij

เป็นโปรแกรมสำหรับอ่านตัวบทและวิเคราะห์ความสัมพันธ์ของคำภายในตัวบทนั้น สามารถดาวน์โหลดมาใช้เพื่อการศึกษาได้ฟรีที่ http://wordij.net/ หลักการเหมือนกับการหา word collocation เราสามารถกำหนดระยะห่างระหว่างคำและความถี่ขั้นต่ำ โปรแกรมจะสร้างไฟล์ output ย่อยๆ ทั้งที่เป็นไฟล์รายการคำเดี่ยว ไฟล์รายการคู่คำ (bigram) ไฟล์ที่คำนวณความสัมพันธ์โดยใช้ค่าสถิติต่าง ๆ และไฟล์ที่แสดงภาพเครือข่ายความสัมพันธ์ (pajek format) :ซึ่งสามารถดูได้จากในโปรแกรมนี้เองโดยดูผ่าน ViSij



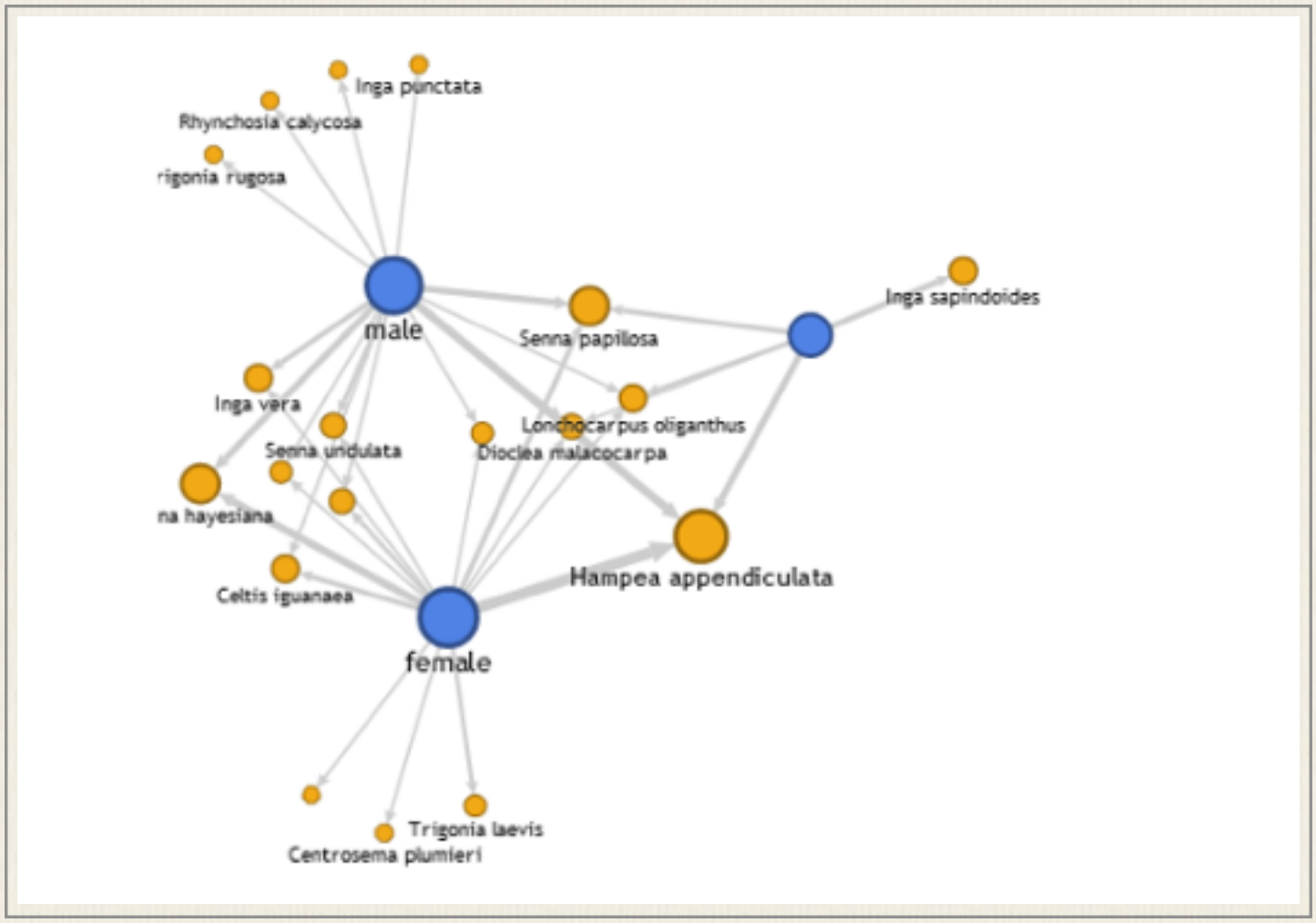
ไฟล์ข้อมูลที่ต้องการวิเคราะห์ต้องเป็นไฟล์เดียว  หากคลังข้อมูลประกอบด้วยไฟล์ย่อยๆ หลายไฟล์ให้ต่อไฟล์เป็นไฟล์เดียวด้วยคำสั่ง copy เช่น  copy *.txt NewfileName.txt  หรือ cat *.txt > NewfileName.txt

WordLink เป็นแท็บที่ใช้สร้างเครือข่ายความสัมพันธ์ของคำ เพียง upload ไฟล์ข้อมูล  และกำหนดไฟล์รายการคำยกเว้นหรือคำที่ไม่ต้องการนำมาวิเคราะห์  (สามารถใช้ไฟล์ droplist.txt ที่มาพร้อมกับโปรแกรมได้)  ระบุระยะห่างของคำและความถี่ขั้นต่ำ  ผลลัพธ์ที่ได้จะถูกเก็บเป็นไฟล์ชื่อเดียวกับไฟล์ข้อมูลแต่มีนามสกุลต่าง ๆ ออกมา  ไฟล์ที่เป็นนามสกุล .net  หมายถึง network ที่เก็นในรูปแบบ pajek ไฟล์  สามารถนำไปใช้กับโปรแกรม visualize network ต่างๆ เช่น VOSViewer, Gephi เป็นต้น

**Google Fusion Table**

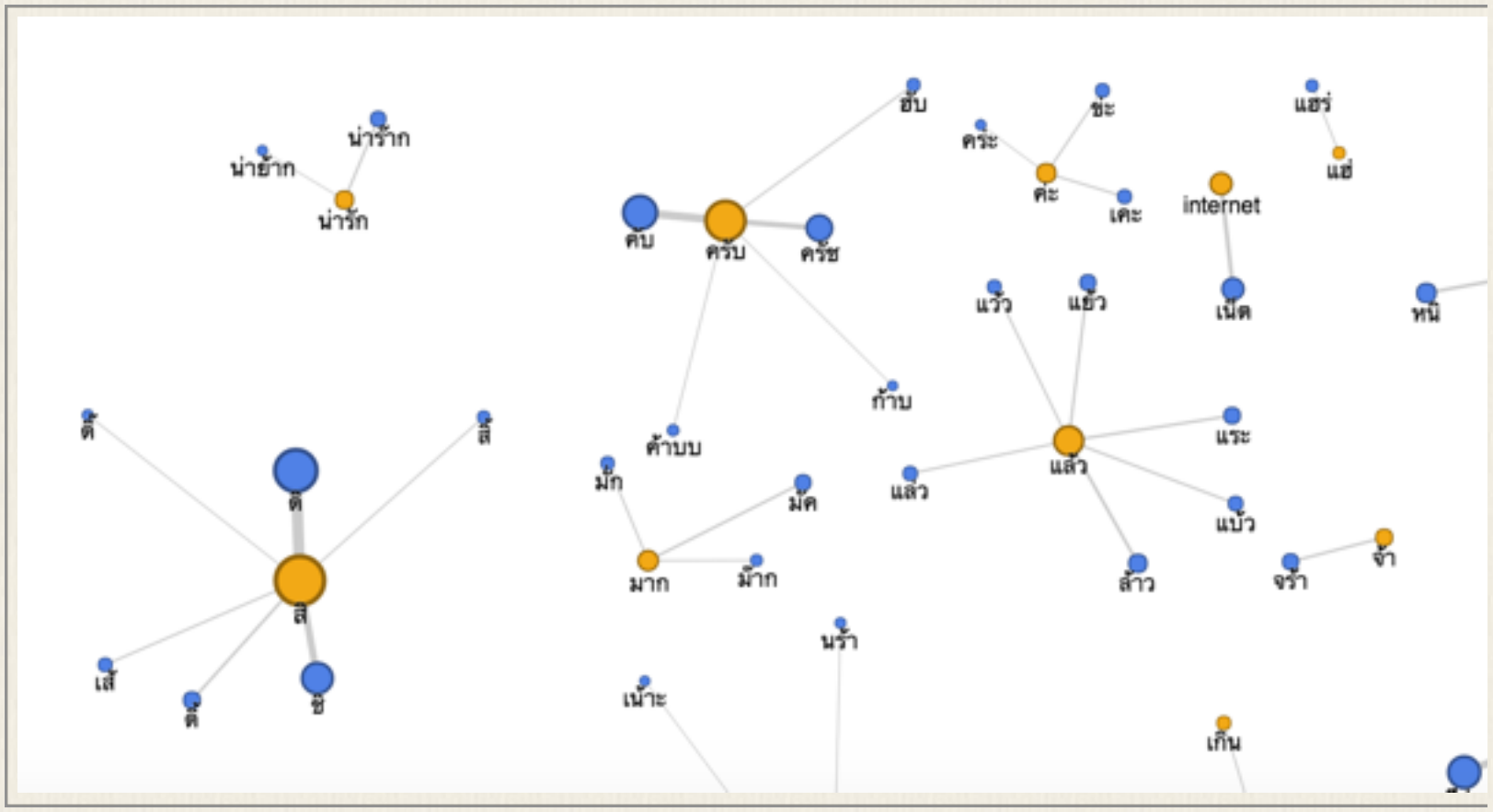
เป็นบริการของ Google ที่สร้างในการวิเคราะห์และแสดงภาพข้อมูลที่อยู่ในรูปตาราง  สามารถ upload ไฟล์ที่เป็น csv  สามารถใช้แสดงกราฟแบบต่าง ๆ รวมถึง network  และการแสดงข้อมูลบนแผนที่

https://sites.google.com/site/fusiontablestalks/home





ตัวอย่างข้างล่างแสดงคำที่แผลงจากรูปมาตรฐานแวเงในรูป network ตามขนาดความถี่ที่พบว่ามีการเขียนในรูปแบบนั้นเท่าใด




ข้อมูลที่แสดงมาจากไฟล์ csv ที่ upload เข้า Google Fusion Table

1-100 of 340

| Form | P+F | Intended |
|---|---|---|
| แฮร่ | 59 | แฮ่ |
| เห้อ | 486 | เฮ่อ |
| ฮรือ | 211 | ฮือ |
| ห้ะ | 153 | ฮ่ะ |
| ไอติม | 290 | ไอศครีม |
| ติม | 212 | ไอติม |
| โอ | 1947 | โอเค |
| เค | 2299 | โอเค |
| เคร | 382 | โอเค |
| โอเช | 29 | โอเค |
| แอด | 365 | แอดมิดชั่น |

# โปรแกรมสกัดข้อมูลจากเว็บ

### โปรแกรม Httrack

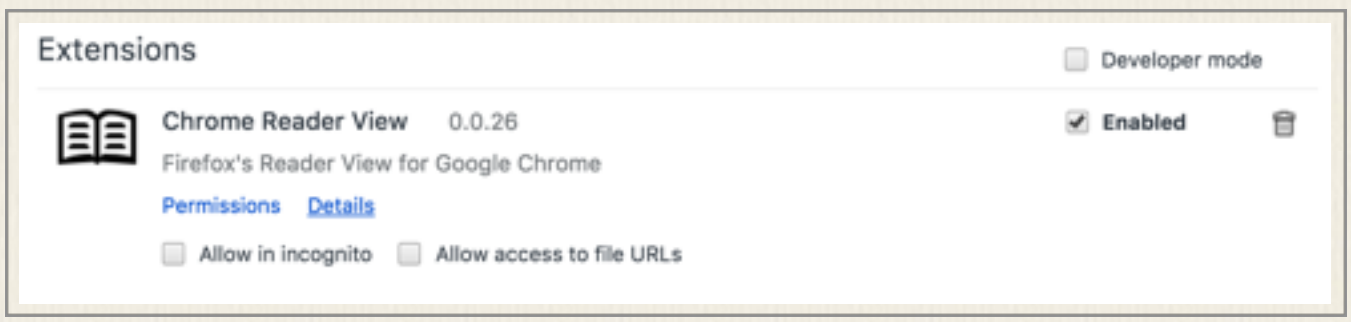
(https://www.httrack.com/) เป็นโปรแกรม copy website เหมาะสำหรับการคัดลอกข้อมูลบนเว็บมาเก็บไว้ในเครื่องเพื่อใช้งาน offline Httrack เป็นโปรแกรมสำหรับติดตั้งบน Windows สะดวกในการเก็บข้อมูลจากเว็บใดเว็บหนึ่ง

### ใช้บริการ boilerpipe

(https://boilerpipe-web.appspot.com/) เป็นบริการ web api ที่แสดงการใช้งาน boilerpipe Java library ในการดึงตัวบทจากเว็บที่ระบุ สามารถใช้ดึงข้อมูลตัวบทออกมาโดยทิ้งส่วนที่เป็นเมนูด้านข้าง เพื่อให้สามารถเห็นตัวบทและคัดลอกตัวบทได้ง่ายขึ้น

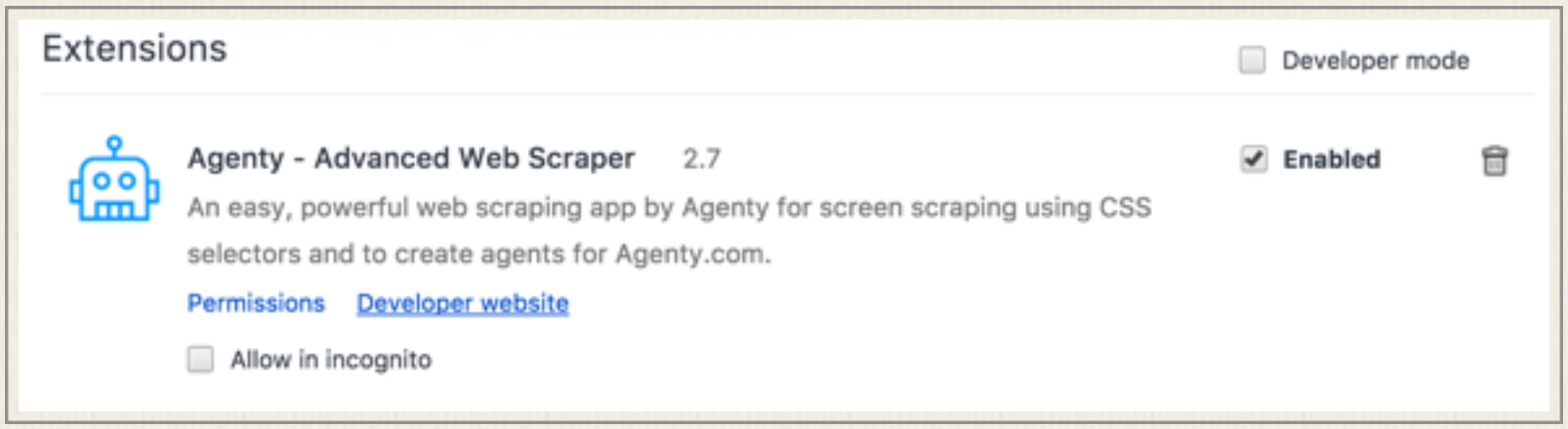
### ใช Chrome Extension

1. ใช้ Reader View addon เพื่อแปลงหน้าเว็บให้อยู่ในรูป text based เป็นหลักก่อน addon นี้ใช้เพื่อให้สะดวกต่อการอ่าน ซึ่งก็ทำให้สะดวกต่อการ copy-paste เฉพาะข้อความที่ต้องการออกมาได้ วิธีนี้เป็นการ copy หน้าเว็บเองเป็นหลัก




2. ติดตั้ง Advanced Web Scraper เมื่อพบเว็บที่มีข้อมูลที่ต้องการ ให้กดเรียก extension นี้ แล้วเลือกข้อมูลที่ต้องการโดยการกด * แล้วคลิกตำแหน่งข้อมูลที่ต้องการบนหน้าเว็บนั้น จะเห็นแถบสีระบายข้อความที่จะถูกสกัดออกมา จากนั้นให้ตั้งชื่อ field ที่แทนข้อมูลนั้น ทำเช่นนี้กับข้อมูลส่วน

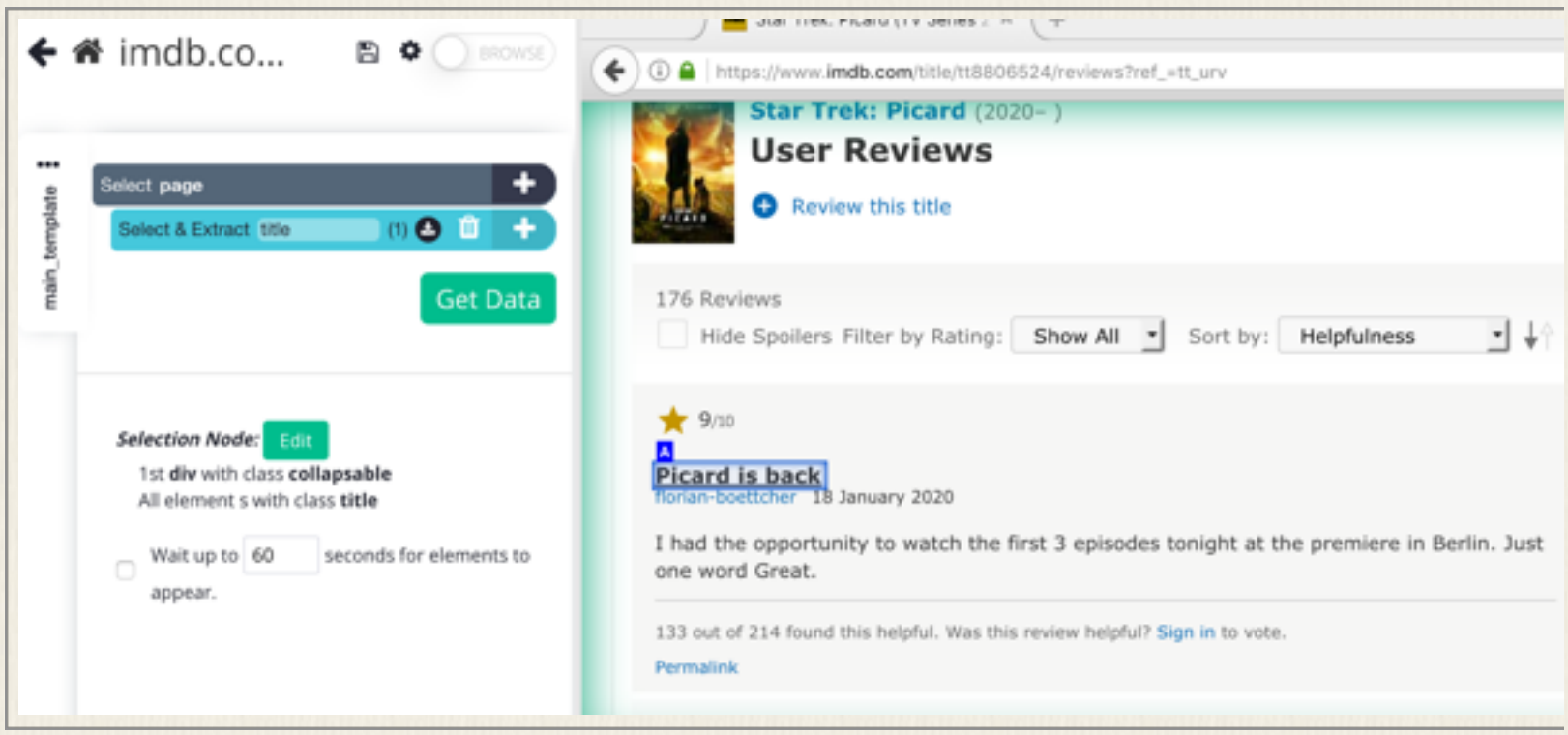
อื่นจนครบทุก filed ข้อมูลที่ต้องการ จากนั้นกด preview เพื่อดูข้อมูลที่จะ extract แล้วเลือก option ว่าจะบันทึกข้อมูลเป็น csv, json โปรแกรมนี้เหมาะสำหรับการดึงเว็บที่มีข้อมูลหลายรายการที่ดึงออกมาเป็นแต่ละ record ได้ เช่น หน้าเว็บวิจารณ์เรื่องต่างๆ ที่สามารถดึง review แต่ละคนออกมาเป็น 1 รายการได้ แต่ข้อมูลที่ดึงจะยังคงติด html tag อยู่บ้าง ต้องมาหาทางตัดออกเองภายหลังหากไม่ต้องการ




## โปรแกรม Parsehub

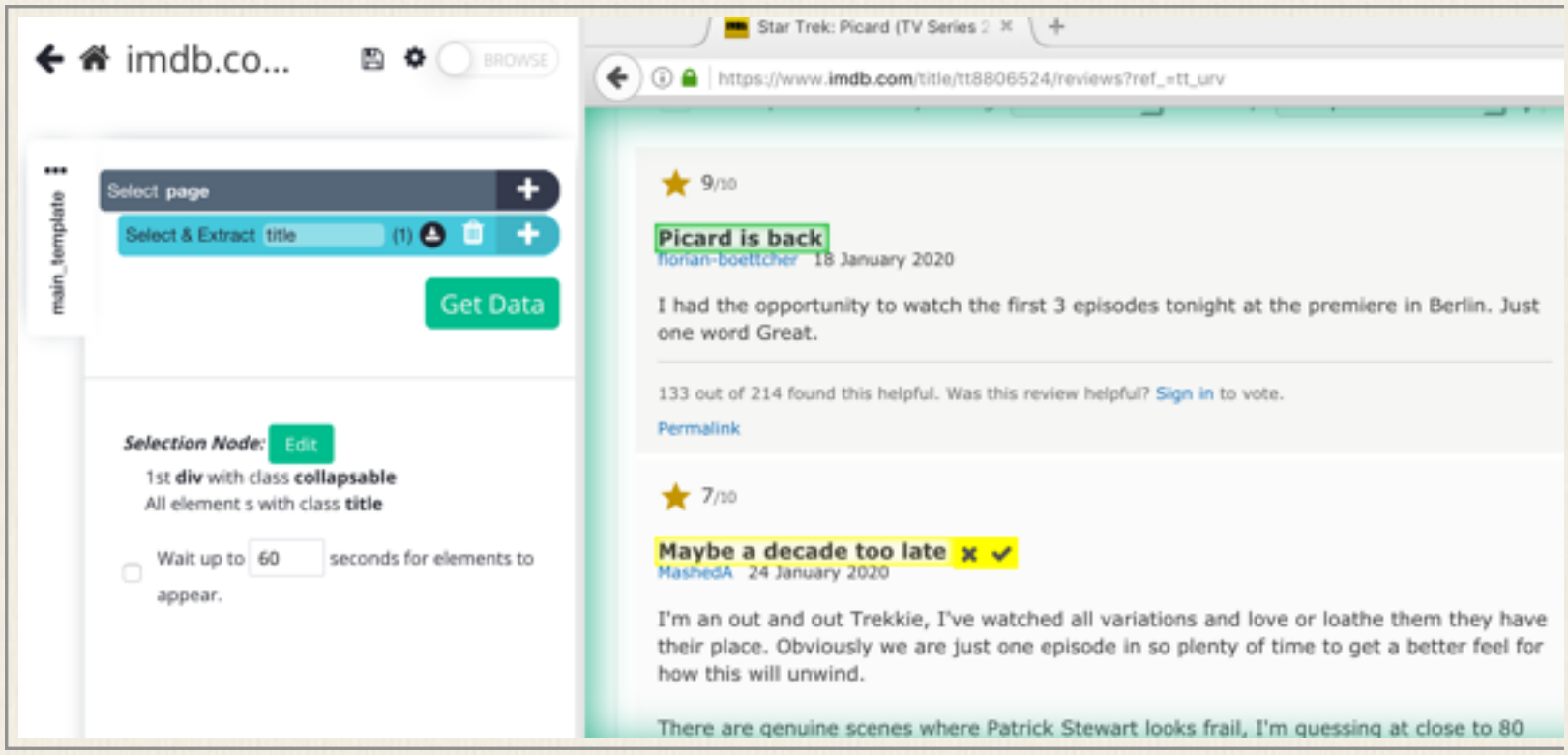
Parsehub เป็นอีกโปรแกรมหนึ่งที่มี UI ให้ออกแบบการสกัดข้อมูลจากหน้าเว็บได้ สามารถสกัดข้อมูลที่มีลักษณะเป็นรายการซ้ำ ๆ กัน หรือมีปรากฏในหลายหน้าได้ โปรแกรมมีทั้ง version ที่เป็น Windows, OS X, Linux สามารถดาวน์โหลดได้จาก https://parsehub.com/ version ที่ฟรีมีข้อจำกัดปริมาณของการสกัดข้อมูลว่าได้ไม่เกิน 200 หน้า หากต้องการใช้สกัดข้อมูลจำนวนมากเป็นประจำจะมีค่าใช้จ่ายรายเดือนที่ $149

ผู้ใช้ต้องเริ่มต้นด้วยการสำรวจหน้าเว็บที่ต้องการสกัดข้อมูลว่า ข้อมูลบริเวณที่ต้องการนั้นอยู่ที่ใด มีแท็ก html ใดในโครงสร้างที่ใช้ระบุตำแหน่งข้อความนั้นได้ไหม เมื่อตั้งโครงงานใหม่ ให้ใส่ url ของหน้าเว็บที่ต้องการลงไป จากนั้นเลือก action ต่าง ๆ เป็นลำดับขั้นตอน ดังตัวอย่างที่แสดงนี้

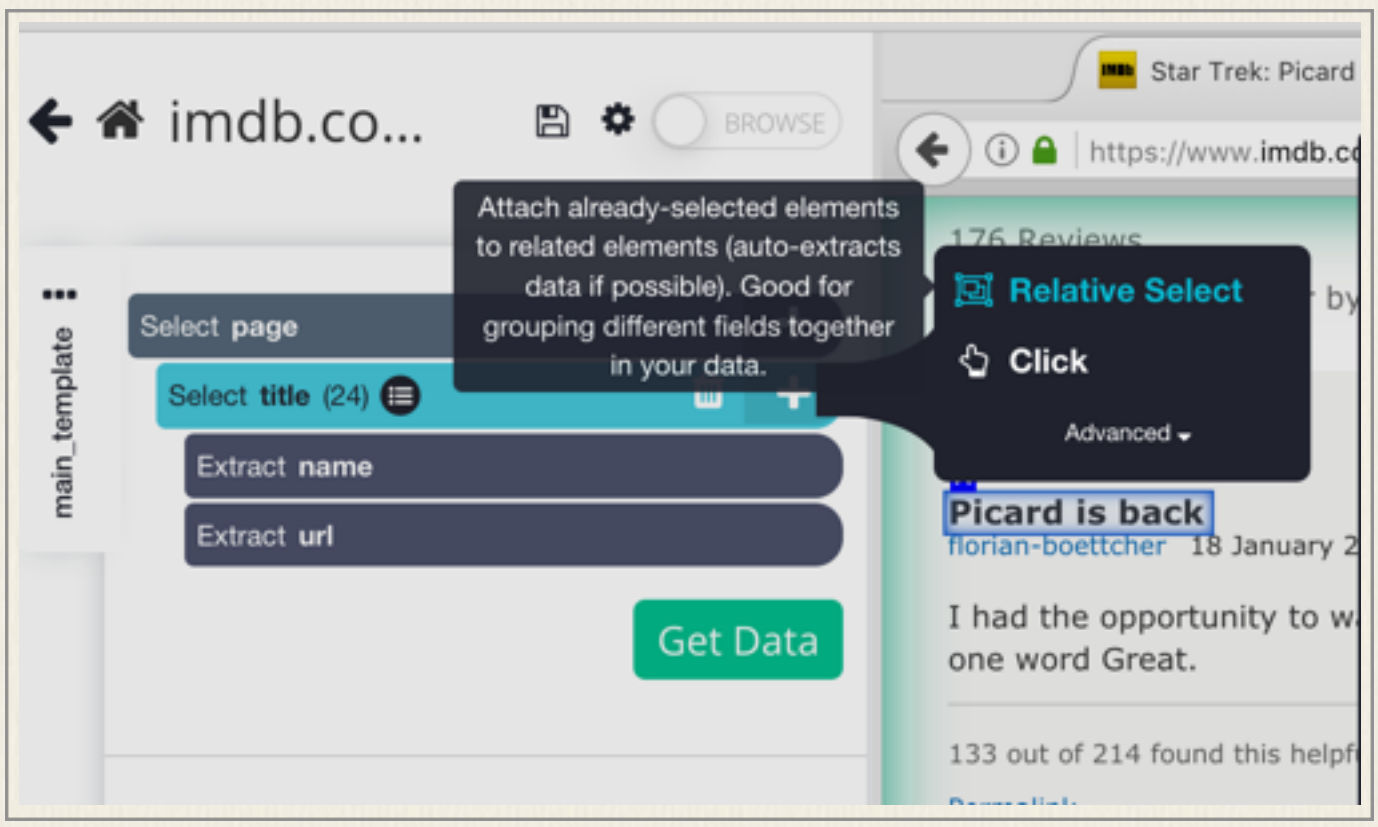



ขั้นแรกสุด Select Page จะขึ้นมาเอง เว็บตัวอย่างนี้เป็น review ภาพยนต์ซึ่งมีหลายคนมาเขียนวิจารณ์ เราต้องการชื่อบทวิจารณ์ที่ตั้ง (เช่น Picard is back) ตัวบทวิจาณ์ และคะแนนที่ให้ให้เลือก Select และเลือกตัวชื่อบทวิจารณ์อันนึงก่อน จะเห็นเป็นแถบสี และเนื่องจากข้อมูลมีลักษณะเป็นรายการหลาย ๆ อัน บทวิจารณ์อันอื่น ๆ จะเห็นเป็นสีเหลืองให้กดที่ชื่อบทวิจารณ์อีกอันนึงเพื่อบอกโปรแกรมให้รู้ว่าเราต้องการสกัดข้อมูลที่มีหลายรายการ จะสังเกตให้จำนวนตัวเลข

ที่เพิ่มขึ้นมากกว่าหนึ่ง  สำรวจว่าข้อมูลชื่อคำวิจารณ์ได้ครบทั้งหน้า  และจั้งชื่อ field ตามต้องการ เช่น “title”

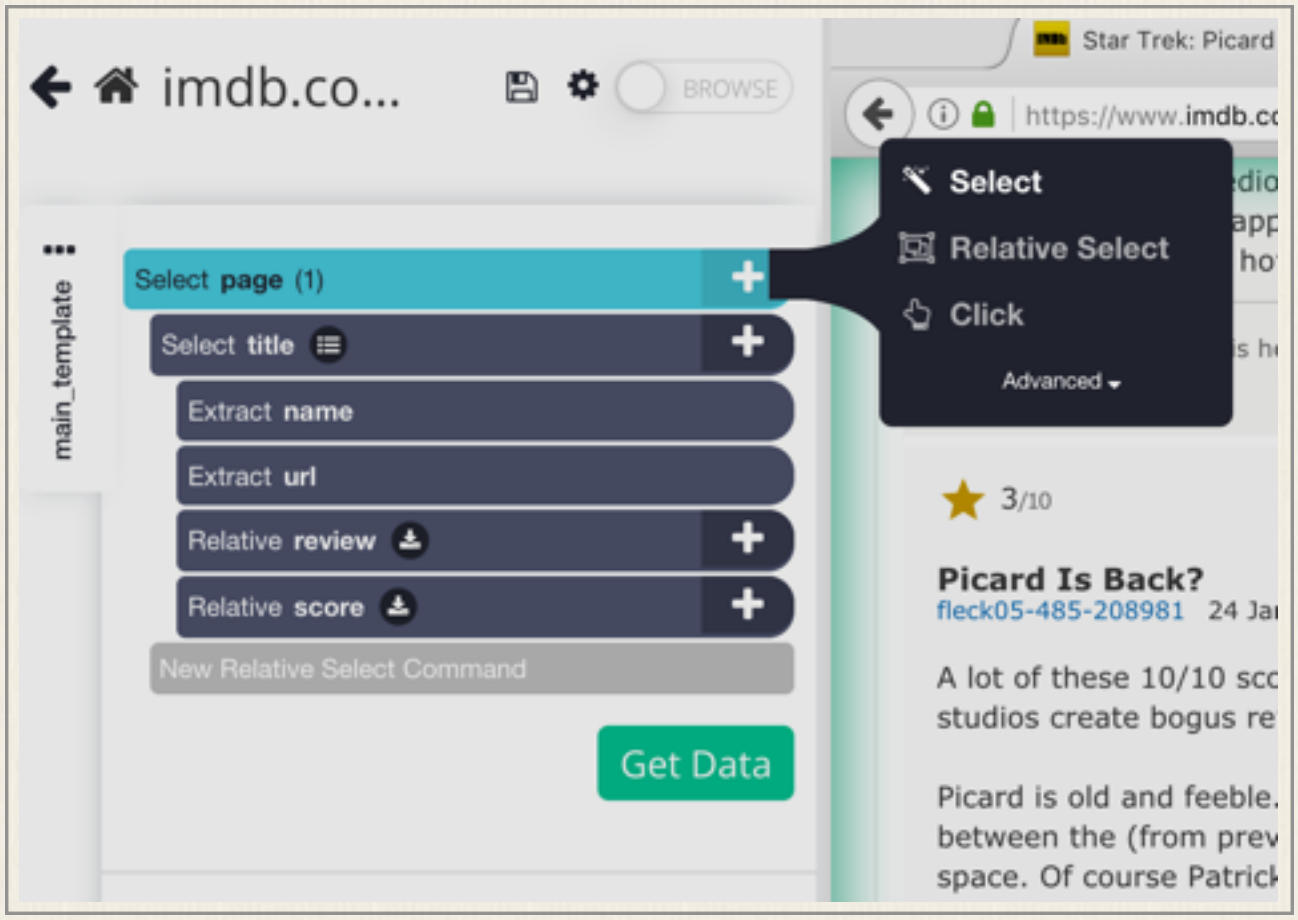


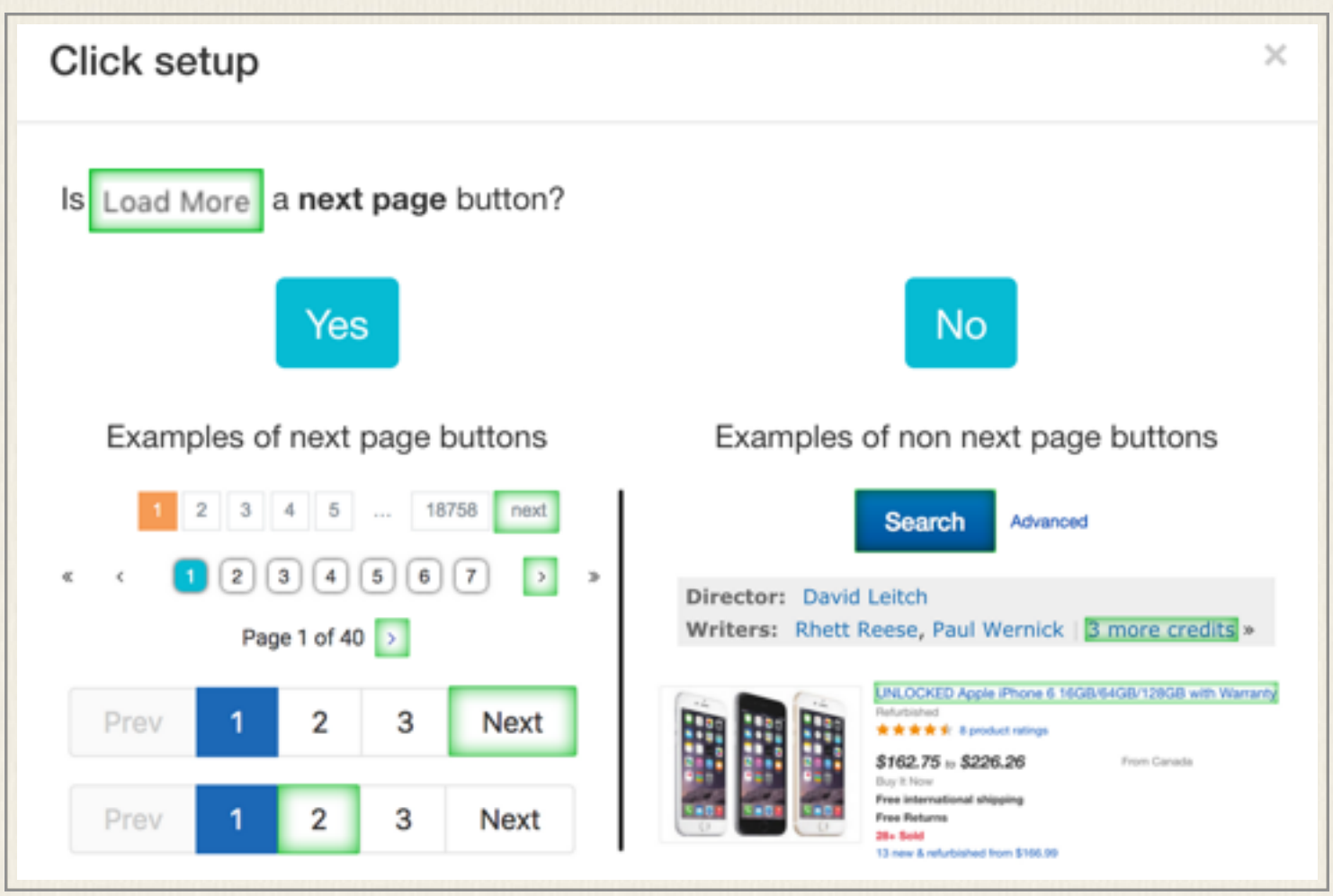

จากนั้นจึงเลือกข้อมูลชิ้นต่อไปที่สัมพันธ์กับชื่อ title กรณีนี้ให้เลือก Relative Select เพื่อโยงชื่อบทวิจารณ์นั้นกับชื่อ title โดยกดเลือกที่ชื่อ title ก่อนแล้วลากลูกศรไปที่บทวิจารณ์  ทำเช่นนี้กับส่วนที่เป็นคะแนนด้วยเพื่อเก็บคะแนนที่ผู้เขียนแต่ละคนให้  พร้อมทั้งตั้งชื่อ field แต่ละอัน เช่น review, score

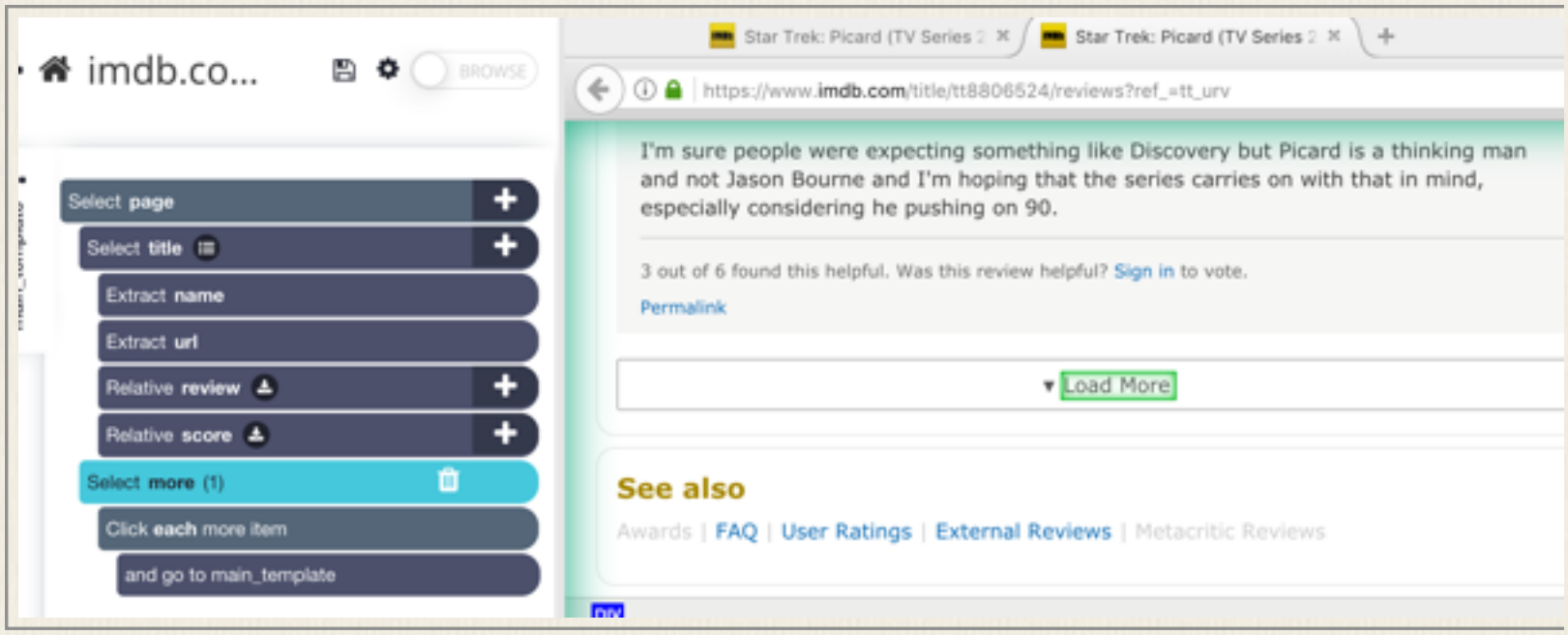




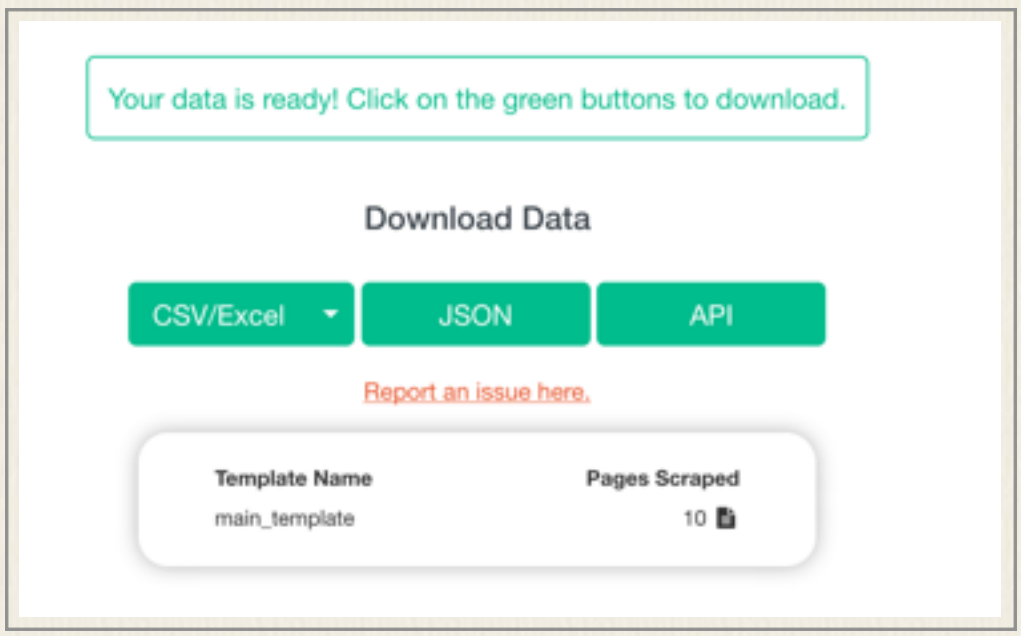
ข้อมูลที่สกัดได้ของแต่ละบทวิจารณ์ก็จะมีทั้ง title, review และ score โดยจะสกัดทั้งหมดเป็นข้อมูลรายการที่มี field ทั้งสามนี้ แต่ในหน้าเพจนี้จะเห็นว่ามีปุ่มให้กด Load More ซึ่งเป็นการเรียกหน้าถัดไปขึ้นมา กรณีนี้ให้เลือก Select อีก item เพิ่มและให้เลือกเป็น Click และจึงบอกว่า Click นั้นเป็นการ click แบบดึงหน้าถัดไป แล้วเลือกให้กลับไปใช้ main template ให้ทำซ้ำ ก็จะเป็นการวนรอบทำงานได้


imdb.co...
BROWSE
main_template
Select page (1)
Select title
Extract name
Extract url
Relative review
Relative score
New Relative Select Command
Get Data
Star Trek: Picard
Select
Relative Select
Click
Advanced
3/10
Picard Is Back?
fleck05-485-208981
A lot of these 10/10 scc
studios create bogus re
Picard is old and feeble.
between the (from prev
space. Of course Patrick


Click setup
Is Load More a next page button?
Yes
No
Examples of next page buttons
Examples of non next page buttons
1 2 3 4 5 ... 18758 next
« ‹ 1 2 3 4 5 6 7 ›
Page 1 of 40 ›
Prev 1 2 3 Next
Prev 1 2 3 Next
Search
Advanced
Director: David Leitch
Writers: Rhett Reese, Paul Wernick | 3 more credits »
UNLOCKED Apple iPhone 6 16GB/64GB/128GB with Warranty
Refurbished
8 product ratings
$162.75 to $226.26
Buy It Now
Free international shipping
Free Returns
28+ Sold
13 new & refurbished from $166.99
From Canada




ภาพขั้นตอนการสกัดข้อมูลทั้งหมดจะเห็นเป็นขั้นตอนคำสั่งตามตัวอย่างที่เห็น จากนั้นจึงสั่ง RUN เพื่อให้โปรแกรมสกัดข้อมูลตามที่ตั้งค่าไว้นี้ ผลลัพธ์ที่ได้สามารถเก็บเป็น CSV เปิดใน Excel หรือเป็น JSON ก็ได้



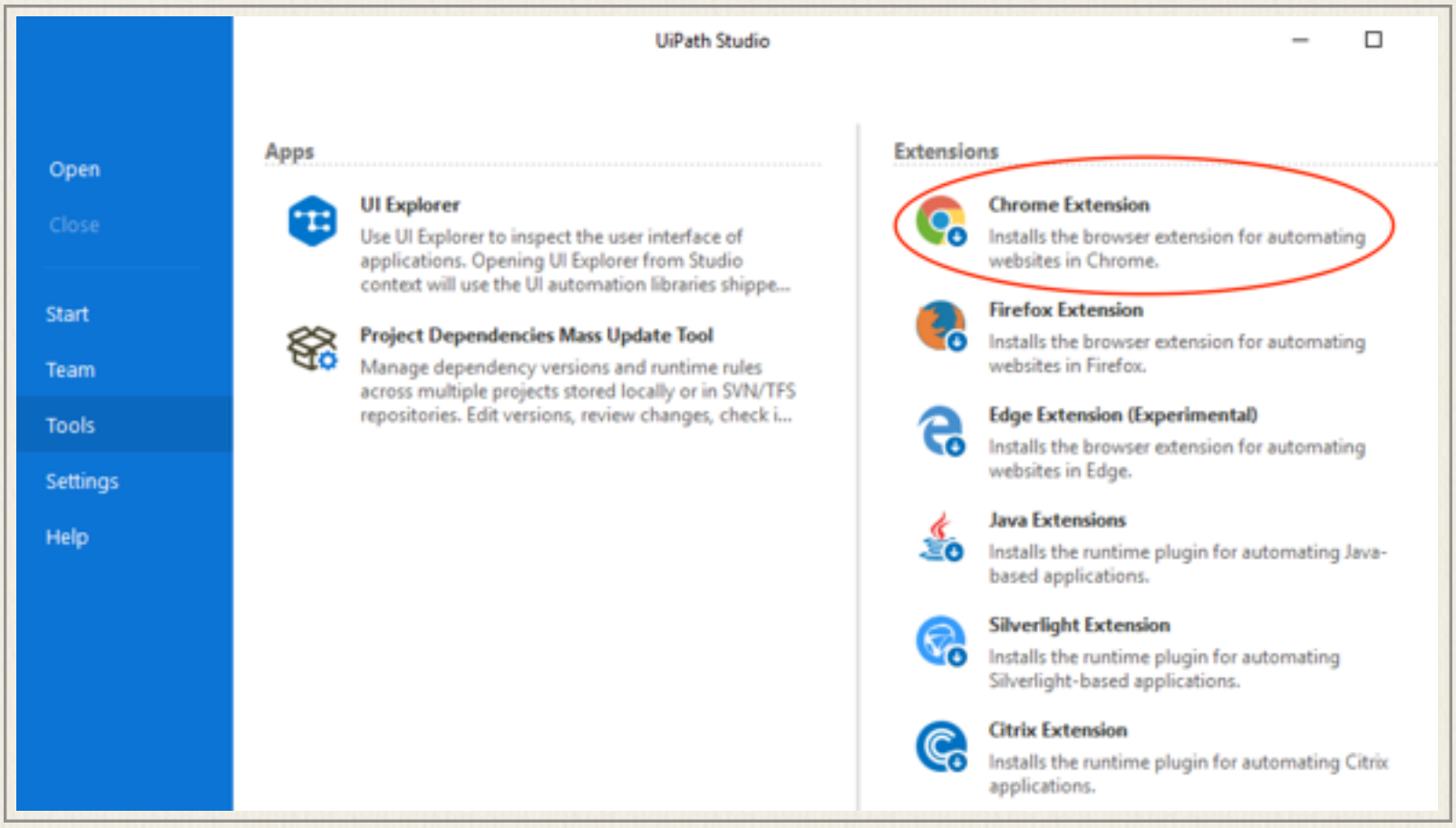
| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | title_name | title_url | title_review | title_score | |
| 2 | Picard is back | https://www.imdb.com/re | I had the opportunity to watch the first 3 episodes tonight at the premiere in Berlin. Just o | 9 | |
| 3 | Overly exceeded expectations! | https://www.imdb.com/re | I just finished watching the first episode world premiere (01/23/20 -12:00 a.m.) and I | 10 | |
| 4 | Picard is back! | https://www.imdb.com/re | Very minor spoilers in this review for the most part, but before I start my | 10 | |
| 5 | Maybe a decade too late | https://www.imdb.com/re | I'm an out and out Trekkie, I've watched all variations and love or loathe them they have | 7 | |
| 6 | Off to a very intriguing start. M | https://www.imdb.com/re | Forget anything negative you may have heard. This is very, very good. And it's definitely | 8 | |
| 7 | Make it so..... | https://www.imdb.com/re | I grew up watching Next Generation and was apprehensive about Picard. I was so wrong. | 10 | |
| 8 | Eh. Better than Discovery. | https://www.imdb.com/re | So far, eh. | 6 | |
| 9 | Great new beginning for Picar | https://www.imdb.com/re | Someone wrote "what were they thinking" because Patrick Stewart will be 80 this year. W | 9 | |
| 10 | This is so good! Finally they go | https://www.imdb.com/re | I'm giving this show a 10/10. From the start on you now that they put in a lot of work! Aft | 10 | |
| 11 | Satisfying, But It's Not Your St | https://www.imdb.com/re | Long-form storytelling isn't new to Star Trek but many of the most enduring & memorable | 9 | |
| 12 | Picard isn't about... Picard. | https://www.imdb.com/re | Of course in 2020 Picard can't be about Picard. Instead it has to be about a strong, man-be | 3 | |
| 13 | Picard Is Back? | https://www.imdb.com/re | A lot of these 10/10 scores share the same phrases, like: Picard is back! I've wondered if | 3 | |
| 14 | A good start | https://www.imdb.com/re | Ok so its not standard star trek. But times have moved on and 26 episode seasons just | 8 | |
| 15 | Jury is still out / On the fence | https://www.imdb.com/re | First episode in. I'm a massive TNG fan (as so many are) and my main concern is they're | 7 | |
| 16 | Breaks my heart to give such a | https://www.imdb.com/re | There are many moment in life when you realising that something has changed. Maybe | 3 | |
| 17 | Loving it. | https://www.imdb.com/re | Very pleased with what I have seen so far please continue. | 9 | |
| 18 | This has potential | https://www.imdb.com/re | I'm a huge old Trek fan. TNG will always be the greatest, followed by DS9. I actually | 7 | |
| 19 | What a let down | https://www.imdb.com/re | Star Trek: The Next Generation was successful in being true to T.O.S. both in format | 4 | |
| 20 | These 10/10 Reviews are not ( | https://www.imdb.com/re | I think they have people writing these reviews, they are all worded basically the same. | 3 | |
| 21 | Too old, too jaded | https://www.imdb.com/re | I have been a life long trek fan, my collection spands 50 years, I have perspective. | 5 | |
| 22 | In name only | https://www.imdb.com/re | Same old story they use a great Iconic character only to focus on all the women in the sho | 1 | |
| 23 | Why oh why | https://www.imdb.com/re | Don't know why people are reviewing this when it hasn't been aired and writting load of r | 10 | |
| 24 | TNG is dead and so is any hop | https://www.imdb.com/re | First of all if you are looking for a follow up to TNG ... it is not here . | 5 | |
| 25 | Left me flat and not bothered | https://www.imdb.com/re | The filming looks good. The design sharp. The dreams were slightly intriguing. The charact | 6 | |
| 26 | Picard is back | https://www.imdb.com/re | I had the opportunity to watch the first 3 episodes tonight at the premiere in Berlin. Just o | 9 | |
| 27 | Overly exceeded expectations! | https://www.imdb.com/re | I just finished watching the first episode world premiere (01/23/20 -12:00 a.m.) and I | 10 | |
| 28 | Picard is back! | https://www.imdb.com/re | Very minor spoilers in this review for the most part, but before I start my | 10 | |

**โปรแกรม UiPath**

โปรแกรม UiPath เป็นโปรแกรมกลุ่ม RPA (robotic process automation) คือเป็นโปรแกรมที่ช่วยในการสร้าง bot สำหรับมาทำงานที่ซ้ำ ๆ เราเพียงแต่กำหนดลำดับขั้นตอนว่าต้องทำอะไรก่อนหลังอย่างไร การสกัดข้อมูลจากเว็บจึงเป็นงานหนึ่งที่สามารถใช้ UiPath นี้ได้ โปรแกรม UiPath ที่เป็น community edition สามารถดาวน์โหลดมาใช้งานได้ (https://www.uipath.com/developers/community-edition) โปรแกรมทำงานบน Windows

หลังติดตั้งโปรแกรมแล้ว ให้เรียก UiPath Studio ที่ใช้ในการออกแบบ process automation ในการสกัดข้อมูลจากเว็บ ตัวอย่างจากการสกัดข้อมูลบทวิจารณ์ภาพยนตร์ของเว็บ IMDB ให้เลือกหนังที่ต้องการ ไปที่หน้าที่มีบท review เรียงเป็นรายการ เช่น https://www.imdb.com/title/tt0073486/reviews?ref_=tt_ov_rt

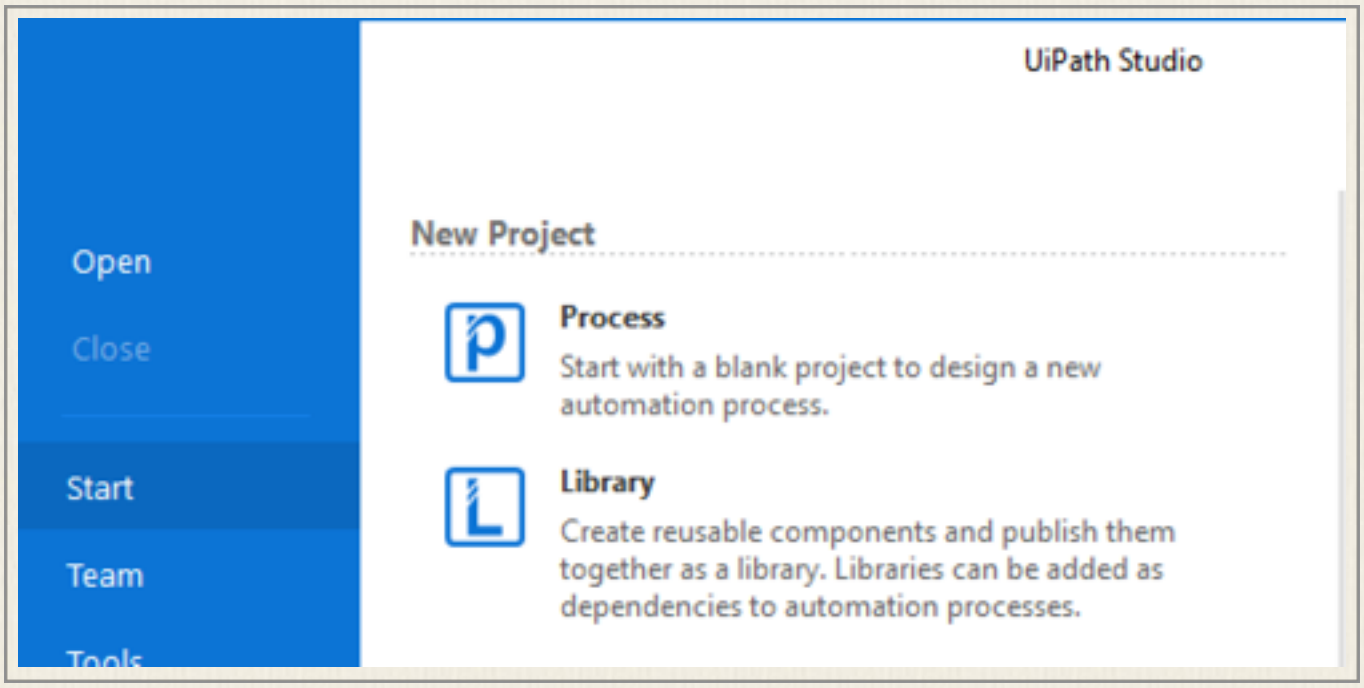
แต่ก่อนจะใช้งานโปรแกรม ให้ติดตั้ง extension ของ browser ที่จะใช้ก่อน เพราะ UiPathมีโมดูลที่ช่วย scrap data จากหน้าเว็บ จึงต้องติดตั้ง extension ของ UiPath ก่อน



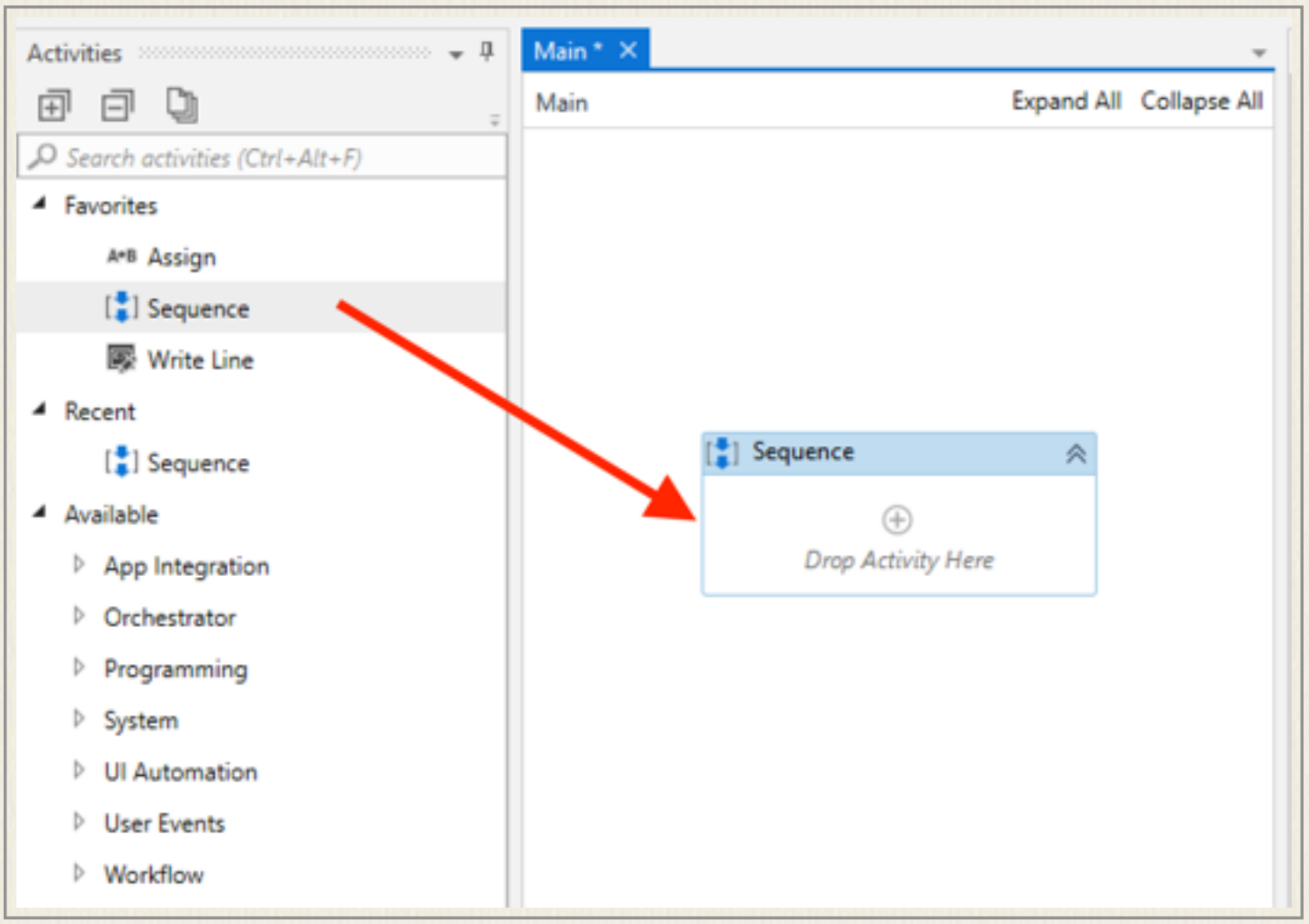

เมื่อติดตั้ง extension เรียบร้อยแล้วให้ enable extension นี้ใน browser ก่อนจะใช้งาน

โปรแกรม UiPath Studio ช่วยในการออกแบบ automation ให้เลือกสร้าง new process และตั้ง

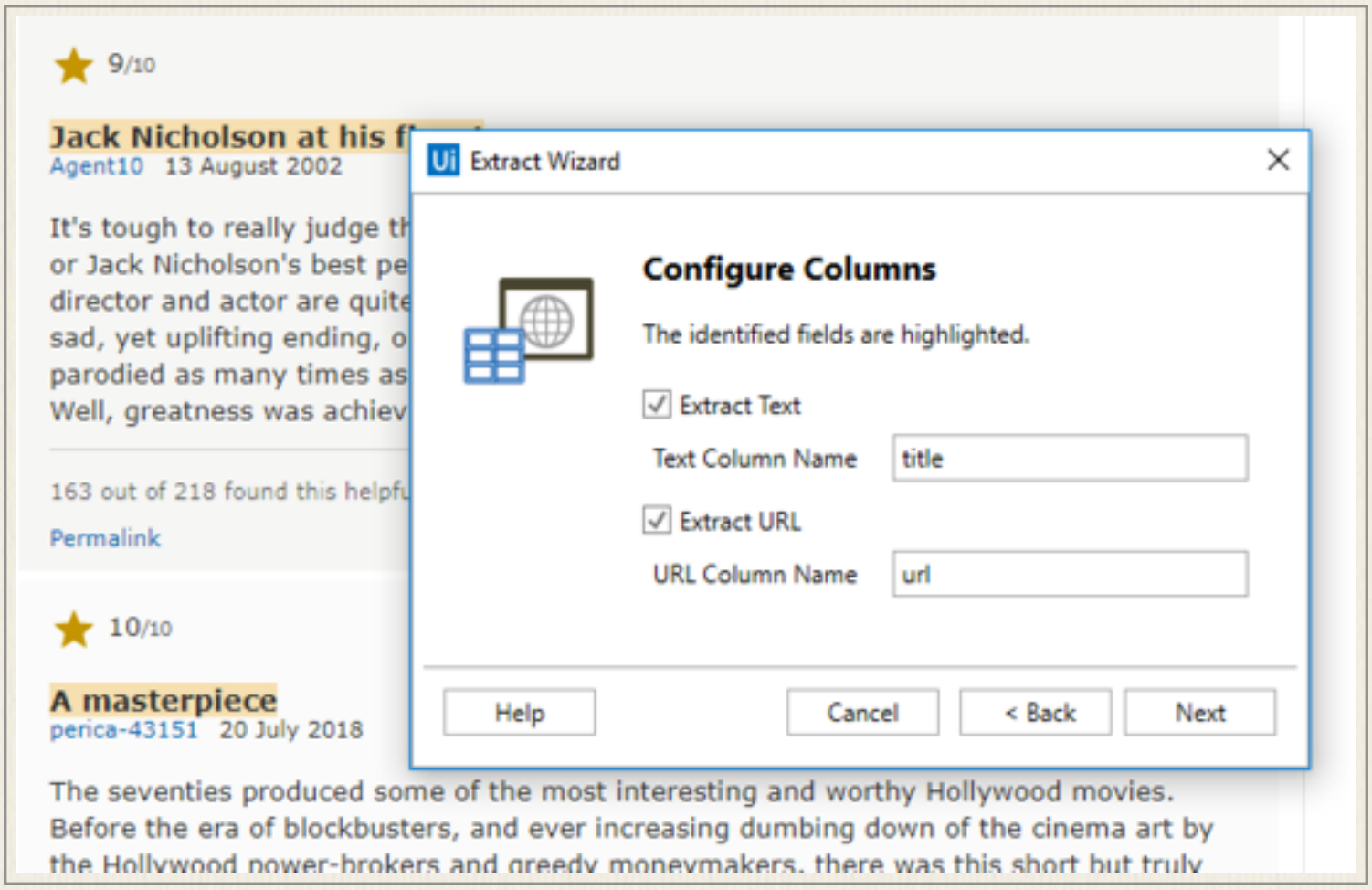
ชื่อ project ใหม่นี้  แต่หากการติดตั้ง extension มีปัญหาหรือไม่สามารถเลือกข้อมูลหน้าเว็บได้ ก็ให้ใช้ Internet Explorer แทนโดยไม่ต้องใช้ extension



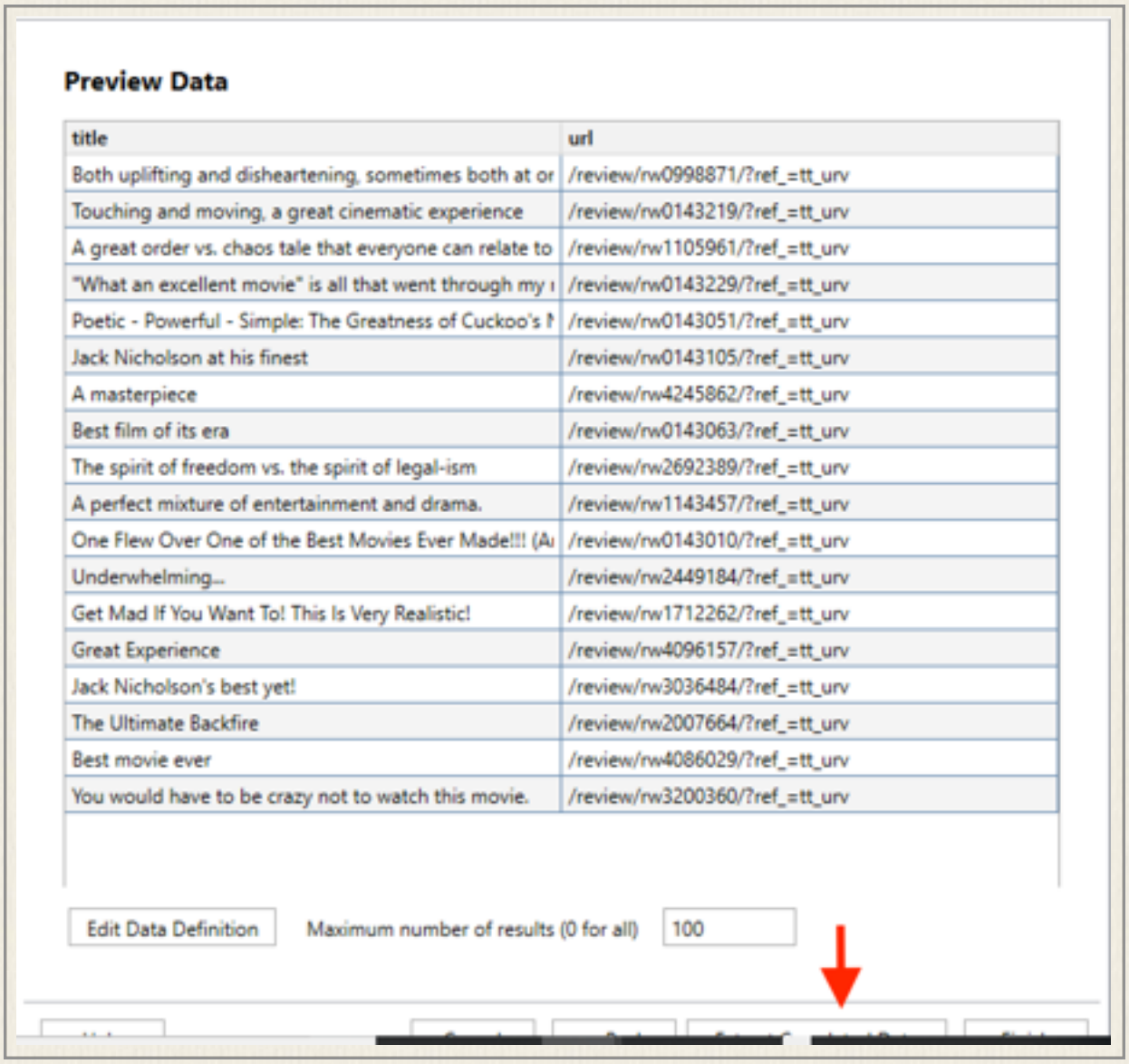

เลือก New - Sequence  จะเห็นกล่อง operation ที่เราสามารถ add process ต่างๆ เข้าไปได้ในกล่องนั้น  ทางซ้ายเป็นตัวเลือกเครื่องมือต่าง ๆ ที่สามารถลากมาใช้ได้  (อยู่ภายใต้หมวด activity) ให้เลือก Sequence เข้ามาเพื่อสร้างกล่องลำดับกระบวนการ  จากนั้นจึงเลือก Data Scraping เพื่อใช้สกัดข้อมูลหน้าเว็บ



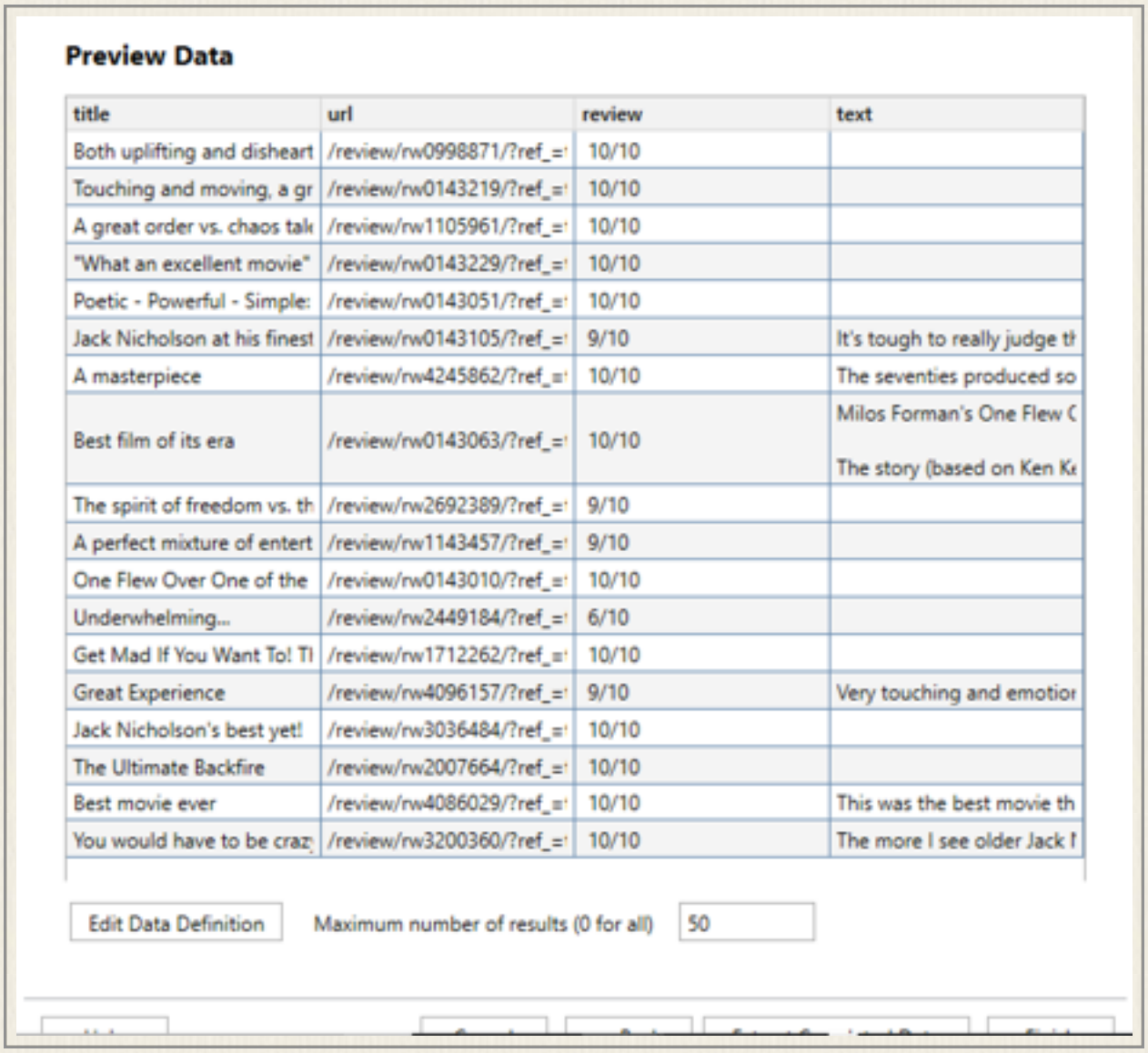
เมื่อเลือก Data Scraping แล้วจะมีหน้าจอถามให้เราเลือก element ที่ต้องการในหน้าเว็บ ข้อมูลที่มีลักษณะเป็นรายการซ้ำๆ กันนี้ เราจะต้องเลือก second element ของข้อมูลประเภทเดียวกันด้วย จะเห็นว่าข้อมูลส่วนที่ถูกเลือกจะมีระบายสีให้เห็น UiPath จะวิเคราะห์ดูว่าสิ่งที่ต้องการสกัดคืออะไร ให้เราตั้งชื่อ field ข้อมูล ในที่นี้ใช้ title กับ url เพื่อสกัดเอา link ของบทวิจารณ์นั้นมาด้วย กรณีที่เราบังเอิญเลือกข้อมูลสองตัวใน field นั้นไม่ตรงกัน โปรแกรมจะฟ้องว่าสกัดไม่ได้ จึงต้องดูให้ชัดว่าส่วนที่ระบายสีทั้งของ first และ second element เป็นข้อมูล field เดียวกัน



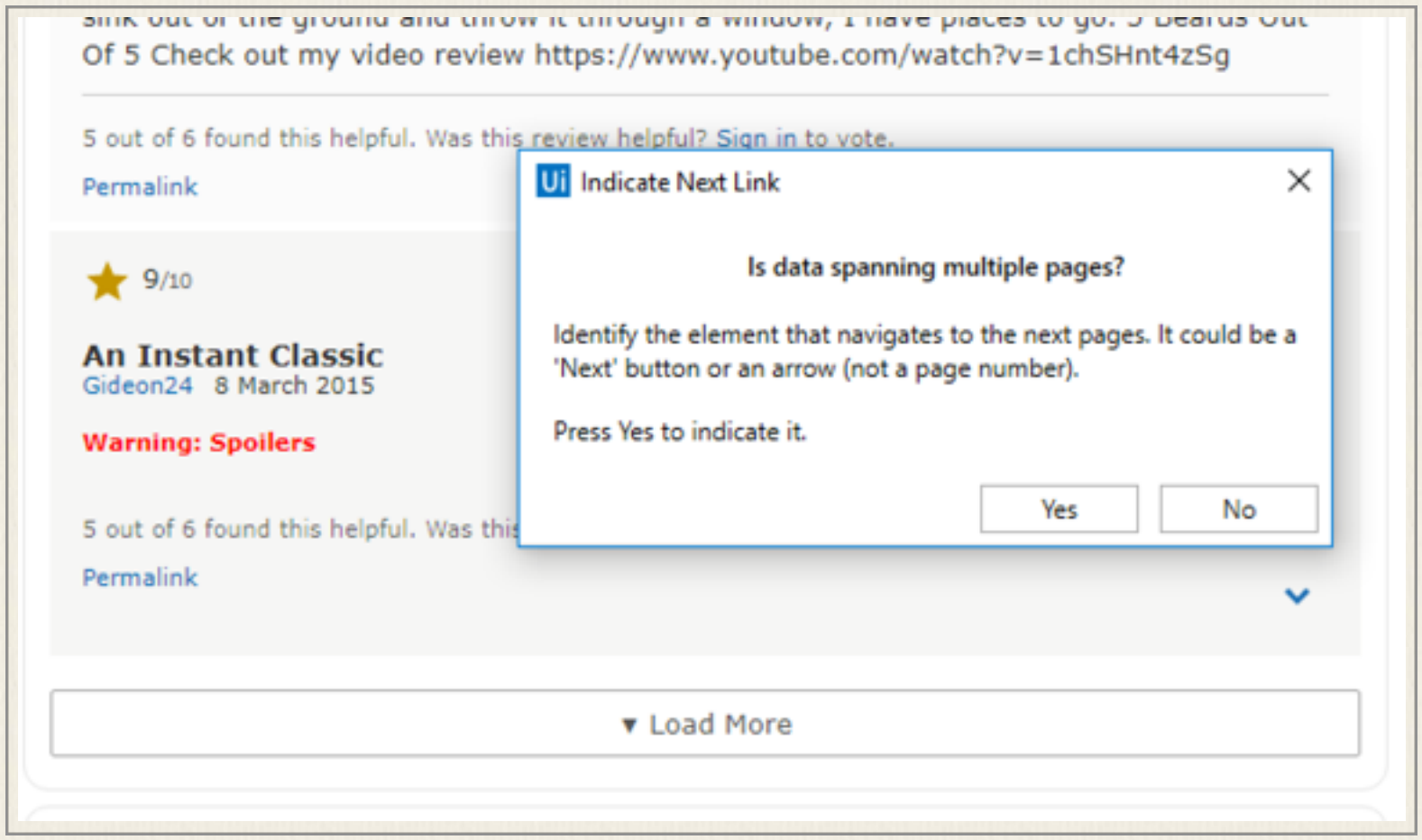

หลังจากเลือกข้อมูลชุดแรกแล้ว ถ้ายังมีข้อมูลอื่น ๆ ในรายการที่ต้องการ เช่น ในที่นี้ต้องการคะแนน review และตัวบทของการ review ด้วย ก็ต้องทำซ้ำกระบวนการนี้อีกสองหน โดยในหน้า Preview data ให้เลือกว่ายังมี related data อยู่อีก นอกจากนี้จำนวนรายการข้อมูลที่ต้องการสกัดก็สามารถระบุได้ ค่า default ตั้งไว้ที่ 100 สามารถเพิ่มหรือลดจำนวนตามความต้องการได้



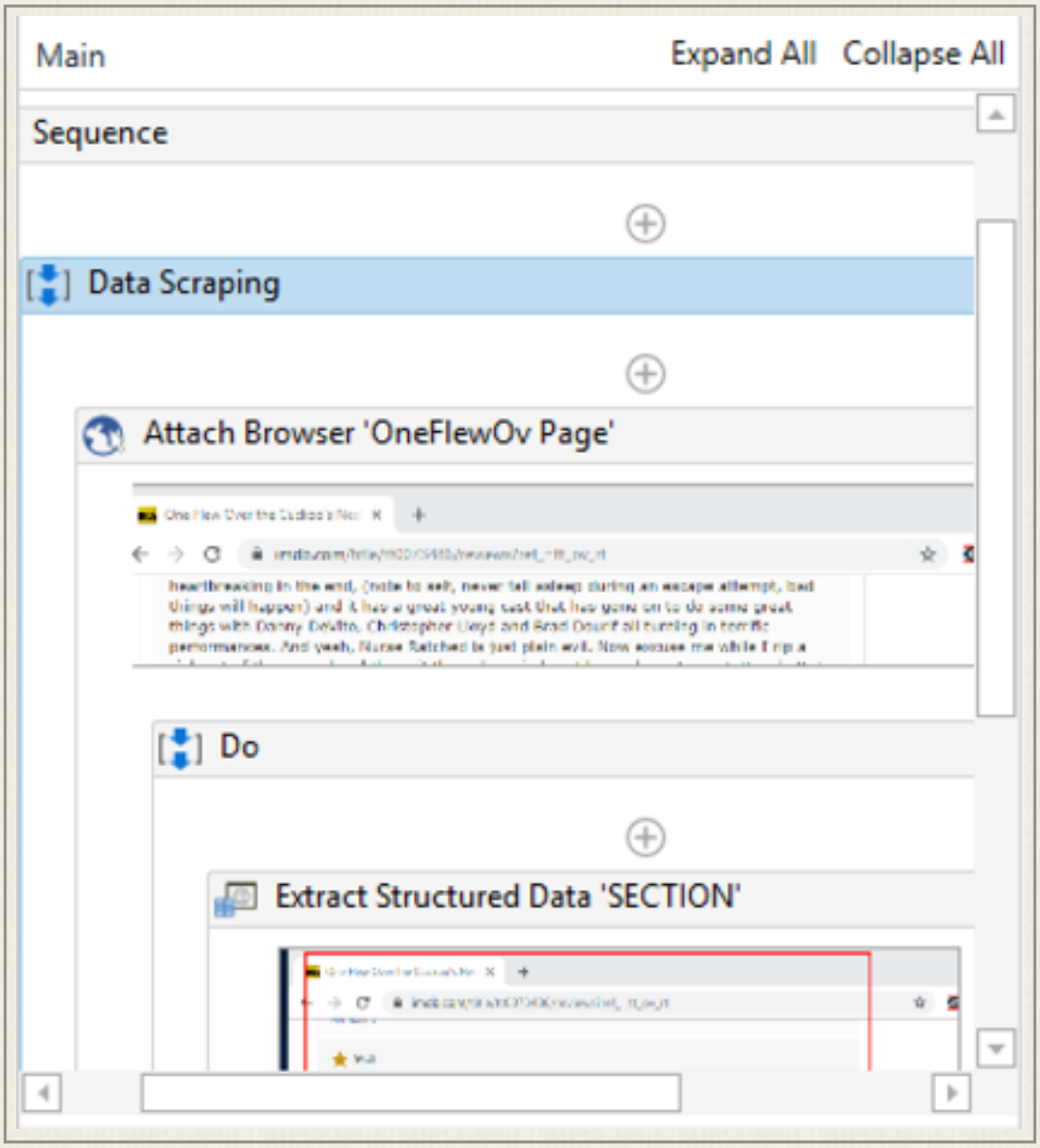

ทำการเลือกข้อมูลที่เกี่ยวข้องในรายการนั้นจนครบที่ต้องการ  จึงกด Finish ในหน้า Preview Data



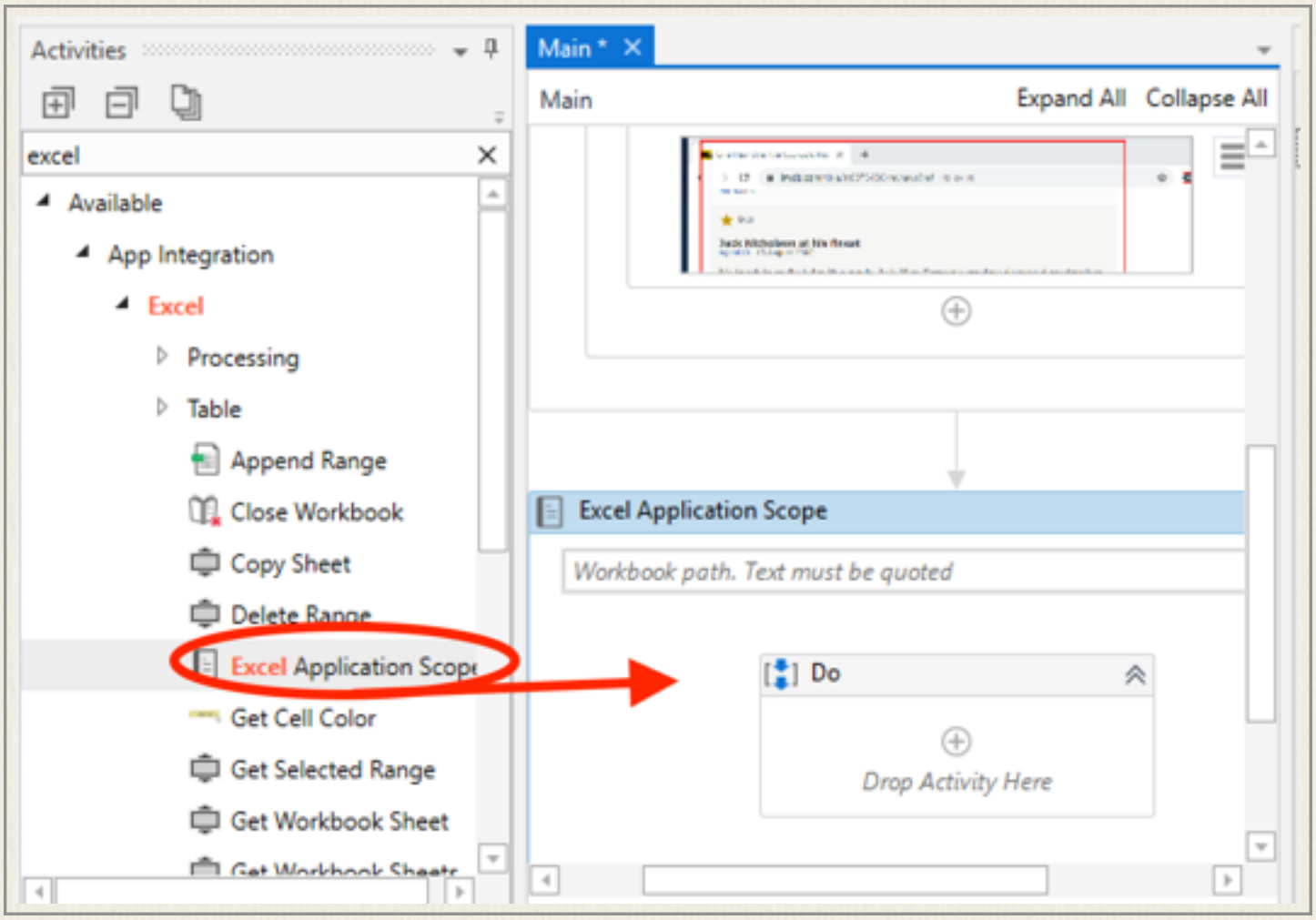

หลังจากนั้น จะมีหน้าจอถามว่าหน้าเว็บนี้เป็นประเภทที่มีหลาย ๆ หน้าไหม คือ ยังมีให้กด Next Page หรือ More อีกไหม  ถ้ามีก็จะไปกดหน้าถัดไปและไปสกัดข้อมูลที่ต้องการจนได้จำนวนรายการครบตามที่ตั้งค่าไว้  แต่ก่อนจะกดเลือกขั้นนี้ ควรกลับไปที่หน้าเว็บแล้วเลื่อนให้เห็นปุ่มสำหรับกด Next Page ก่อน




เมื่อกด Yes แล้ว ให้ไปที่หน้า Browser เพื่อบอกว่าปุ่ม Next Page หรือ More อยู่ที่ไหน  เป็นอันเสร็จสิ้นกระบวนการสกัดข้อมูล  เราจะเห็น process ที่กำหนดเป็นลำดับหน้าจอและคำสั่งเหมือนเป็น visual programming เพื่อให้เห็นว่ากระบวนการนี้มีขั้นตอนอย่างไร




ข้อมูลที่สกัดได้จะเป็น Data Table ที่มี field ต่างๆ ตามชื่อที่ตั้งไว้คือ title, url, review, text ขั้นตอนต่อไปคือการเก็บข้อมูลลงไฟล์ ซึ่งสามารถเก็บเป็น Excel ได้ ให้พิมพ์หา Excel ในกล่องเครื่องมือจนเจอ Excel Application Scope แล้วลากมาตอนท้ายของกล่องของ Data Scraping เพื่อบอกว่าให้ใช้ Excel Application และในกล่อง Excel นี้ให้ใส่ชื่อไฟล์ที่ต้องการเก็บข้อมูล ในที่นี้ใส่ test.xlsx จากนั้นลาก Write Range แล้วบอกว่า Input ที่เป็น DataTable มาจากไหน ก็คือเอามาจากข้อมูลที่สกัดจากเว็บ เนื่องจากเราไม่ได้เปลี่ยนชื่อตัวแปรใน Output ของ process ก่อน ตัวแปรที่เก็บข้อมูลคือ ExtractDataTable (กรณีที่ถ้าเราต้องการใช้ตัวแปรที่อยู่นอกกล่อง ให้ไปดูที่ตัวแปรนั้นแล้วเปลี่ยน scope ให้ครอบคลุมถึงกล่องที่ต้องการใช้ได้) ข้อมูลจำนวนที่ต้องการสกัดจะถูกเก็บลงใน Excel ไฟล์ตามที่ต้องการ กรณีที่ต้องการ Header row ในไฟล์ Excel ด้วยก็ให้เลือก AddHeaders







ตัวอย่างที่ยกมาเป็นตัวอย่างง่าย ๆ ที่ข้อมูลทั้หงมดอยู่ในหน้าเว็บมีการแบ่ง field ต่างๆ ชัดเจน ข้อมูลที่มีหลายหน้าก็สามารถสกัดได้ แต่ในบางเว็บ เช่น ในเว็บตัวอย่างนี้ หากดูข้อมูลที่สกัดได้ จะเห็นว่าบางรายการไม่มีตัวบทวิจารณ์มา ทั้งนี้เพราะในหน้าเว็บไม่ได้แสดงข้อมูล field นี้ทั้งหมด ผู้

ใช้ต้องกดอ่าน full review อีกที  หากจะทำต่อจริง ๆ ก็ต้องสร้าง automation ต่อให้ดูว่ารายการที่มีตัวบท  ให้ไปเอา url ที่ดึงมา  ไปสร้าง address ส่งให้เว็บแล้วสกัดบทวิจารณ์นั้นมาใส่เพิ่มในรายการนั้น

### โปรแกรม Octoparse

Octoparse เป็นโปรแกรมสกัดข้อมูลบนเว็บอีกโปรแกรม แต่ใช้งานได้เฉพาะบน Windows เป็นโปรแกรมที่มี UI ช่วยให้ง่ายต่อการเลือกส่วนของข้อมูลที่ต้องการสกัดนำมาใส่เป็น field ต่าง ๆ ได้ มีข้อดีคือสามารถสกัดข้อมูลที่ปรากฏเป็นรายการซ้ำ ๆ ในหน้าเว็บและที่มีรายการอยู่มากกว่าหนึ่งหน้าก็ได้ ใน version ฟรีสามารถใช้สกัดข้อมูลได้ในปริมาณจำกัดได้ไม่เกิน 10,000 รายการ หากต้องการใช้งานจำนวนมากจะมีค่าใช้จ่ายรายเดือน $75

ขั้นตอนการใช้งานจะคล้ายกับ Parsehub คือมี 3 ขั้นตอน ดังนี้

1. ใส่ url ของหน้าเว็บที่มีข้อมูลที่ต้องการ
2. ให้รายละเอียดว่าจะสกัดข้อมูลส่วนไหนบ้าง จะทำซ้ำในหน้าเว็บ จะสกัดหลาย ๆ หน้าหรือไม่
3. สั่งให้ดำเนินการสกัด และ save ข้อมูลซึ่งสามารถเก็บเป็น csv, excel หรือ json ก็ได้

สำหรับรายละเอียดขั้นตอนการใช้งานสามารถศึกษาเพิ่มเติมได้จาก tutorial ของโปรแกรมเอง

### ใช้ R package สำหรับ web crawler

เราสามารถใช้ package ใน R เพื่อช่วยในการสกัดข้อมูลที่ต้องการจากเว็บเพื่อมาใช้ในงานที่ต้องการได้ package หนึ่งที่เป็นประโยชน์เพื่อการนี้คือ rvest ตัวอย่างข้างล่างดัดแปลงจาก

http://stat4701.github.io/edav/2015/04/02/rvest_tutorial/

เริ่มด้วยการติดตั้ง package rvest ของ R และเรียกใช้ package นี้

```
> install.packages("rvest")
>library(rvest)
```

load ข้อมูลจากหน้าเว็บ review ภาพยนตร์ Arrival มาเก็บไว้

```
> arrival_movie <- html("http://www.imdb.com/title/tt2543164/")
```

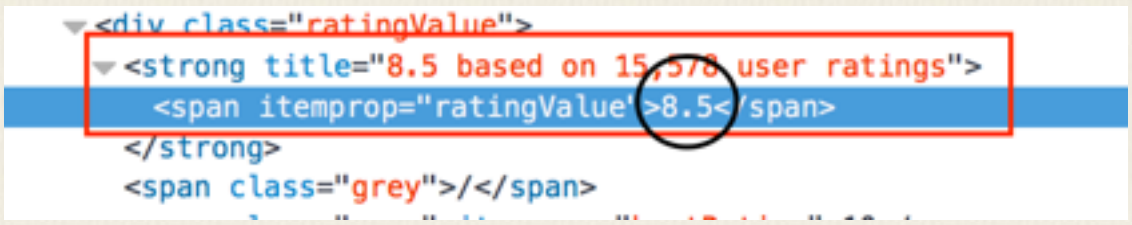
ดึงข้อมูล rating ออกมาเก็บไว้ที่ตัวแปร rating ข้อมูลนี้อยู่ในแท็ก strong ซึ่งภายใต้มีแท็ก span อยู่แล้ว สกัด text ในนั้นออกมา แปลงเป็นตัวเลข เครื่องหมาย %>% เป็นการ pipe หรือส่งผ่านข้อมูลต่อ ซึ่งจะเห็นว่าในคำสั่งให้เอาข้อมูลเว็บที่เก็บใน arrival_movie มาสกัด node ที่มีแท็ก strong แล้วก็ span เอาข้อมูลที่ได้ส่งต่อให้ html_text() ดึง text ออกมาส่งต่อให้มาแปลงป็นตัวเลข

```
> rating <- arrival_movie %>%
+   html_nodes("strong span") %>%
+   html_text() %>%
+   as.numeric()
> rating
```

หากดูภายในเว็บจะเห็นข้อมูลการให้คะแนนอยู่ในแท็กนี้

<span

itemprop="ratingValue">8.5




คำสั่งต่อมาสั่งให้หาโหนดที่มี attribute value เป็น titleCast (มีเครื่องหมาย # นำหน้า) แล้วดูต่อหา attribute ชื่อ itemprop (มีเครื่องหมาย . นำหน้า) แล้วต่อไปจนพบแท็ก span จึงส่งข้อมูลต่อให้ดึง text ออกมาซึ่งจะเป็นชื่อนักแสดง และเนื่องจากพบข้อมูลแบบนี้มากกว่าหนึ่ง จึงดึงทั้งหมดออกมาเป็นรายการชื่อทั้งหมดได้

```
> cast <- arrival_movie %>%
+   html_nodes("#titleCast .itemprop span") %>%
+   html_text()
> cast
 [1] "Amy Adams"          "Jeremy Renner"      "Michael Stuhlbarg"  "Forest Whitaker"
"Sangita Patel"
 [6] "Mark O'Brien"       "Abigail Pniowsky"   "Tzi Ma"             "Nathaly Thibault"   "Ruth
Chiang"
```

[11] "Jadyn Malone"  "Julia Scarlett Dan" "Russell Yuen"  "Anana Rydvald" "Leisa Reid"


จะเห็นว่าการใช้ rvest ช่วยให้เราสามารถดึงข้อมูลส่วนที่ต้องการในหน้าเว็บมาใช้ได้เลย  เพียงแต่เราต้องรู้ว่าข้อมูลนั้นอยู่ที่ไหน ภายใต้แท็กอะไร เพื่อจะได้เขียนคำสั่งสกัดให้ถูกตำแหน่งได้

**ใช้ Python package สำหรับ web crawler**

Python มีโมดูล webscrapy ที่ช่วยในการสกัดหน้าเว็บที่ต้องการได้  และมักใช้ร่วมกับF,ดูล BeautifulSoup ที่ช่วยให้เข้าถึงแท็กต่าง ๆ ใน html เพื่อเข้าถึง text ที่ต้องการได้  ตัวอย่างการเขียนโปรแกรม Python มีดังนี้

ส่วนแรกคือ import library ที่จำเป็นต้องใช้คือ url  กับ BeautifulSoup

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
import csv
import re
```

กำหนด url ของ pages ที่ต้องการสกัดข้อมูล ใน list ในที่นี้ให้เพียงหนึ่ง url สำหรับทดลองดู

```
quote_page = ['https://www.imdb.com/title/tt5028340/reviews?ref_=tt_ql_3']
```

ขั้นต่อไปคือวนรอบดึงแต่ละ url มา  urlopen ใช้สำหรับเปิด url นั้น  แล้วนำเข้าข้อมูล html โดยใช้ BeautifulSoup  เก็บโครงสร้าง html  ไว้ในตัวแปร  soup  จากนั้น เราต้องดูข้อมูลใน page นั้นว่าส่วนที่ต้องการสกัดอยู่ภายใต้แท็กชื่ออะไร  บทวิจาร์แต่ละอันอยู่ในแท็ก  ที่มี attribute

“class” ที่มีค่าเป็น 'review-container'  วนลูปด้วยการใช้ find_all ของ BeautifulSoup  จากนั้นในแต่ละ review เราจะสกัดสองอย่าง คือ rating 1-10  และตัวบทที่เป็นความเห็นของคนวิจารณ์ ตัวความเห็นอยู่ในแท็ก  ที่มี attribute “class” ที่มีค่าเป็น 'text show-more__control'  ดึงออกมา (ตัวอย่างคือ This film will be a classic one day! .....)  ส่วนคะแนน review อยู่ภายใต้  ..  (ตัวอย่างคือ 8/10)  จึงใช้ find ซ้อนกันสองครั้ง  การสกัดข้อมูลจากแต่ละเว็บจึงต้องเขียนแบบเจาะจงให้เข้ากับโครงสร้างข้อมูลแท็กภายในไฟล์ html นั้น ๆ

```
data = []
for pg in quote_page:
 # query the website and return the html to the variable 'page'
    page = urlopen(pg)
    # parse the html using beautiful soap and store in variable `soup`
    soup = BeautifulSoup(page, 'html.parser')
    # get the review item
    for item in soup.find_all('div', attrs={'class': 'review-container'}):
        ## get the review content
        review_box = item.find('div', attrs={'class':'text show-more__control'})
        review = re.sub(r'<.+?>','',str(review_box))
        ## get rating number
        rate_box = item.find('span').find('span')
        rate = re.sub(r'<.+?>','',str(rate_box))
        print(rate,'=>',review)
        data.append((rate,review))
```

หลังจากนั้นจึงเขียนข้อมูลลงไฟล์ csv เพื่อนำไปใช้ต่อ  ซึ่งจะทำให้ได้ข้อมูลตามตัวอย่าง

```
with open('reviews.csv', 'a') as csv_file:
```

```
  writer = csv.writer(csv_file)
# The for loop
  for rate, review in data:
      writer.writerow([rate, review])
```

| A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | Growing up Mary Poppins was one of my favorite films. When Saving Mr. Banks came out a few years back I re-watched the ori | | | | | | | | | |
| 10 | This film will be a classic one day! It was great! Emily Blunt as Mary Poppins was the only person I would have picked for the rol | | | | | | | | | |
| 8 | Greetings again from the darkness. The 1964 classic Disney film MARY POPPINS is much beloved and has been shared across ge | | | | | | | | | |
| 10 | Mrs. Shullivan and I treated ourselves to a special early Christmas present by attending the opening night of Mary Poppins Retur | | | | | | | | | |
| 3 | With the multiple references to the original Mary Poppins throughout this movie, and even in the title, my hope was this movie | | | | | | | | | |
| 10 | This is profoundly one of the years very best pictures! It is filled with pure bliss and the magic Disney has been aiming to posses | | | | | | | | | |
| 2 | Excellent performances of thoroughly unimaginative music, a pale reflection of the original, which was all about the music. It's a | | | | | | | | | |
| 3 | I can't begin to say how excited I was for this movie. I've been waiting over a year for it. We brought my daughter to see it with | | | | | | | | | |
| 2 | Long and boring! Fell asleep several times and I just wanted it to end. None of the songs made any sense and they are totally fo | | | | | | | | | |
| 2 | The acting was good. The characters were likeable but the music let it down. This film wasn't necessary. | | | | | | | | | |
| 1 | I went to see this, against my basic instinct, because I thought the original film was so marvellously good that they couldn't go s | | | | | | | | | |
| 1 | Why every time they need to sing every damn thing I have never been more annoyed with a movie as I have watching this crap. | | | | | | | | | |
| 3 | First things | | | | | | | | | |
| 2 | It took all of me not to get up and walk out...and I've sat through some bad movies! But the story & music were just awful. | | | | | | | | | |
| 2 | Since Disney is incompetent of coming up with new ideas, and must resort to using older stories they did years ago, they certain | | | | | | | | | |
| 1 | The great | | | | | | | | | |
| 1 | I believe that 'Mary Poppins Returns' is the soulless Disney nostalgia cash-grab number 112? I don't know, I lost count long ago.I | | | | | | | | | |
| 1 | This film was a huge disappointment. It has a cast of classic actors including Emily Blunt, Meryl Street, Colin Firth and Angela Ln | | | | | | | | | |
| 1 | Honestly, I was cringing by the end of this movie and would have left if I weren't with a large group. Mary Poppins didn't really d | | | | | | | | | |
| 1 | It was always going to be hard to go up against a cherished classic but this film is just dire. Don't waste your time. I had to walk | | | | | | | | | |
| 2 | ***Warning Mild Spoilers***Where was the joy? The charm? The humor? Mary Poppins Returns is such a slog of a movie and p | | | | | | | | | |

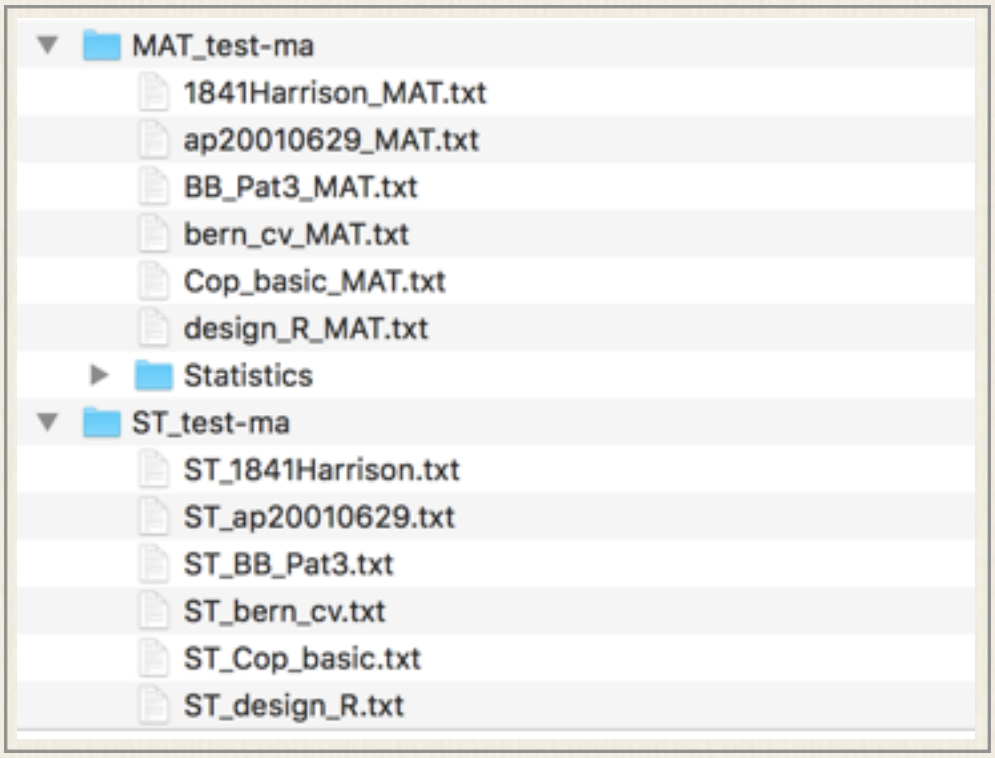
# โปรแกรมสำหรับทำ Dimension Reduction

การวิเคราะห์โดยใช้ multivariate analysis เป็นวิธีการทางสถิติที่ได้รับความนิยมมากขึ้นในงานวิจัยที่ข้อมูลมีตัวแปรต้น (independent variable) หลายตัว ในกลุ่มนี้ เทคนิคการลด dimension เป็นวิธีการที่ถูกนำมาใช้ในการศึกษาภาษา งานที่เป็นที่รู้จักดีคืองานของ Biber (1993) ที่ใช้ Factor Analysis ลดตัวแปรทางภาษาจำนวนมากที่พบใน text ลงเหลือ 5-6 dimension ที่สามารถใช้แยกความต่างทาง genre ของตัวบทได้ หรืองานของ Binongo (2003) ที่ใช้วิธีการลด dimension เพื่อหาลักษณะการเขียนของแต่ละคนและสรุปว่างานชิ้นที่สงสัยควรเป็นงานที่เขียนโดยคนไหน โปรแกรมที่ใช้สำหรับทำงานทั้งสองแบบมีผู้พัฒนาให้ใช้ ดังนี้

**โปรแกรม** Multidimensional Analysis Tagger

เป็นโปรแกรมที่เขียนขึ้นโดย Andrea Nini เพื่อวิเคราะห์ตัวบทโดยอาศัยลักษณ์ต่าง ๆ ที่พบในตัวบทตามแนวการวิเคราะห์ที่ Douglas Biber ใช้ในการวิเคราะห์และพัฒนาโปรแกรม Variation across Speech and Writing tagger โปรแกรมทำงานบน Windows และสามารถดาวน์โหลดได้จากเว็บ https://sites.google.com/site/multidimensionaltagger/

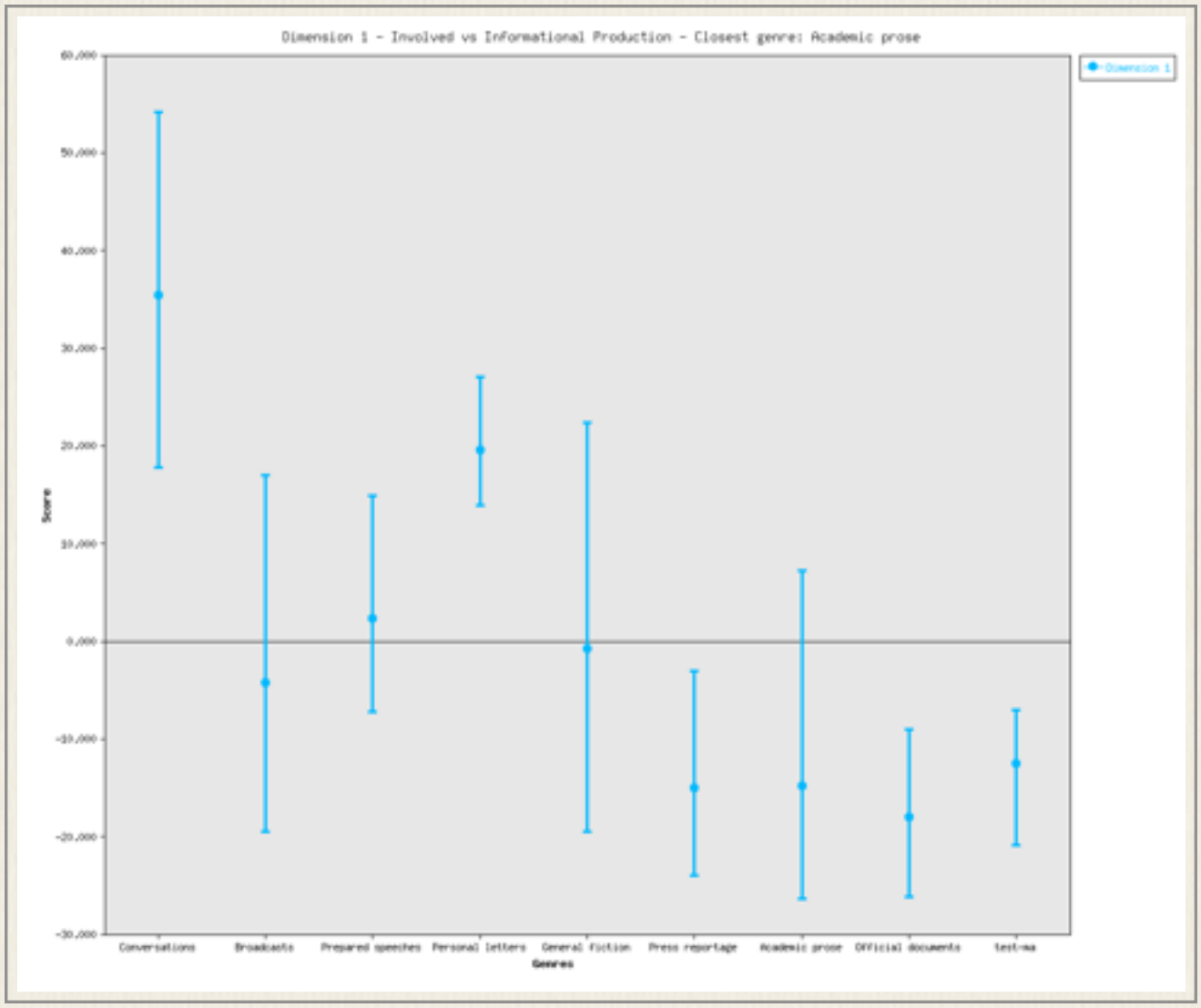
สิ่งที่โปรแกรมทำคือ run Tagger และ Analyze Tagger ที่ใช้ในส่วน POS Tagger ใช้ Stanford Tagger ตามด้วย จากนั้นจึง tag ลักษณ์ทางภาษาอื่น ๆ แล้วทำ multidimensional analysis ตามแบบ Biber โดยลด dimension ออกมาเป็น 6 dimension ตัวอย่างข้างล่างแสดงผลการวิเคราะห์จะเห็น folder ย่อยที่ถูกสร้างขึ้นสองแฟ้มภายใต้ folder ที่เราเก็บข้อมูล




folder ST_ เก็บข้อมูลที่มีการแท็ก POS  folder MAT_ เก็บข้อมูลที่แท็กลักษณ์ภาษาเพิ่มเติม แสดงใน [ ]  ภายใต้ folder MAT_  จะมี folder statistics ที่เก็บผลการวิเคราะห์ไว้  ในตัวอย่าง cor-pus นี้มีไฟล์ 6 ไฟล์  แต่ละไฟล์จะถูกวิเคราะห์ว่าเมื่อมองใน dimension 1-6 แล้วมีค่าคะแนนเท่าไร และควรจัดเป็นตัวบทประเภทใด เช่น learned exposiiton, scientific exposition  และถ้ามองภาพรวมทั้ง corpus ควรเป็นประเภทใด  ในตัวอย่างนี้ได้ใส่ text ที่มีความต่างประเภทอยู่

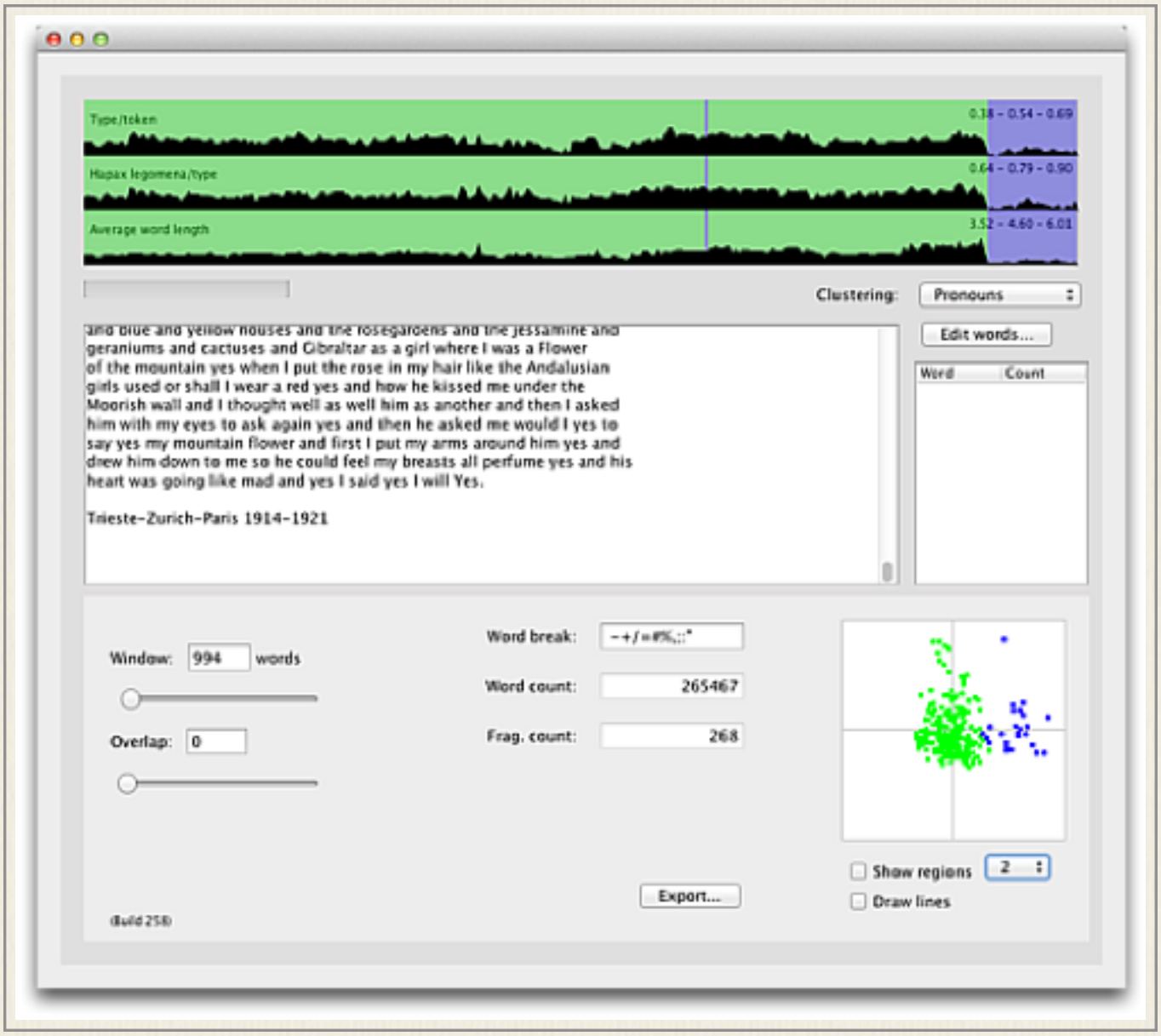
| Filename | Dimension1 | Dimension2 | Dimension3 | Dimension4 | Dimension5 | Dimension6 | Closest Text Type |
|---|---|---|---|---|---|---|---|
| BB_Pat3 | -7.11 | -2.27 | 6.42 | 2.51 | 6.53 | -0.43 | Scientific exposition |
| Cop_basic | -17.23 | -4.31 | 6.61 | -0.64 | 3.19 | -1.96 | Learned exposition |
| bern_cv | -20.98 | -3.3 | 9.22 | 1.66 | 5.67 | 0.69 | Learned exposition |
| 1841Harrisor | -8.86 | 0.99 | 9.88 | 2.42 | 2.97 | 1.07 | General narrative exposition |
| design_R | -9.35 | -2.99 | 6.98 | 1.28 | 7.74 | -1.57 | Scientific exposition |
| ap20010629 | -11.84 | 1.97 | -0.11 | -4.65 | -2.72 | -2.05 | General narrative exposition |
| CORPUS | -12.56 | -1.65 | 6.5 | 0.43 | 3.9 | -0.71 | Scientific exposition |

และเมื่อแสดงผลเป็นกราฟตาม dimension ต่าง ๆ รูปข้างล่างเป็นผลใน dimension 1 “Involved vs Informational Production” corpus ที่ทอลองเป็นเส้นขวาสุดเทียบกับตัวบทมาตรฐานที่เป็นตัวแทนตัวบทต่างๆ เช่น conversations, broadcasts, prepared speeches, personal letters, general fictions, press reportages, academic prose, และ official documents




## โปรแกรม Text Variation Explorer

โปรแกรม Text Variation เป็นอีกโปรแกรมที่อาศัยการวิเคราะห์แบบ dimension reduction มีการแปลง text เป็นเว็กเตอร์และลดมิติด้วยเทคนิค PCA เพื่อใช้วิเคราะห์เปรียบเทียบระหว่าง text ได้ โปรแกรมนี้พัฒนาโดย Harri Siirtola ในโครงการ DAMMOC project สามารถดาวน์โหลดได้จาก http://www.uta.fi/sis/tauchi/virg/projects/dammoc/tve.html เป็นโปรแกรม Java จึงใช้ได้กับทั้ง Windows และ Mac




ตัวอย่างการใช้งานโปรแกรมนี้คืองานของ Binongo (2003) ที่ต้องการศึกษาว่า หนังสือที่เป็นประเด็นตั้งคำถามกันว่า หนังสือ The Royal Book of Oz เป็นหนังสือเล่มสุดท้ายที่เขียนโดย Lyman Frank Baum หรือเป็นเล่มแรกที่ Ruth Plumly Thompson เขียนกันแน่ วิธีการตรวจสอบคือการนำหนังสือที่แน่ใจแล้วว่าเขียนโดย Baum และที่เขียนโดย Thompson แยกเป็นข้อมูลสองชุด และเลือกคำที่ไม่แปรไปตามเนื้อเรื่องแต่เป็นคำที่สามารถใช้บ่งบอกลักษณะสไตล์การเขียนของแต่ละคนได้ เช่น คำกลุ่มคำไวยากรณ์ ใช้คำเหล่านี้ที่ใช้มากสุด 50 คำมาเป็นตัวแทนสร้างเว็กเตอร์ 50 มิติของ text แต่ละส่วน จากนั้นใช้ PCA (Pinciple Componential Analysis) ลดเหลือ 2 มิติเพื่อพล็อต text แต่ละส่วนลงไป ซึ่งจะเห็นว่าจากสไตล์การเจียนที่มีลักษณะการใช้คำไวยากรณ์ต่างกัน text ของนักเขียนคนเดียวกันจะกระจุกอยู่ใกล้กัน ในขณะที่ text จาก Baum และ Thompson จะแยก

จากกัน  จากนั้นใช้วิธีการเดียวกันวิเคราะห์ text ในหนังสือที่สงสัยแล้วพล็ตดู ก็จะเห็นว่าใกล้กับกลุ่ม ของ Thompson มากกว่า

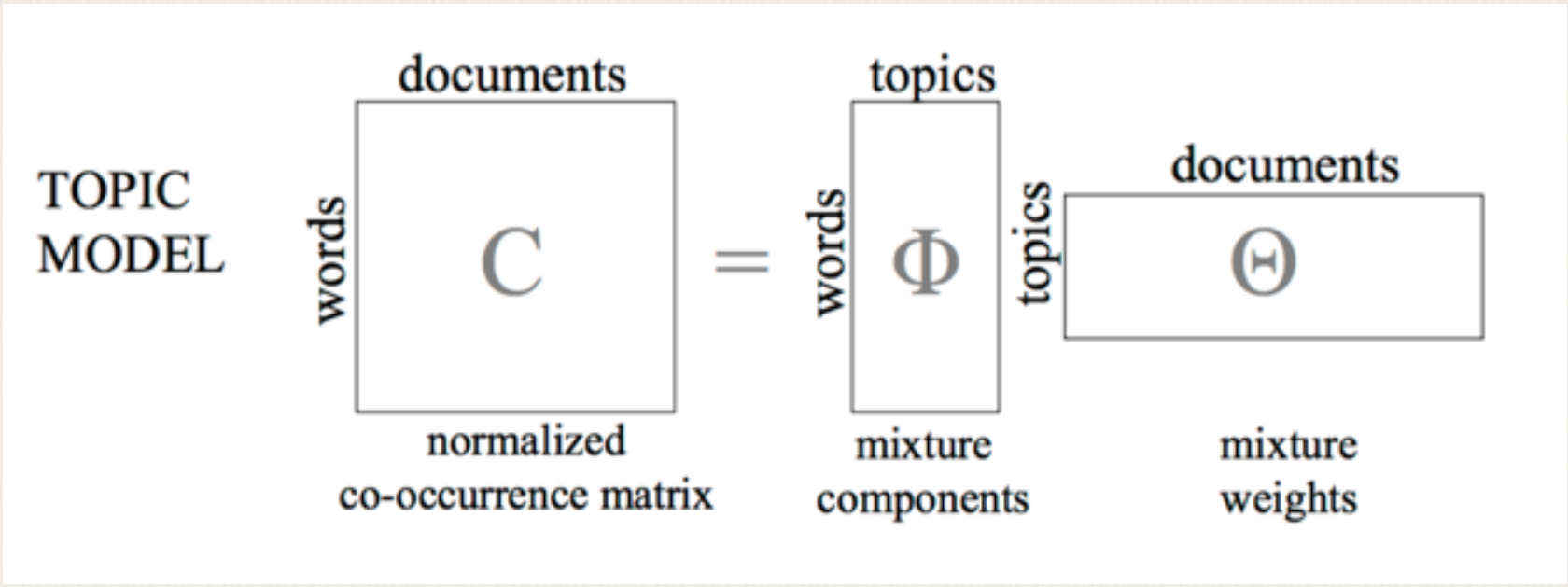

- Biber, Douglas. (1993). The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An overview of Methodology and Findings.  Computers and the Humanities 26: 331-345.
- Binongo, José Nilo G. (2003). Who wrote the 15th Book of Oz? An application of multivariate analysis to authorship attribution. Chance 16(2): 9–17.

SECTION 11

# โปรแกรม Topic Modeling

เป็นโปรแกรมสำหรับวิเคราะห์เอกสารต่าง ๆ ที่มี เพื่อมองหาว่าในเอกสารนั้นมี topic หรือหัวข้ออะไรบ้าง และหัวข้อนั้น ๆ แสดงออกผ่านทางรูปคำอะไรบ้าง มีโปรแกรมที่สามารถดาวน์โหลดมาใช้ได้ฟรีทั้งบน Windows และ Mac เช่น โปรแกรม GUI สำหรับ MALLET Topic Modeling (https://github.com/senderle/topic-modeling-tool) ซึ่งช่วยให้การใช้งาน MALLET topic modeling (http://mallet.cs.umass.edu/topics.php) ทำได้สะดวกขึ้น หรือโปรแกรม Topic modeling ของ Stanford สามารถดาน์โหลดได้ที่ https://nlp.stanford.edu/software/tmt/tmt-0.4/

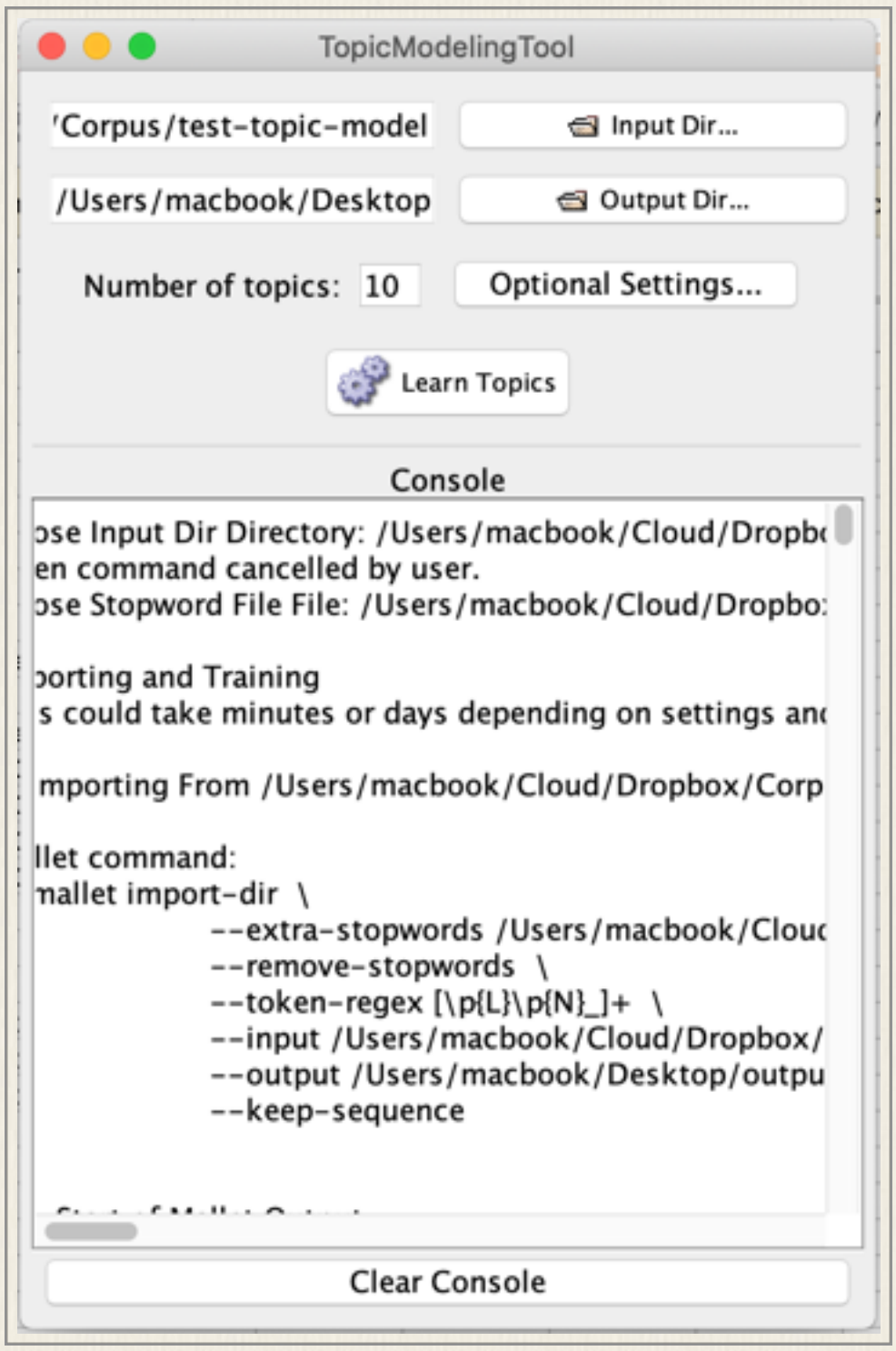
Topic modeling เป็นการใช้วิธีการทางสถิติเพื่อหาว่าในเอกสารต่างๆที่มีนั้นมีหัวข้อหรือ topic อะไรอยู่บ้าง อาศัยหลักการสร้างเมทริกซ์ดูการกระจายตัวของคำต่างๆในเอกสารสร้าง Document-Term matrix ออกมา สมติว่ามีเอกสาร d ชิ้น และมีจำนวนรูปศัพท์ในเอกสารทั้งหมด n ศัพท์ ก็จะสามารถสร้างเมทริกซ์ขนาด d x n ได้ ซึ่งเมทริกซ์ d x n นี้ สามารถมองได้ว่าเป็นผลมาจากการรวมกันของเทริกซ์ขนาด d x k และ k x n ได้ ซึ่งถ้าเรามองว่า k นี้คือจำนวน topic ทั้งหมดที่มีในข้อมูล เราก็สามารถหาคำตอบนี้ได้ เพราะจะได้ผลที่บอกว่า คำที่เกี่ยวข้องกับ topic ต่างๆ (1..k) มีคำอะไรบ้าง เกี่ยวข้องมากน้อยเพียงใด ส่วนผลอีกตารางจะบอกว่าเอกสารแต่ละชิ้นมีความเกี่ยวข้องกับ topic ต่างๆ กี่เรื่องและเกี่ยวข้องมากน้อยแค่ไหน อัลกอลิทึมพื้นฐานที่มักใช้กันในการทำ topic modeling คือ latent Dirichlet allocation (LDA; Blei, Ng, and Jordan 2003)




` เพื่อให้เห็นผลที่ชัดเจนการการทำ topic modeling ในตัวอย่างนี้จึงทดลองใช้ไฟล์ข้อมูลจากโดเมนที่แตกต่างกันขัดเจน ได้แก่ รายงานภาวะเศรษฐกิจ 7 ไฟล์, เอกสารทรัพย์สินทางปัญญา 7 ไฟล์, คำกล่าวสุนทรพจน์ 7 ไฟล์, นิยายของเจน ออสติน 3 เล่ม แล้วใช้โปรแกรม GUI ของ MALLET Topic modeling กับข้อมูลชุดนี้ โดยตั้งจำนวน topic ไว้ที่ 10 topic เมื่อกำหนด input directory และ output directory แล้ว ผลเก็บในรูป csv และ html ใน folder output_csv และ output_html โปรแกรมจะบอกว่า topic ต่าง ๆ นั้นมีคำอะไรที่สำคัญใน topic นั้น โดยให้ชื่อเป็น topic 0-n ตามจำนวน topic ที่ตั้งไว้ตอนต้น

การระบุว่า topic นั้นเกี่ยวกับเรื่องหรือประเด็นอะไร เป็นสิ่งที่ผู้วิจัยต้องใช้ความรู้ของตนระบุชื่อเอง ถ้าต้องการสำรวจผลที่ได้ว่า จำนวน topic ที่กำหนดและจำนวนรอบการ run นั้นให้ผลดีพอหรือยัง ก็ควรเปิด folder output_html เพราะมี link ที่โยงข้อมูลให้คลิกดูได้สะดวกมากกว่า ส่วนใน output_csv จะเห็นไฟล์ csv ที่สรุปความสัมพันธ์ระหว่าง word, topic, doc เมื่อเปิดดูไฟล์ top_in_doc.csv จะเห็นว่าแต่ละ doc นั้นมีเนื้อหาเกี่ยวกับ topic อะไรมากน้อยกว่ากันโดยคำนวณมาเป็นน้ำหนัก จากข้อมูลที่นำมาทดสอล จะเห้นว่าไฟล์เนื้อหาทางด้านเศรษฐกิจถูกจัดว่ามีน้ำหนักเป็น topic 1 และ 6 มากสุด ไฟล์สุนทรพจน์มีเนื้อหาหนักไปทาง topic 5 ไฟล์นวนิยายมีเนื้อหาเป็น topic 7

ส่วนไฟล์ top_in_word.csv แสดงถึงรายการคำสำคัญในแต่ละ topic ที่วิเคราะห์มาโดยเรียงตามลำดับความสำคัญของคำต่าง ๆ



| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | docId | filename | toptopics... | | | | | | | | | |
| 2 | 1 | :-model/Econ6.txt | 1 | 0.885509 | 6 | 0.094686 | 2 | 0.011777 | 9 | 0.006541 | 5 | 0.001452 |
| 3 | 2 | :-model/Econ7.txt | 1 | 0.925717 | 6 | 0.067827 | 2 | 0.00507 | 7 | 9.67E-04 | 9 | 3.80E-04 |
| 4 | 5 | :-model/Econ5.txt | 1 | 0.936576 | 6 | 0.04541 | 2 | 0.014489 | 9 | 0.003474 | 8 | 1.76E-05 |
| 5 | 7 | :-model/Econ4.txt | 1 | 0.940114 | 6 | 0.035583 | 2 | 0.024246 | 9 | 1.69E-05 | 8 | 1.39E-05 |
| 6 | 3 | .905-Roosevelt.txt | 5 | 0.724357 | 2 | 0.141972 | 7 | 0.115547 | 6 | 0.016541 | 9 | 5.37E-04 |
| 7 | 11 | .885-Cleveland.txt | 5 | 0.842149 | 2 | 0.149401 | 1 | 0.004623 | 6 | 0.003191 | 9 | 2.47E-04 |
| 8 | 16 | odel/1909-Taft.txt | 5 | 0.746463 | 2 | 0.159577 | 8 | 0.035656 | 1 | 0.02883 | 6 | 0.021531 |
| 9 | 17 | /1889-Harrison.txt | 5 | 0.778569 | 2 | 0.111491 | 7 | 0.039536 | 9 | 0.034806 | 6 | 0.02164 |
| 10 | 22 | 1901-McKinley.txt | 5 | 0.769398 | 2 | 0.197317 | 8 | 0.016379 | 7 | 0.009352 | 9 | 0.004828 |
| 11 | 23 | .893-Cleveland.txt | 5 | 0.808717 | 2 | 0.137304 | 7 | 0.021332 | 9 | 0.016723 | 6 | 0.011997 |
| 12 | 24 | 1897-McKinley.txt | 5 | 0.771384 | 2 | 0.180006 | 1 | 0.027493 | 9 | 0.018826 | 7 | 0.002054 |
| 13 | 10 | :-model/Econ1.txt | 6 | 0.902746 | 1 | 0.086751 | 2 | 0.005236 | 8 | 0.002615 | 9 | 0.002495 |
| 14 | 12 | :-model/Econ3.txt | 6 | 0.907203 | 1 | 0.087003 | 2 | 0.005106 | 8 | 6.35E-04 | 9 | 2.07E-05 |
| 15 | 15 | :-model/Econ2.txt | 6 | 0.926648 | 1 | 0.064479 | 2 | 0.008677 | 9 | 1.46E-04 | 8 | 1.73E-05 |
| 16 | 18 | %20Sensibility.txt | 7 | 0.443415 | 3 | 0.303965 | 2 | 0.251887 | 5 | 5.05E-04 | 9 | 1.89E-04 |
| 17 | 20 | del/Persuasion.txt | 7 | 0.424923 | 4 | 0.328804 | 2 | 0.241655 | 9 | 0.002042 | 3 | 0.001333 |
| 18 | 21 | d%20Prejudice.txt | 7 | 0.446169 | 0 | 0.322595 | 2 | 0.230289 | 9 | 5.97E-04 | 3 | 2.07E-04 |
| 19 | 6 | odel/BB_CP_4.txt | 8 | 0.917078 | 9 | 0.045114 | 2 | 0.035551 | 6 | 0.001026 | 0 | 9.00E-04 |
| 20 | 8 | -model/.DS_Store | 8 | 0.356986 | 5 | 0.35613 | 2 | 0.101019 | 9 | 0.052641 | 6 | 0.051425 |

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Topic Id | Top Words... | | | | | | | | | | | |
| 2 | 0 | elizabeth darcy bennet bingley jane miss wickham collins lydia mother catherine hope sister friend father gardiner lizzy lor | | | | | | | | | | | |
| 3 | 1 | percent market rate 1 1999 growth prices 2 2000 rates quarter 1998 4 financial federal debt economic real inflation investme | | | | | | | | | | | |
| 4 | 2 | time day world home house found attention visit passed business situation determined brought told subject expected dou | | | | | | | | | | | |
| 5 | 3 | elinor marianne edward mother dashwood jennings willoughby miss sister lucy john brandon colonel ferrars middleton ho | | | | | | | | | | | |
| 6 | 4 | anne captain elliot wentworth charles russell walter mary musgrove louisa miss bath father elizabeth lady friend uppercro | | | | | | | | | | | |
| 7 | 5 | people government public congress laws american country policy law citizens united business executive power national fr | | | | | | | | | | | |
| 8 | 6 | sales district prices demand activity reported contacts continued construction report strong percent manufacturing growth | | | | | | | | | | | |
| 9 | 7 | lady young dear sir sister project feelings happy cried morning family replied left love pleasure mind gutenberg letter ever | | | | | | | | | | | |
| 10 | 8 | copyright rights author law owner protection property act literary licence moral trade united article intellectual artistic king | | | | | | | | | | | |
| 11 | 9 | information confidence public confidential law defendant breach employee employer duty plaintiff obligation nature cont | | | | | | | | | | | |
| 12 | | | | | | | | | | | | | |

# โปรแกรม Corpus Annotator

**โปรแกรม annotator** คือโปรแกรมที่ช่วยในการ markup ข้อมูลเพิ่มเติมในตัวบท โปรแกรมบางตัวออกแบบมาเพื่อใช้เหมือนเป็นการทำ note เพิ่มเติมในเอกสาร เช่น โปรแกรม Annotator (http://annotatorjs.org) ที่เขียนด้วยภาษา Javascript ให้ผู้ใช้สามารถเติมข้อความเพิ่มเติมไปในเอกสารผ่านทางเว็บได้ ไม่ว่าจะเป็นหน้าเว็บที่เป็น html หรือ pdf หรือ EPUB จึงเหมาะสำหรับการ markup ทั่วไปแต่ไม่เหมาะกับการกำกับข้อมูลทางภาษาศาสตร์เพื่อสร้างข้อมูลที่จะนำมาวิเคราะห์ต่อ

**โปรแกรม brat rapid annotation tool** (http://brat.nlplab.org)

เป็นโปรแกรมที่ทำงานบนเว็บสำหรับกำกับข้อมูลทางภาษา เช่น ชื่อเฉพาะ หมวดคำ ความสัมพันธ์ระหว่างคำ จึงเป็นโปรแกรมที่เหาะกับการกำกับข้อมูลภาษาศาสตร์เพื่อนำมาวิเคราะห์ต่อมากกว่า

โปรแกรมนี้ต้องติดตั้งบน server ทำงานด้วยโปรแกรม Python 2.5 ขึ้นไป (ใช้บน Python 3 ไม่ได้) ขั้นตอนนี้จะต้องให้ผู้ดูแล server เป็นคนติดตั้งให้ก่อนจึงจะใช้งานได้ เมื่อติดตั้งเรียบร้อย ผู้ที่ต้องการใช้ จะต้องมีไฟล์ข้อมูล และไฟล์ configuration ที่กำหนดแท็กต่างๆ และความสัมพันธ์ที่จะมีระหว่างแท็กได้ แล้ว upload ไฟล์เหล่านี้ขึ้นไปที่ server เพื่อใช้งานได้

ไฟล์กำหนด the configuration ของการทำ annotation มีสี่ไฟล์ คือ

**annotation.conf: annotation type configuration**
**visual.conf: annotation display configuration**
**tools.conf: annotation tool configuration**
**kb_shortcuts.conf: keyboard shortcut tool configuration**

ไฟล์ annotation.conf ประกอบด้วยสี่ส่วน คือ [entities]  [relations] [events] และ [attributes]  [entities] ใช้กำหนดแท็กสำหรับกำกับแต่ละหน่วย เช่น Person, Organization, Location หน่วยที่จะมีทั้งหมดปรากฏล่าง [ENTITIES]  หนึ่งบรรทัดต่อหนึ่งหน่วย  Entities อาจมีลำดับชั้นได้ด้วยการจัดเรียงแล้วเคาะแท็บหน้ากลุ่มเป็นลูกกลุ่ม

```
[entities]
Person
Location
Organization
```

```
[entities]
Living-thing
                Person
                Animal
                Plant
Nonliving-thing
                Building
                Vehicle
```

[relations] ใช้กำหนดความสัมพันธ์ระหว่าง 2 entity  เช่น  กำหนด Argument1 เป็น Person  Argument2 เป็น Organization สำหรับความสัมพันธ์ Employment  ตามตัวอย่างนี้

```
[relations]
Family        Arg1:Person, Arg2:Person
Employment    Arg1:Person, Arg2:Organization
```

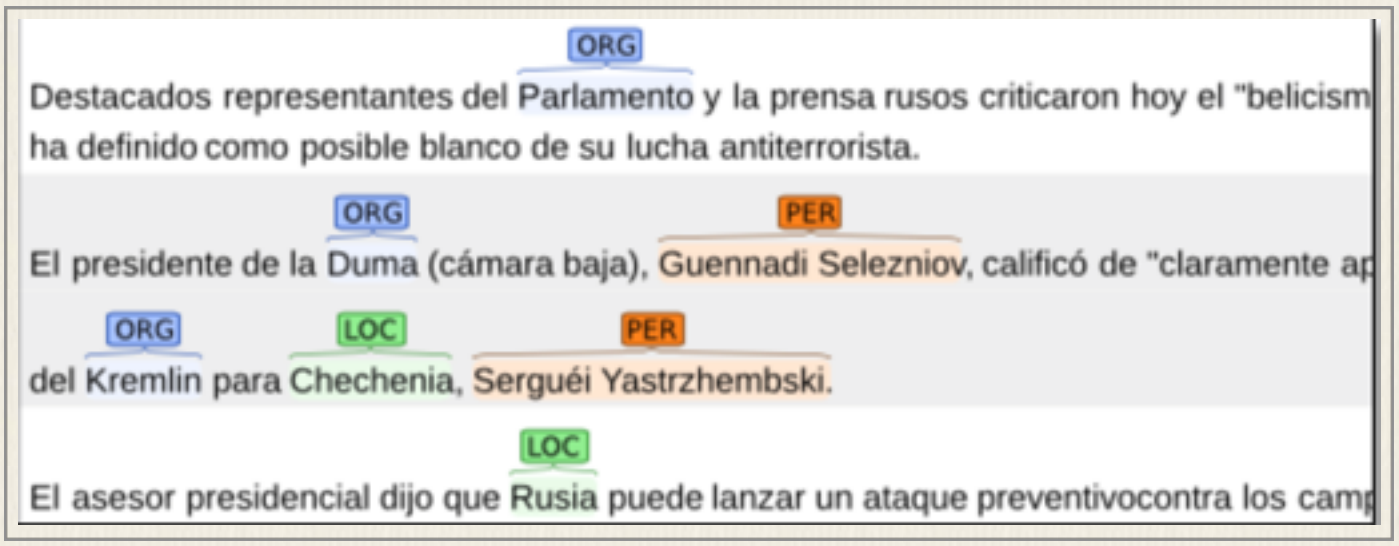
[events] ใช้กำหนดความสัมพันธ์ของ Entity หรือ Relation ที่มีใน Event นั้น จะมีองค์ประกอบหนึ่งตัวหรือมากกว่าสองก็ได้  มีลักษณะคล้ายยการกำหนด Frame ของเหตุการณ์ว่าประกอบด้วยอะไรบ้าง  องค์ประกอบเขียนในรูป Role:Type เช่น Participant1:Person  Org:Company  เช่น ตัวอย่างนี้

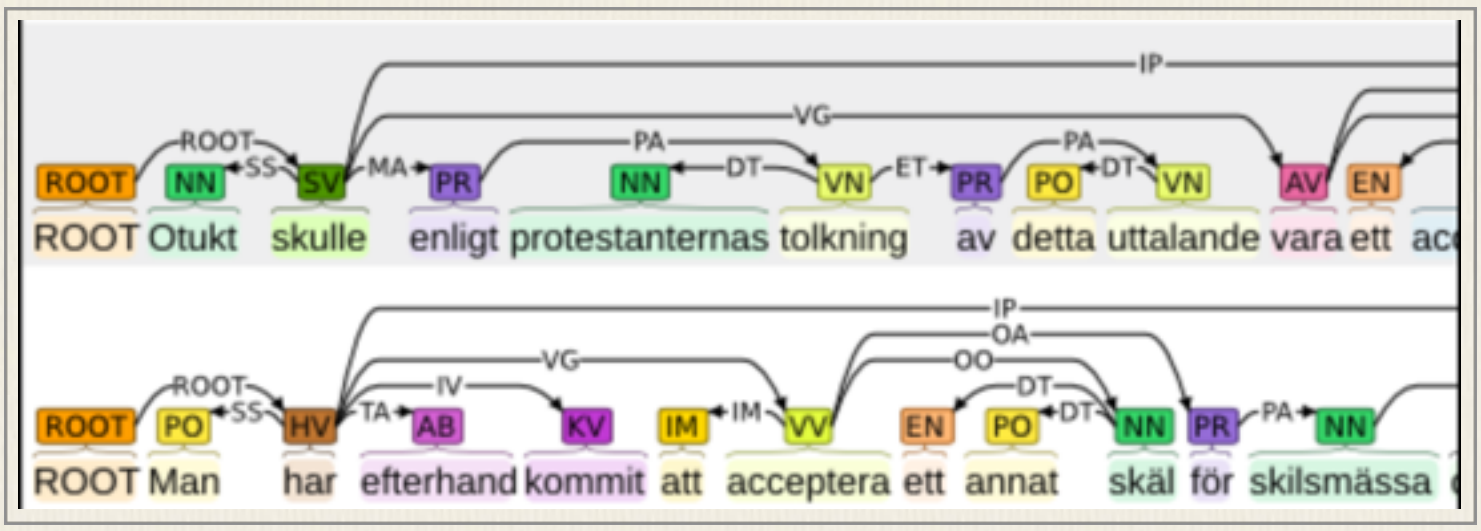
```
[events]
Marriage      Participant1:Person, Participant2:Person
Bankruptcy    Org:Company
```

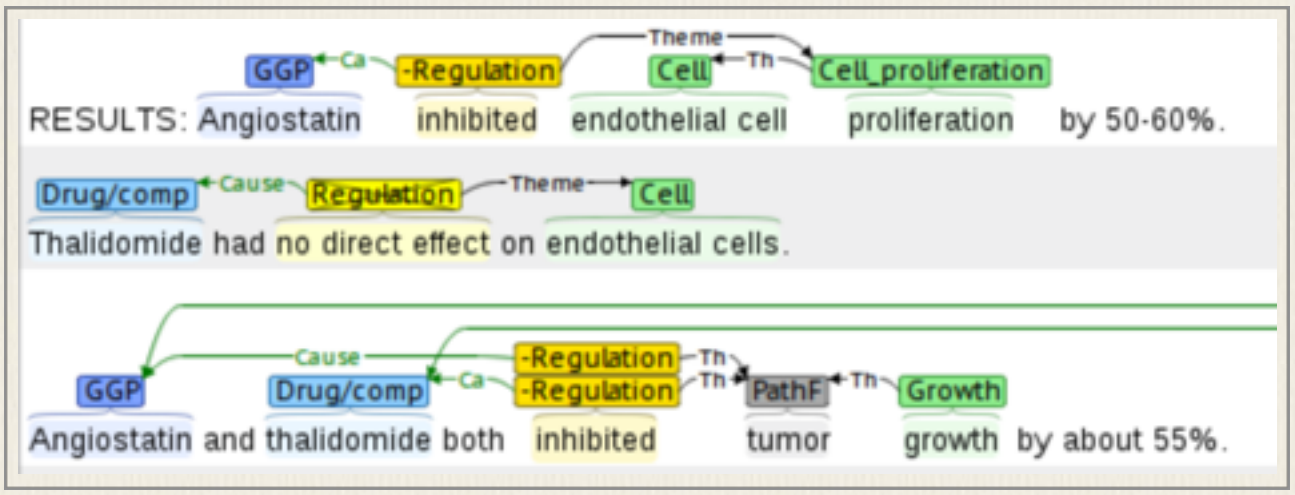
[attributes] ใช้กำหนด value ให้กับ annotation หรือแท็กต่างๆ ที่กำกับ จะมีค่าเป็น Binary หรือมีค่าเป็น ตัวเลือกต่างๆที่กำหนดก็ได้

```
[attributes]
Negated       Arg:<EVENT>
Confidence    Arg:<EVENT>, Value:L1|L2|L3
```

สำหรับ configuration อื่นเป็นเรื่องการกำหนดการแสดงผลหน้าจอและสีต่างๆ ที่แสดง รายละเอียดสามารถดูได้จากเว็บโครงการเอง Brat annotation นี้ได้ถูกใช้สำหรับกำกับข้อมูลทางภาษาศาสตร์หลายเรื่อง เช่น การกำกับชื่อเฉพาะ การกำกับโครงสร้างต้นไม้แบบพึ่งพา การกำกับ Event ดังตัวอย่างที่แสดงข้างล่างนี้






## การติดตั้งโปรแกรม Brat

เราสามารถใช้งาน Brat บนเครื่องตัวเองได้โดยให้ run server บนเครื่องตัวเอง ขั้นตอนคือให้ download Brat version 1.3

http://weaver.nlplab.org/~brat/releases/brat-v1.3_Crunchy_Frog.tar.gz

จากนั้นจึงกระจายไฟล์ออกมาโดยใช้คำสั่งนี้

```
tar xzf brat-v1.3_Crunchy_Frog.tar.gz
```

เมื่อได้ folder ออกมาแล้ว ให้วาง folder ในที่ที่ต้องการและติดตั้งโปรแกรมด้วยคำสั่งนี้ (เครื่องต้องมี Python 2 อยู่แล้ว)

```
./install.sh -u
python2 standalone.py
```

เมื่อเสร็จแล้ว จะเห็นว่า server start ให้ทดลองไปที่ browser แล้วเรียก http://127.0.0.1:8001 จะเห็นหน้าจอของ Brat ขึ้นมา จะมีตัวอย่างให้ทดลองใช้ ให้ login เป็น admin รหัสผ่าน admin-brat (ถ้าต้องการใช้ชื่อหรือเพิ่มชื่อ user อื่นให้ไปแก้ไขไฟล์ config.py)

จากนั้นแก้ไขไฟล์ annotation.conf ให้รู้จักแท็กต่างๆ ที่ต้องการใช้ เช่น

```
[spans]
ADJ
ADP
ADV
```

AUX
CCONJ
DET
INTJ
NOUN
NUM
PART
PRON
PROPN
PUNCT
SCONJ
SYM
VERB
X

[relations]
X Arg1:<ENTITY>, Arg2:<ENTITY>
acl Arg1:<ENTITY>, Arg2:<ENTITY>
advcl Arg1:<ENTITY>, Arg2:<ENTITY>
advmod Arg1:<ENTITY>, Arg2:<ENTITY>
amod Arg1:<ENTITY>, Arg2:<ENTITY>
appos Arg1:<ENTITY>, Arg2:<ENTITY>
aux Arg1:<ENTITY>, Arg2:<ENTITY>
case Arg1:<ENTITY>, Arg2:<ENTITY>
cc Arg1:<ENTITY>, Arg2:<ENTITY>
ccomp Arg1:<ENTITY>, Arg2:<ENTITY>
clf Arg1:<ENTITY>, Arg2:<ENTITY>
compound Arg1:<ENTITY>, Arg2:<ENTITY>
conj Arg1:<ENTITY>, Arg2:<ENTITY>
cop Arg1:<ENTITY>, Arg2:<ENTITY>
csubj Arg1:<ENTITY>, Arg2:<ENTITY>
dep Arg1:<ENTITY>, Arg2:<ENTITY>
det Arg1:<ENTITY>, Arg2:<ENTITY>
dislocated Arg1:<ENTITY>, Arg2:<ENTITY>
fixed Arg1:<ENTITY>, Arg2:<ENTITY>
flat Arg1:<ENTITY>, Arg2:<ENTITY>
iobj Arg1:<ENTITY>, Arg2:<ENTITY>
list Arg1:<ENTITY>, Arg2:<ENTITY>
mark Arg1:<ENTITY>, Arg2:<ENTITY>
nmod Arg1:<ENTITY>, Arg2:<ENTITY>
nsubj Arg1:<ENTITY>, Arg2:<ENTITY>

```
nummod Arg1:<ENTITY>, Arg2:<ENTITY>
obj Arg1:<ENTITY>, Arg2:<ENTITY>
obl Arg1:<ENTITY>, Arg2:<ENTITY>
orphan Arg1:<ENTITY>, Arg2:<ENTITY>
parataxis Arg1:<ENTITY>, Arg2:<ENTITY>
punct Arg1:<ENTITY>, Arg2:<ENTITY>
reparandum Arg1:<ENTITY>, Arg2:<ENTITY>
xcomp Arg1:<ENTITY>, Arg2:<ENTITY>

[events]
# none

[attributes]
Abbr Arg:<ENTITY>, Value:isAbbr
Foreign Arg:<ENTITY>, Value:isForeign
NameType Arg:PROPN, Value:Com|Geo|Giv|Oth|Prs|Sur
NounType-Class Arg:NOUN, Value:isClass
NumType-Mult Arg:NUM, Value:isMult
PartType Arg:PART, Value:Emp|Int|Neg|Res
Prefix Arg:<ENTITY>, Value:isPrefix
PronType Arg:PRON, Value:Prs|Rcp
Root Arg:<ENTITY>, Value:isRoot
VerbType-Cop Arg:VERB, Value:isCop
```

และกำหนดค่าต่างๆ ใน visual.conf กับ tool.conf ตามที่ต้องการ

## การนำเข้าข้อมูล

ข้อมูลสามารถนำเข้าผ่านการใส่ plain text ผ่านโปรแกรมโดยตรงได้ โปรแกรมจะไปแยกส่วนต่างๆ ตามที่ระบุใน tool.conf แต่ในกรณีภาษาไทย เราคงต้องการ preprocess ข้อมูลก่อน เช่น เราต้องการข้อมูลที่ผ่านการตัดคำและใส่ POS ตามแบบ UD เราต้องเตรียมไฟล์ ann กับ .txt ให้โปรแกรม ไฟล์ .txt เป็นข้อมูลดิบ ส่วน .ann เป็นไฟล์ตาม format ของ Brat

สมมติเราต้องการนำข้อมูลที่ได้จาก tltk.nlp.pos มาใช้

**คน/NOUN|ที่/SCONJ|นับถือ/VERB|สิ่ง/NOUN|เหล่านี้/DET|เป็น/VERB|คน/NOUN|ที่/SCONJ|น่าวิตก/VERB|มาก/ADV|<s/>/PUNCT|ถ้า/SCONJ|เป็น/VERB|ชาย/NOUN|ก็/SCONJ|เคย/AUX|**

**บวชเรียน/VERB|มา/VERB|แล้ว/ADV|<s/>/PUNCT|หญิง/NOUN|ก็/SCONJ|เคย/AUX|เข้า/VERB|วัด/NOUN|ทำบุญ/VERB|<s/>/PUNCT|ฟังเทศน์/NOUN|ได้/AUX|<s/>/PUNCT|ทำ/VERB|การ/NOUN|ต่างๆ/DET|<s/>/PUNCT|อย่าง/NOUN|ที่/SCONJ|พุทธศาสนิกชน/VERB|ทำ/VERB|กัน/PRON|<s/>/PUNCT|มี/VERB|พระพุทธรูป/NOUN|กราบ/VERB|ไหว้/VERB|แต่/CCONJ|ก็/SCONJ|นับถือ/VERB|ของ/ADP|พวก/NOUN|นี้/DET|<s/>/PUNCT|ถ้า/SCONJ|ไป/VERB|ถาม/VERB|เขา/PRON|ว่า/SCONJ|ทำไม/PART|ถึง/VERB|นับถือ/NOUN|<s/>/PUNCT|ก็/SCONJ|ตอบ/VERB|ว่า/SCONJ|<s/>/PUNCT|ของ/ADP|เหล่านี้/DET|เป็น/VERB|ไสยศาสตร์/NOUN|มี/VERB|เหตุ/NOUN|อธิบาย/VERB|ไม่ได้/AUX|จึง/SCONJ|นับถือ/VERB|<s/>/PUNCT|<Fail>ที่นับถือเพราะขลั</Fail>/NOUN|<s/>/PUNCT|**

เราก็ต้องเขียนโปรแกรมเพื่อแปลงข้อมูลให้ได้ไฟล์ .ann ตามตัวอย่างนี้ (ใช้โปรแกรม

thpos2brat.py )

**T1 NOUN 0 2 คน**
**T2 SCONJ 2 5 ที่**
**T3 VERB 5 11 นับถือ**
**T4 NOUN 11 15 สิ่ง**
**T5 DET 15 23 เหล่านี้**
.....

เมื่อเตรียมข้อมูลเสร็จก็สามารถใช้ Brat ทำการกำกับข้อมูลได้ เช่น เติม relation ระหว่างคำ ข้อมูลที่กำกับจะถูกเติมเพิ่มในไฟล์ .ann เช่น

T1 NOUN 0 4 **ความ**
A1 Prefix T1
#1 syllables T1 **ความ**
T2 VERB 4 12 **มุ่งหวัง**
#2 syllables T2 **มุ่ง-หวัง**
T3 ADP 12 15 **ต่อ**
#3 syllables T3 **ต่อ**
T4 NOUN 15 21 **ปัจจัย**
#4 syllables T4 **ปัจ-จัย**
.....
R1 X Arg1:T1 Arg2:T2
R2 X Arg1:T4 Arg2:T5
R3 X Arg1:T4 Arg2:T3
R4 X Arg1:T1 Arg2:T4
.....

ผลลัพธ์ที่ได้สามารถแปลงไปเป็น format อื่นๆ ตามที่ต้องการ เช่น แปลงเป็น CONLLU โดยการใช้โปรแกรมแปลง เช่น brat2conllu.py (Chonlathorn Kwankajornkiet) -i บอกชื่อ folder ที่เก็บข้อมูล brat annotation -o บอกชื่อไฟล์ที่เป็น output -p บอกให้สร้างไฟล์เก็บข้อมูลที่เป็น non-projective tree คือมีการไขว้กันของ dependency

**python brat2conllu.py -i ../data/export_ann_20-2-18 -o brat-output**
**#=> output file in CONLLU is brat-output.conllu**

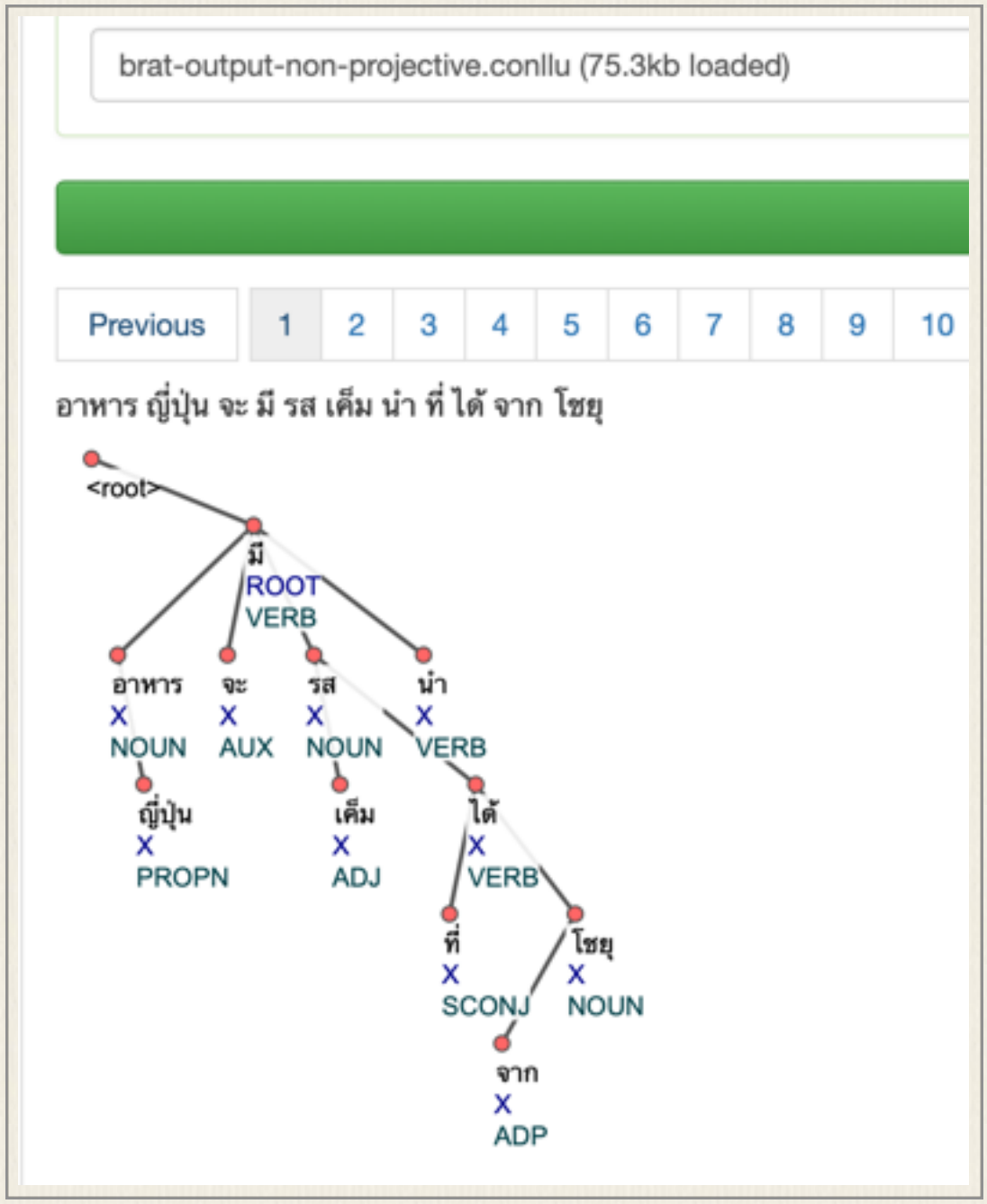
**python brat2conllu.py -i ../data/export_ann_20-2-18 -o brat-output -p**
**#=> create non-projective tree**

ข้อมูลในรูปแบบ CONLLU

```
1   ข้อควรปฏิบัติ   _   NOUN   NOUN   _   0   ROOT   _   SpaceAfter=No
2   เพื่อ   _   ADP   ADP   _   3   X   _   SpaceAfter=No
3   ความ   _   NOUN   NOUN   Prefix=Yes   1   X   _   SpaceAfter=No
4   ปลอดภัย   _   VERB   VERB   _   3   X   _   SpaceAfter=No
5   หาก   _   SCONJ   SCONJ   _   6   X   _   SpaceAfter=No
6   เกิด   _   VERB   VERB   _   1   X   _   SpaceAfter=No
7   วิกฤต   _   NOUN   NOUN   _   6   X   _   SpaceAfter=No
8   น้ำท่วม   _   NOUN   NOUN   _   7   X   _   SpaceAfter=No
```

ในระหว่างการแปลงเป็น conllu หากข้อมูลที่กำกับไม่ถูกต้อง เช่น ลืมโยงความสัมพันธ์ มี head มากกว่าหนึ่ง โปรแกรมจะเก็บข้อผิดนั้นไว้ในไฟล์ที่เป็น output โดยมี -error.txt ต่อท้าย เช่น brat-output-error.txt ให้ตรวจสอบการกำกับข้อมูลคำที่มีปัญหาแก้ไข

กรณีกำกับข้อมูลแล้วไม่ได้เป็น dependency tree ออกมา ไฟล์ที่มีชื่อ -non-projective.conllu จะเก็บข้อมูลที่มีปัญหาเหล่านี้ ให้ใช้โปรแกรมออนไลน์ https://universaldependencies.org/conllu_viewer.html เพื่อ load file "brat-output-non-projective.conllu") ดูต้นไม้ที่มีปัญหา และกลับไปกำกับข้อมูลใหม่ให้ถูกต้อง ตัวอย่างเช่น ต้นไม้ข้างล่างมีการเกยกันของ มี->นำ กับ รส -> ได้




จะต้องแก้ไขใหม่โดยโยง เค็ม -> นำ และตัด มี -> นำ ออกไป จึงจะเป็น projective tree ขึ้นมาได้

นอกจากนี้ยังมีโปรแกรมเขียนและแชร์ไว้ในภาษาต่าง ๆ เช่น Perl, Python ที่ใช้กับข้อมูล conllu

https://github.com/UniversalDependencies/cairo/blob/master/weaver/brat2conllu.pl
https://github.com/Binyamephrem/Amharic-treebank

ข้อมูลที่ได้ในรูป CONLLU นี้สามารถนำไปใช้ train โมเดล UD parser เช่น Malt Parser (http://www.maltparser.org) ได้ Malt Parser เป็นโปรแกรม java ที่รับข้อมูลกำกับภาษาใดก้ได้นำมา train ระบบเพื่อสร้างโมเดลสำหรับทำ UD Parser ได้ เช่น เมื่อสั่ง

**java -jar maltparser-1.9.2.jar -c test -i data/brat-outout.conllu -if conllu -m learn**

โมเดลของ UD Parser จะถูกสร้างและเก็บในชื่อ test.mco  และให้นำโมเดลนี้ไปใช้กับข้อมูลใหม่ต่อไปได้  วิธีการ parse ก็ใช้คำสั่งลักษณะเดียวกัน  ในตัวอย่างข้างล่าง s1.conllu เป็นไฟล์ input ที่คำแต่ละคำพร้อม POS ถูกแปลงเป็นรุปแบบ conllu แล้วส่งให้โมเดล test.mco parse ข้อมูลนี้ output ถูกเก็บไว้ในไฟล์ out.conllu

**java -jar maltparser-1.9.2.jar -c test -i data/s1.conllu -if conllu -o out.conllu -m parse**

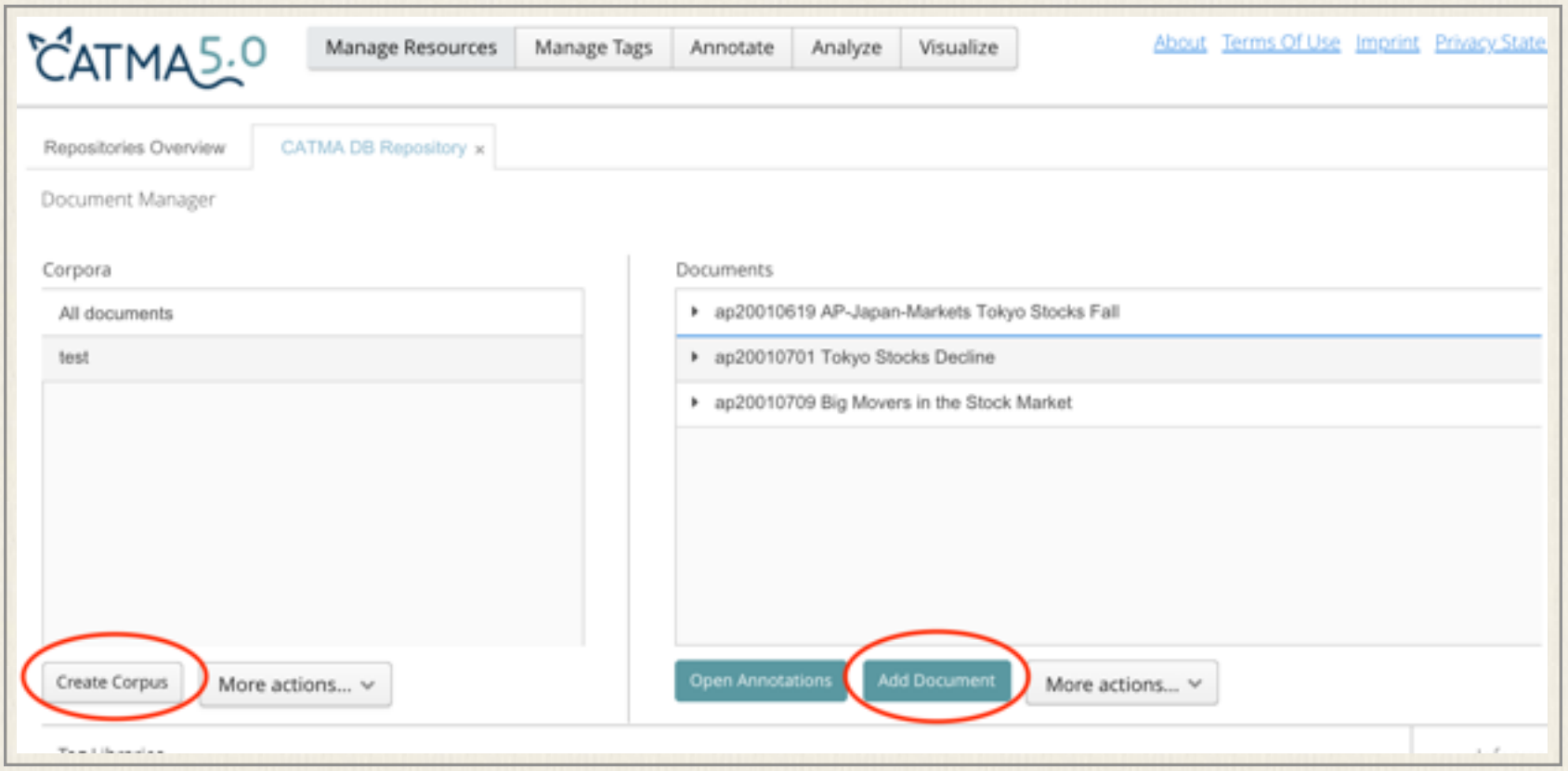
**โปรแกรม CATMA** (https://portal.catma.de/catma/)

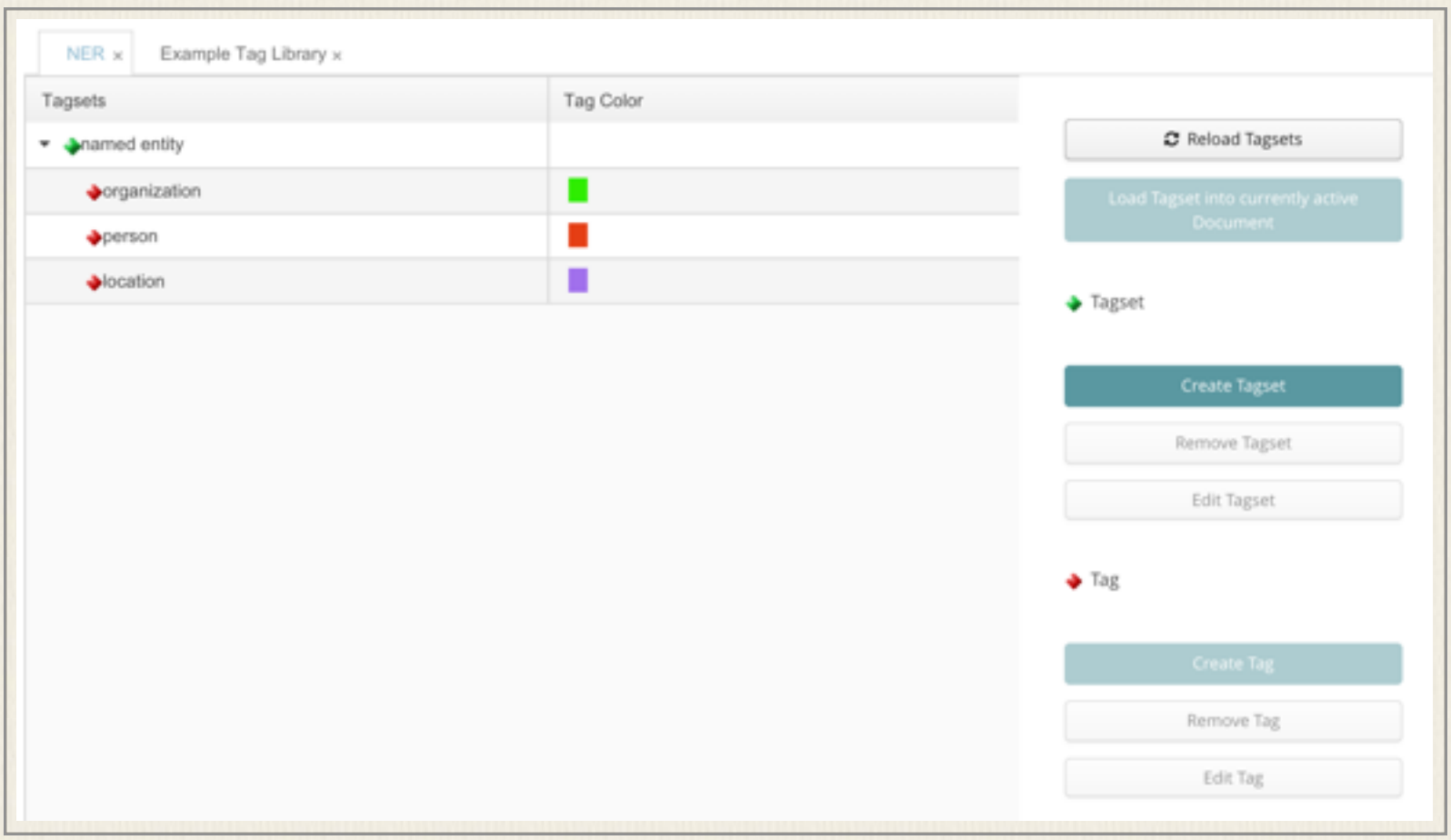
เป็นโปรแกรมสำหรับใช้กำกับข้อมูลใน corpus พร้อมช่วยวิเคราะห์ข้อมูลที่กำกับ ตัวโปรแกรมอยู่ที่ server ของ University of Hamburg ประเทศเยอรมัน CATMA เอง เป็น web-based โปรแกรมจึงสะดวกสำหรับผู้ใช้โดยทั่วไป เพียงแค่สมัครใช้งานก็สามารถเข้าไปทำงานได้เลย จะทำงานคนเดียวหรือแชร์งานทำด้วยกันหลาย ๆ คนก็ได้

CATMA สนับสนุนการกำกับข้อมูลแบบ external stand-off markup คือตัวข้อความกับข้อมูลที่กำกับถูกแยกจากกัน แต่เชื่อมโยงโดยาร cross-link ไปที่ตัว original text ได้ แท็กต่าง ๆ ที่ใช้สามารถกำหนดเองได้ตามวัตถุประสงค์ของงานที่ต้องการ เช่น แท็ก named entities หรือแท็กเพื่อใช้ในการวิเคราะห์ metaphor การกำกับข้อมูลทำง่าย ๆ ด้วยการเลือกข้อความส่วนที่ต้องการแล้วบอกว่าต้องการแท็กเป็นอะไร ข้อมูลที่กำกับแล้วสามารถนำออกมาเป็น XML ไฟล์ตามมาตรฐานของ TEI ได้

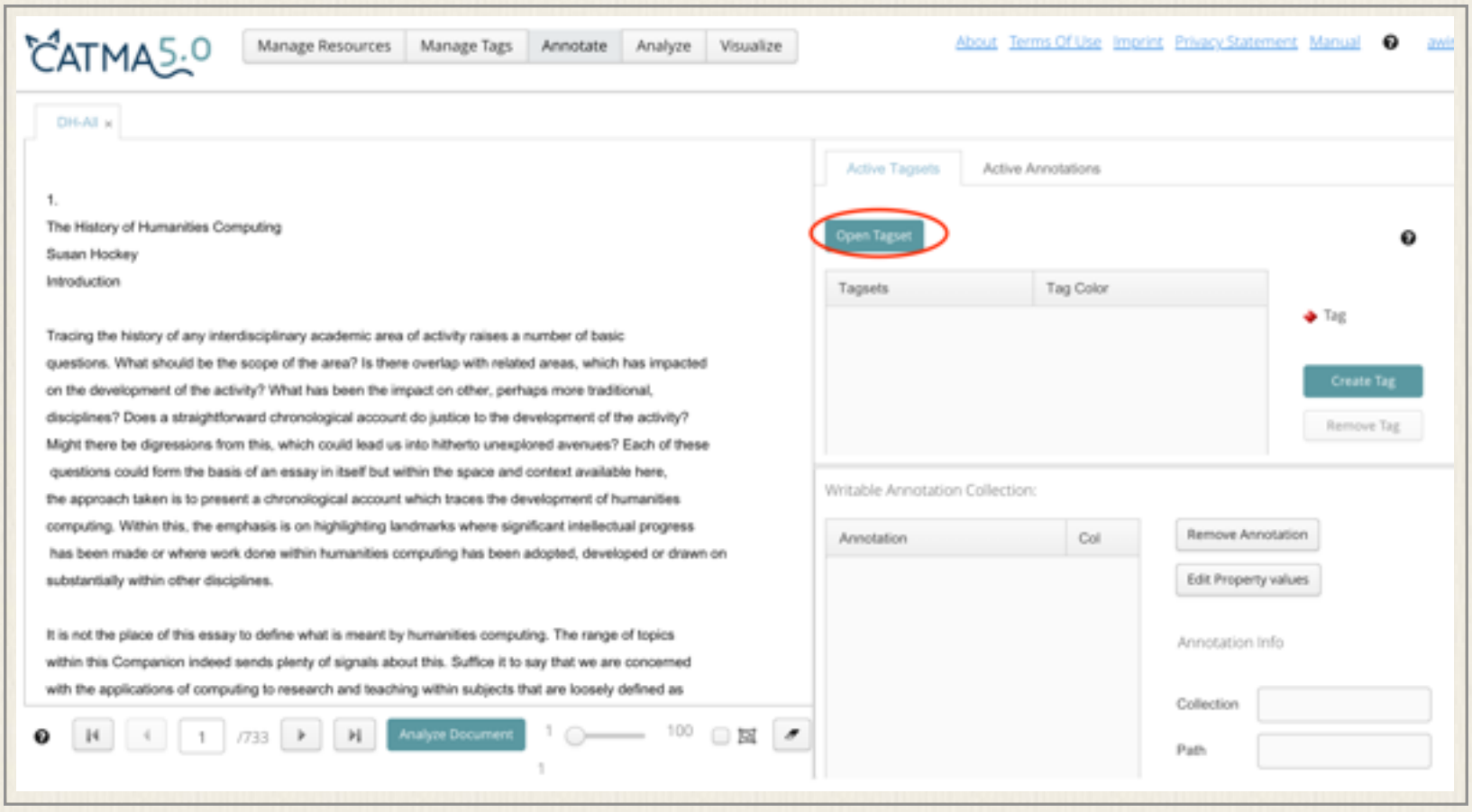
**ขั้นตอนการใช้งานโดยคร่าว ๆ มีดังนี**

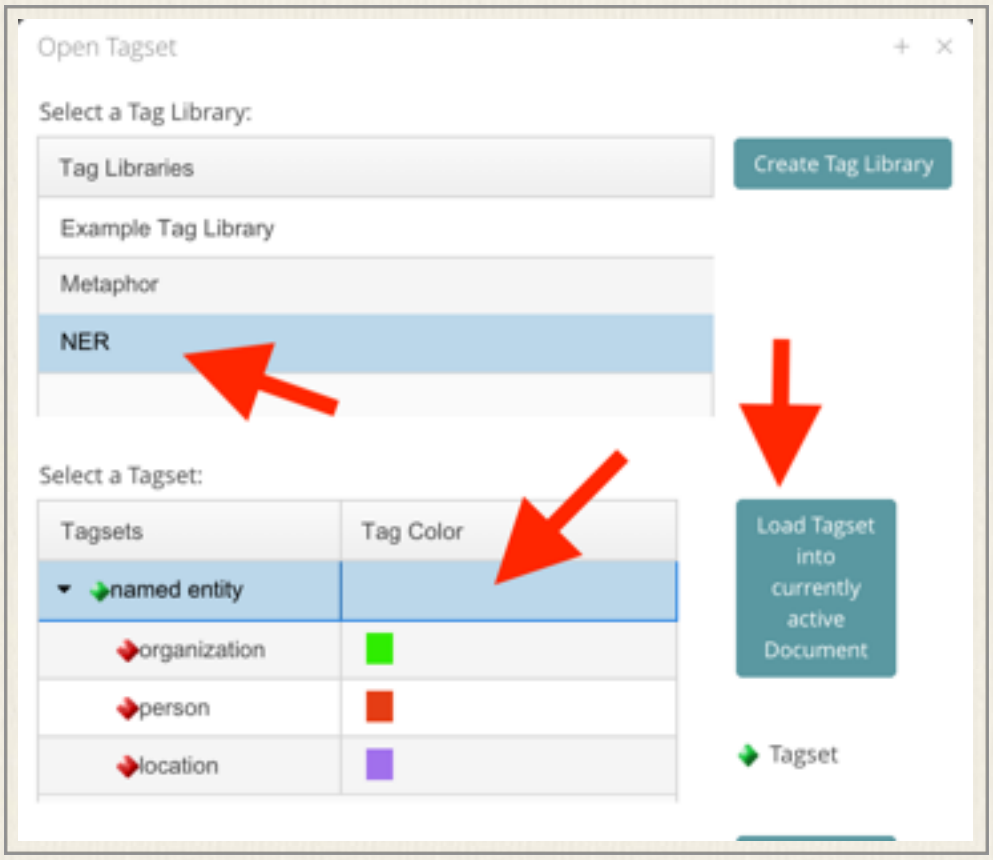
Manage Resource ขั้นแรกให้สร้าง corpus โดยการ upload ไฟล์ต่าง ๆ ที่ต้องการทำงานขึ้นไป อาจใช้วิธีการ upload ไฟล์ที่มีในเครื่องหรือระบุ url ที่มีข้อมูลนั้นก็ได้ หลังจาก upload ไฟล์เข้าไปแล้ว ขั้นต่อไปคือ กำหนดแท็กที่ต้องการใช้ในการกำกับข้อมูล ให้เลือก Create Tag Library แล้วจึง Open Tag Library ที่สร้างขึ้นมา จากนั้นกำหนดชื่อแท็กเซ็ตที่ต้องการ ภายในแท็กเซ็ตจึงตั้งชื่อแท็กแต่ละอันที่จะใช้ในการแท็ก พร้อมระบุสีที่ต้องการให้แสดงในหน้าจอ แท็กต่าง ๆ ที่สร้างขึ้นมานี้คือสิ่งที่จะไปใช้กำกับข้อมูลส่วนที่ต้องการ เช่น ในงาน NER หรือ named entities recognition เราต้องการกำกับข้อมูลชื่อสามประเภท คือ ชื่อบุคคล ชื่อองค์กร และชื่อสถานที่ ก็เลยสร้างแท็กไว้สามตัวในเซ็ตของ NER



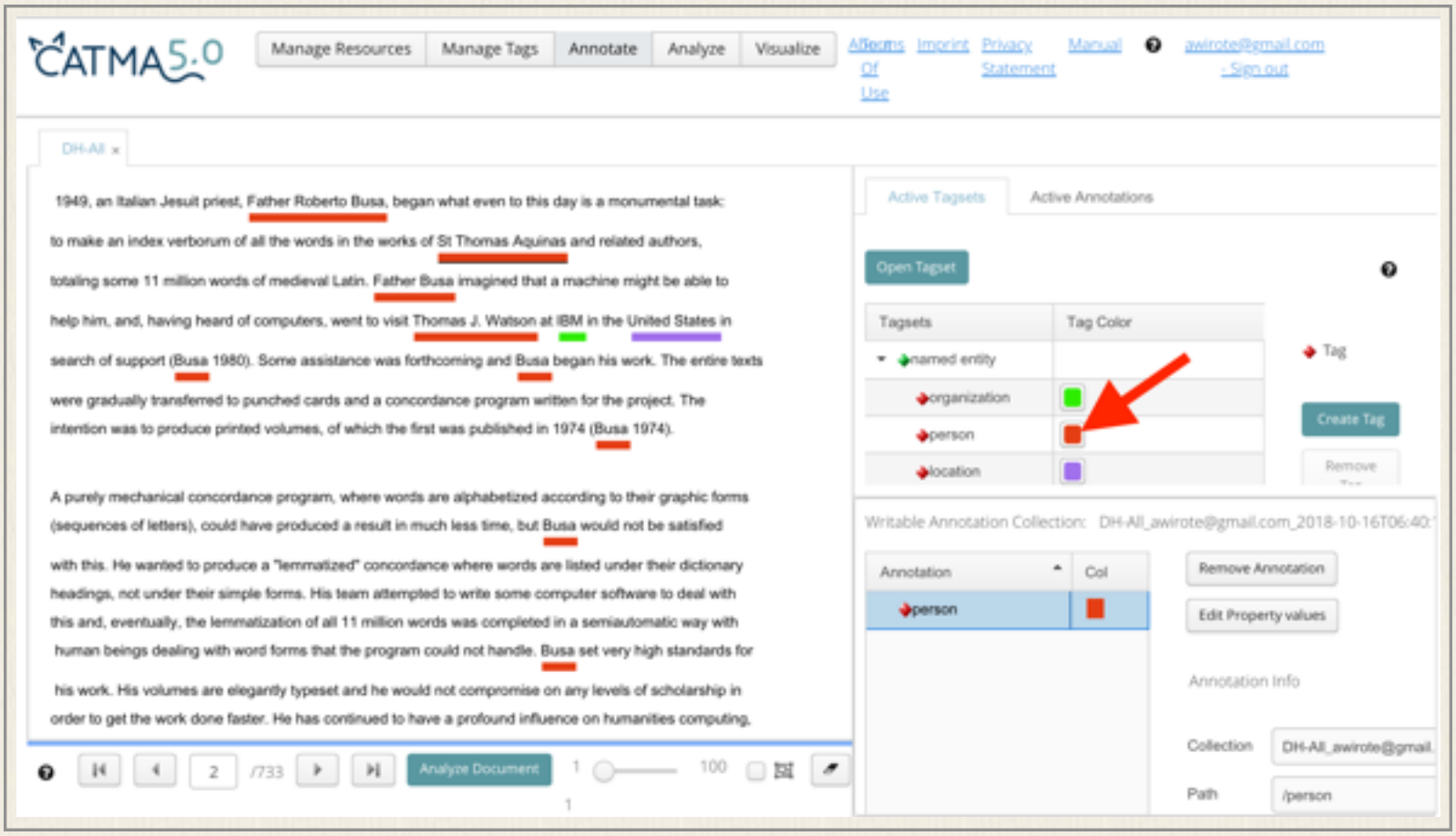




การกำกับข้อมูลทำได้ โดยการเลือก Document ที่ต้องการและสั่ง Open Document เมื่อเปิด Document มาได้แล้ว จะต้องเลือก Tagset ที่ต้องการใช้ ให้เลือก Open Tagset และในหน้าจอต่อไปจึงเลือก Tag Libraries และ Tagset ที่ต้องการใช้ แล้ว load Tagset นั้นมาใช้กับ Document



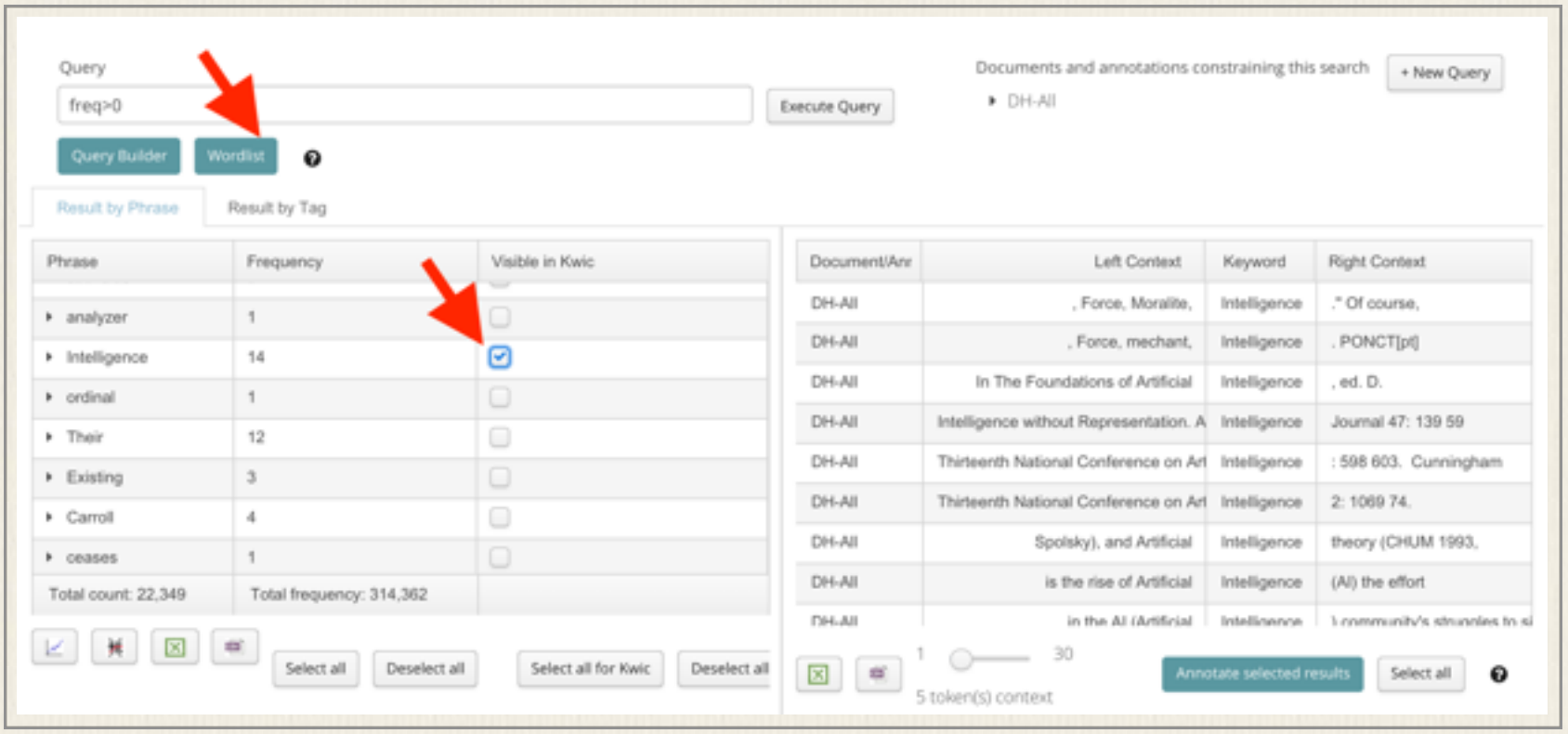




การกำกับข้อมูลทำได้ง่าย ๆ โดยการเลือกส่วนของ text ที่ต้องการใน Document นั้น แล้วไปกดเลือกแท็กที่ต้องการในตารางของ Tagset ให้กดที่ตัวไอคอนสีของแท็กนั้น จะเห็นว่าข้อความที่ถูกกำกับจะมีการระบายสีแท็กนั้นไว้ข้างใต้ให้เห็น ให้ทำการแท็กข้อมูลที่ต้องการไปจนครบใน Document นั้น แล้วจึงจะเข้าสู่ขั้นตอนต่อไป คือ การ Analyze Document

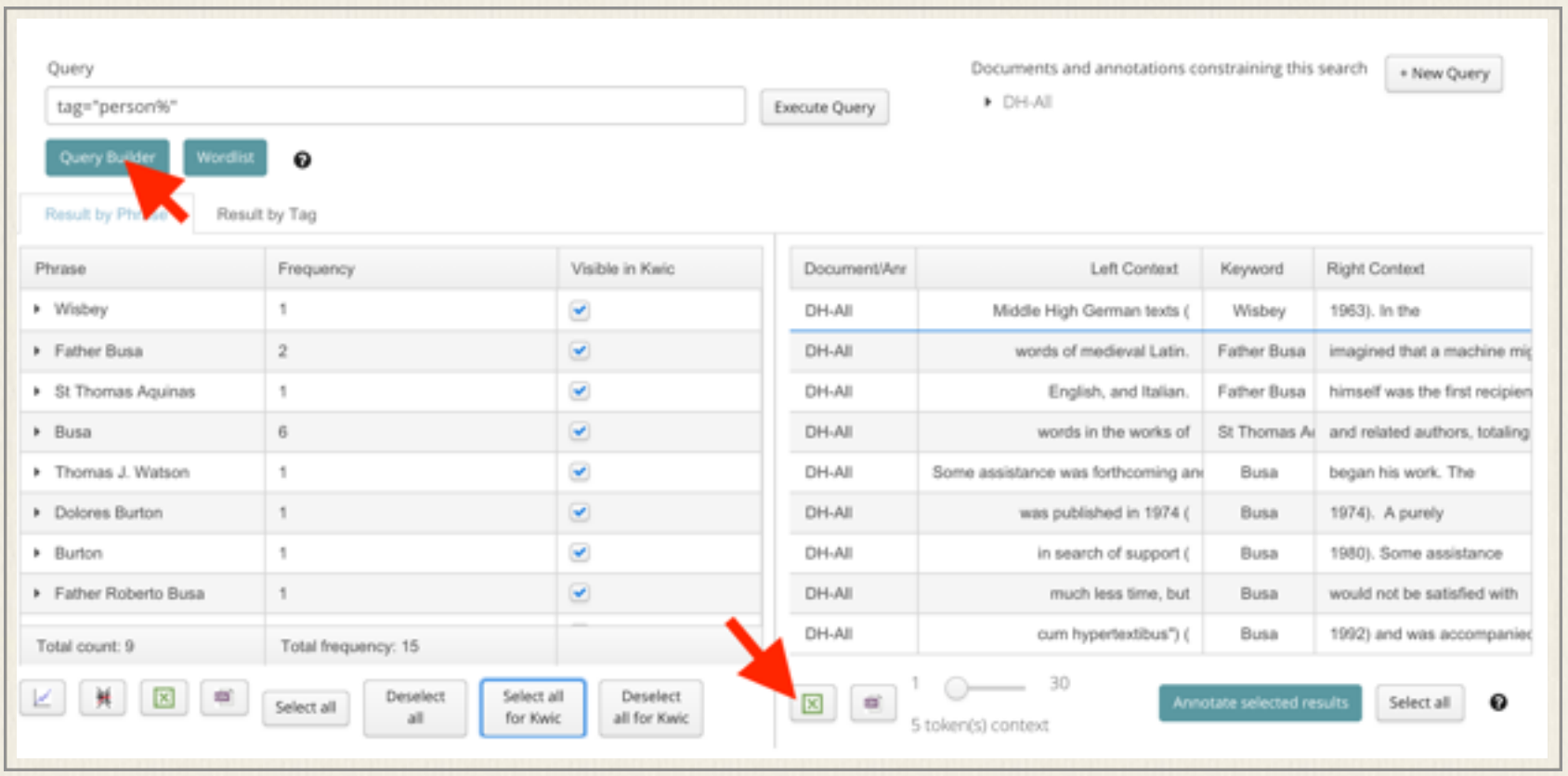



การ Analyze คือการมองหาข้อมูลที่ต้องการใน Document นั้น ทำได้ตั้งแต่การค้นหาคำหรือวลี หรือค้นหาข้อมูลที่ถูกกำกับด้วยแท็กต่าง ๆ สามารถแสดงผลออกมาในรูปที่เป็น kwic และ save ไฟล์ออกมาเป็น Excel ได้

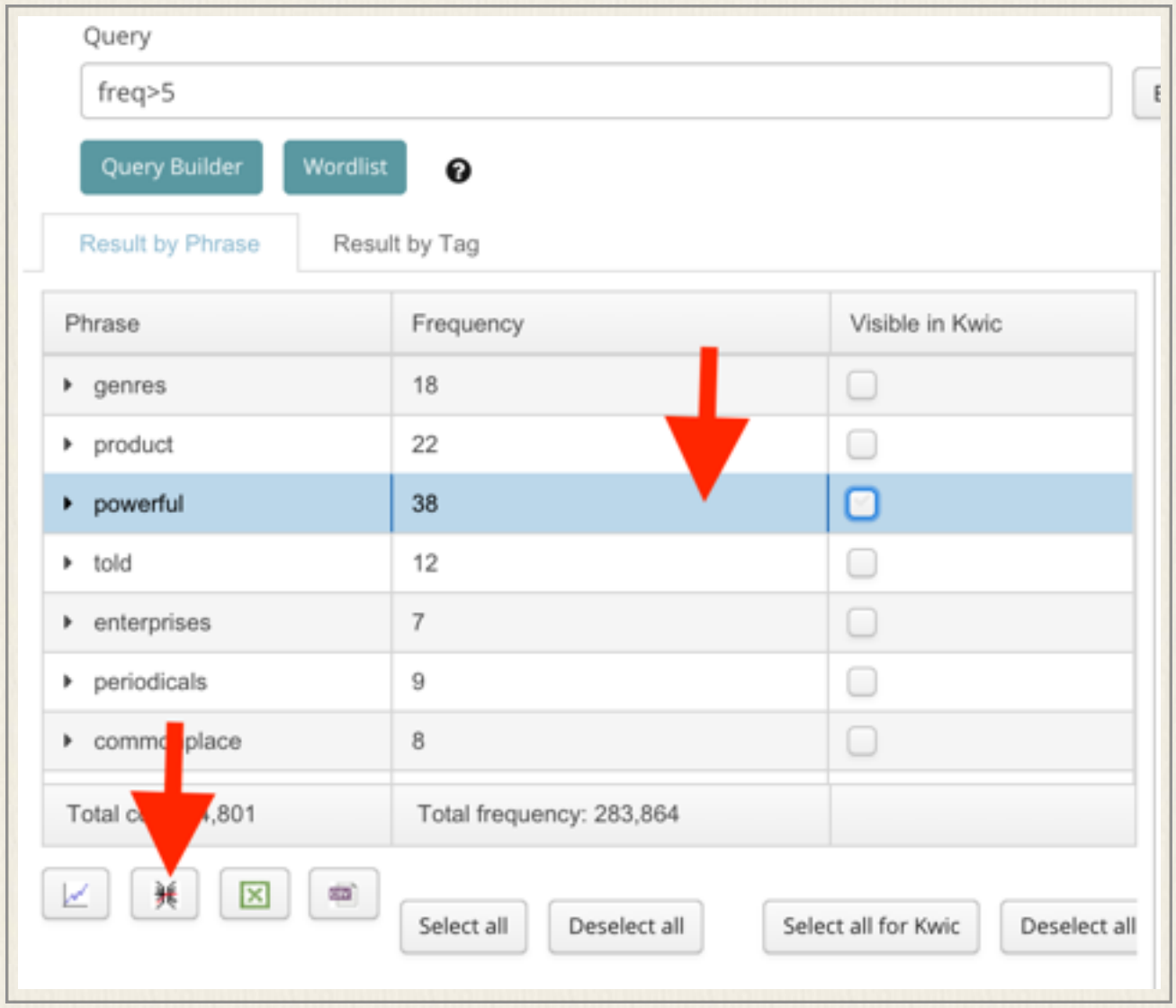
ในขั้นแรกสามารถเลือกจาก wordlist เพื่อจะได้เห็นว่าใน document นั้นมีคำอะไรมากน้อยแค่ไหน เลือกคำที่ต้องการ ก็จะเป็นคำนั้นพร้อมบริบทซ้ายขวา (kwic) ในหน้าจอด้านขวา เราสามารถใช้ขั้นตอนนี้กำกับข้อมูลคำหรือวลีที่เห็นให้เป็นแท็กที่ต้องการด้วยการเลือก Annotate selected result โดยเลือกแถวที่ต้องการกำกับข้อมูลแท็ก แล้วลากแท็กที่ต้องการมายังหน้าต่าง kwic นั้น ตัวคำค้นก็จะถูกแท็กด้วยสิ่งที่ลากมา ทำให้ไม่ต้องเสียเวลากำกับคำหรือวลีนั้นทีละคำก็ได้ หรือจะใช้ประโยชน์เป็นเหมือน concordance โปรแกรมดึงคำพร้อมบริบทออกมาใช้ก็ได้

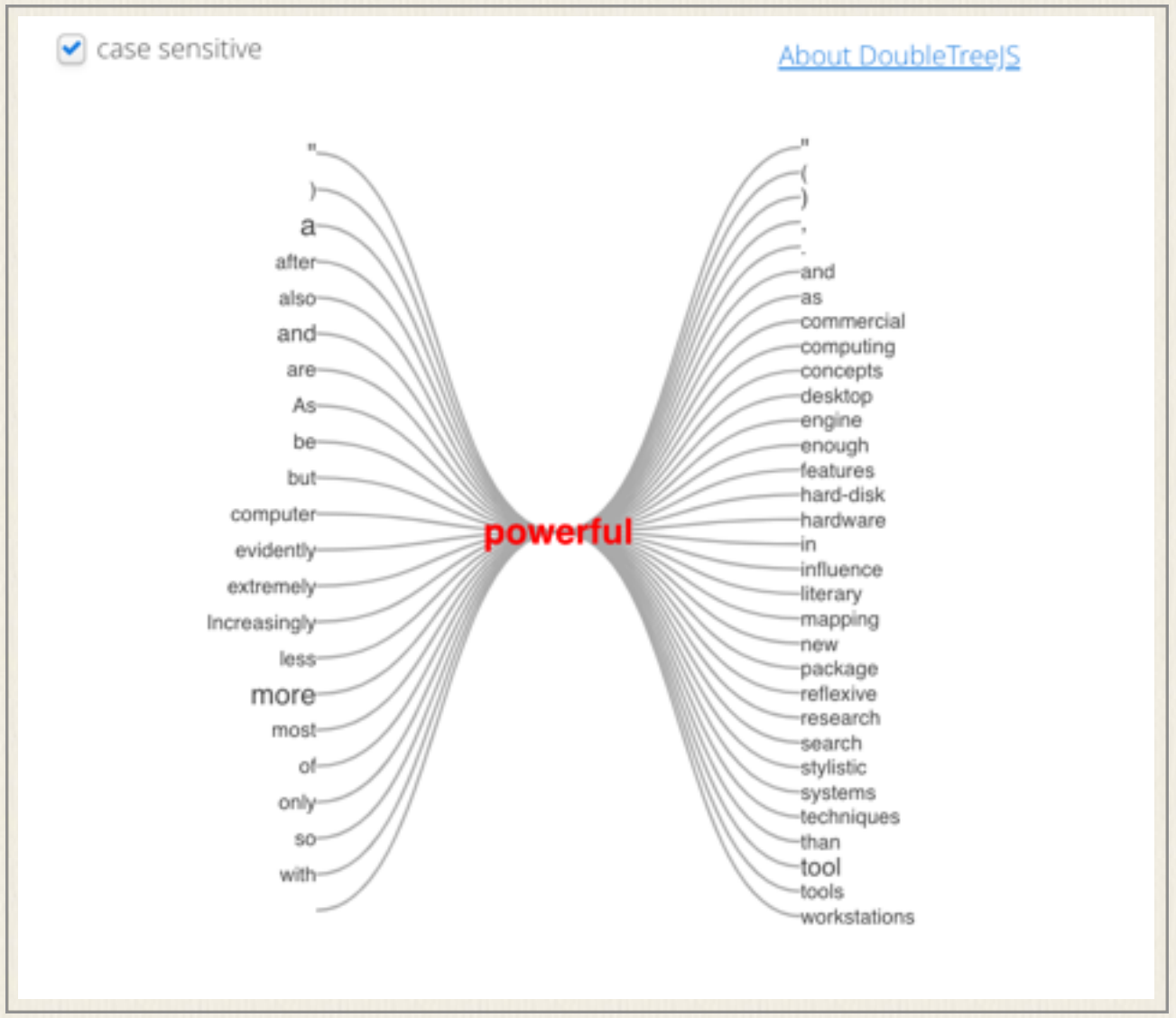



การค้นให้ใช้ Query Builder ได้  แล้วเลือกค้นคำหรือค้นแท็ก หรือจะค้นคำเขียนคล้ายคลึง คำปรากฏร่วมกัน (collocation) ก็สามารถระบุผ่านการใช้ Query Builder ได้




ส่วนสุดท้ายคือ Visualize ซึ่ง ณ ปัจจุบันยังไม่มีอะไรมาก  มีเพียงการแสดงผลของการกระจายความถี่ของคำ  และข้อมูลการปรากฏของคำติดกัน (bigram) ในรูปของ double tree  โดยเราต้องเลือกคำที่ต้องการก่อน  แล้วกดไปคอนที่อยู่ข้างล่าง ถ้ามีความต่างของความถี่มากจะเห็นความใหญ่เล็กของคำชัดเจนขึ้น


Query
freq>5
Query Builder
Wordlist
Result by Phrase
Result by Tag
Phrase
Frequency
Visible in Kwic
genres 18
product 22
powerful 38
told 12
enterprises 7
periodicals 9
8
Total frequency: 283,864
Select all
Deselect all
Select all for Kwic
Deselect all


case sensitive
About DoubleTreeJS
"
)
a
after
also
and
are
As
be
but
computer
evidently
extremely
Increasingly
less
more
most
of
only
so
with
powerful
"
(
)
,
.
and
as
commercial
computing
concepts
desktop
engine
enough
features
hard-disk
hardware
in
influence
literary
mapping
new
package
reflexive
research
search
stylistic
systems
techniques
than
tool
tools
workstations

โปรแกรม CATMA เหมาะสำหรับการกำกับข้อมูล เพื่อนำข้อมูลที่กำกับนั้นมาหา pattern หรืออีกที่เรียกในเว็บอธิบายโปรแกรมว่าคือ model คือการพยายามหาข้อสรุปบางอย่างจากการวิเคราะห์ข้อมูลที่กำกับไว้ เป็นงานเน้นด้าน qualitative literary หรือ text studies ที่อาศัยการวิเคราะห์ text ด้วยผู้วิจัยมากกว่าจะเป็นงานเน้นข้อมูลมาก ๆ แบบ text mining และเนื่องจากการกำกับข้อมูลอาศัย tagset ที่สร้างขึ้นได้ตามที่เรากำหนด เราจึงสามารถกำหนด tageset หลายชุดกำกับข้อมูลเดียวกันในระดับต่างๆ ได้

## การแปลงข้อมูล XML annotation

การกำกับข้อมูลด้วย XML เป็นมาตรฐานที่ได้รับการยอมรับในการใช้กำกับคลังข้อมูลภาษา จึงมีโมดูลหรือเครื่องมือที่ใช้ในการตรวจสอบการกำกับว่า valid หรือไม่ ดังนี้

online tool ที่เป็น XML Validator เช่น

https://www.liquid-technologies.com/online-xml-validator
http://xmlvalidator.new-studio.org/

ถ้าต้องการตรวจสอบมากกว่าหนึ่งไฟล์ ให้เขียนโปรแกรมโดยใช้ Python library เช่น

```
import xml.etree.ElementTree as ET
tree = ET.parse('country_data.xml')
root = tree.getroot()
```

ถ้าไฟล์ไม่ valid ET.parse จะไม่ผ่านและเกิด error ขึ้น หรือจะอ่านข้อมูลทั้งไฟล์มาเป็น string เช่นเก็บใน intxt ก่อน แล้วแปลงเป็น tree ด้วย ET.fromstring ก็ได้ สามารถใช้วิธีนี้ถ้ารู้แน่ว่าไฟล์เดิมมีส่วนของข้อมูลที่ไม่ valid ต้องแปลงข้อมูลภายในก่อน

```
parser = ET.XMLParser()
root = ET.fromstring(intxt, parser=parser)
```

นอกจากนี้ ยังมี Python library “xmltodict” ที่สร้างแปลง xml tree ที่ได้จากการ parse ทำเป็น dictionary ใน Python แล้วแปลงเป็น json ด้วย json.dump ได้ทันที สามารถเขียนข้อมูลที่ได้เป็นไฟล์ json ซึ่งเป็น annotation ที่นิยมในการใช้กับโปรแกรมคอมพิวเตอร์ ทำให้ได้ข้อมูล annotation ที่เป็น json ไปใช้ต่อได้

```
x = xmltodict.parse(intxt)
j = json.dumps(x, ensure_ascii=False)
        ## option “ensure_ascii=False” is to prevent writing Thai utf8 character as \uxxx
File.write(j)
```

แต่ข้อมูลที่ต้องการแปลงระหว่าง XML กับ JSON ทีไม่มาก ก็สามารถใช้ online tools ต่างๆ ที่มีได้ เช่น http://www.utilities-online.info/xmltojson/

## โปรแกรม ELAN

เป็นโปรแกรมสำหรับใช้กำกับข้อมูลเสียงหรือวีดิโอ เป็นโปรแกรมฟรีผลิตโดย Max Planck Institute for Psycholinguistics. https://tla.mpi.nl/tools/tla-tools/elan/

CHAPTER 5

# การดูแลรักษาข้อมูล

การดูแลรักษาข้อมูล

ข้อมูลงานวิจัยไม่ว่าจะเป็นข้อมูลดิบ ข้อมูลที่วิเคราะห์แล้ว ข้อสรุปหรือบทความ วิทยานิพนธ์ รายงานวิจัย ถือเป็นสิ่งสำคัญที่จะต้องปกปักรักษาตลอดช่วงของการทำวิจัย เพราะหากมีการสูญหายของข้อมูลไม่ว่าจะในขั้นตอนใด ก็อาจทำให้งานวิจัยนั้นล้มเหลวไม่สำเร็จตามเวลาที่กำหนด

# การสำรองข้อมูลและการรักษาความปลอดภัย

การสำรองข้อมูลและการรักษาความปลอดภัย

การสำรองข้อมูล เป็นกิจวัตรที่ผู้วิจัยควรต้องกระทำอย่างสม่ำเสมอทุกวัน หรืออย่างน้อยทุกสัปดาห์ ข้อมูลควรจะต้องมีการสำรองไว้หลายชุด และเก็บไว้ในหลายที่ทั้งที่บ้าน ที่ทำงานหรือมหาวิทยาลัย หรือที่เป็น cloud storage เช่น dropbox, google drive หรือ onedrive หรือบางครั้งอาจสำรองข้อมูลเป็นหลาย ๆ version คือไม่ได้ลบไฟล์เก่าทิ้ง แต่เก็บไฟล์ข้อมูลใหม่ วิธีนี้จะทำให้มีข้อมูลก่าเมื่อสัปดาห์ก่อน เมื่อเดือนก่อน อยู่ตลอดเวลาการวิจัย ทำให้หากต้อการย้อนกลับไป ณ เวลาเดิมก็สามารถทำได้ หากจะทำเช่นนี้สิ่งที่จะต้องคำนึงคือการวางระบบให้ชัดเจนว่าข่อมูลชุดไหนเป็นของช่วงเวลาไหน

ในกรณีที่ข้อมูล (ทั้งที่เป็นข้อมูลภาษา ข้อมูลงานวิจัย ฯลฯ) มีจำนวนมาก อาจสะดวกกว่าถ้าใช้วิธีการ sync ข้อมูลแทนที่จะ copy หรือ backup ไฟล์ทั้งหมด โดยสามารถใช้โปรแกรม เช่น Synkron (http://synkron.sourceforge.net/) ซึ่งเป็นโปรแกรมฟรีสำหรับทั้ง PC และ Mac Allway Sync (http://allwaysync.com/) สำหรับ PC ซึ่งมีทั้งที่ฟรีและต้องซื้อ หรือโปรแกรมที่ขายอย่าง GoodSync ได้ http://www.goodsync.com/ แต่ทั้งนี้ต้องรอบคอบและระวังไม่ให้ข้อมูลเก่าทับข้อมูลใหม่ หรือลบข้อมูลใหม่ทิ้ง ควรทดลองใช้โปรแกรมกับข้อมูลทดสอบอื่นๆ ก่อนจนเข้าใจวิธีการใช้โปรแกรมดีแล้วจึงจะใช้กับข้อมูลจริงได้

การ backup ระบบคอมพิวเตอร์ก็จำเป็นกรณีที่เครื่องที่ใช้มีปัญหา จะได้ restore กลับได้โดยง่าย Windows restore เป็นการใช้งานพื้นฐานที่ช่วยให้เราย้อนระบบกลับไปยังวันที่ไม่มีปัญหาได้ แต่ก็อาจจะไม่สำเร็จในบางกรณี Time machine ของ Mac ใช้สำหรับ backup เครื่องเป็นระยะๆ เพื่อสามารถย้อนกลับไปวันก่อนจะมีปัญหาได้

การ clone หรือสร้าง backup image ของ drive ไว้เป็นอีกวิธีที่ช่วยได้ ทำให้สามารถ restore ทั้ง hard drive กลับมาได้เมื่อไม่สามารถแก้ไขปัญหาเครื่องได้ โปรแกรมประเภทนี้ ได้แก่ Norton Ghost, ที่เป็นโปรแกรมฟรีให้ใช้ก็มี เช่น Macrium Reflect Free Edition สำหรับ PC (http://www.macrium.com/ReflectFree.asp), สำหรับ Mac สามารถใช้โปรแกรม SuperDuper (http://www.shirt-pocket.com/SuperDuper/SuperDuperDescription.html)

อย่างไรก็ตาม ก่อนใช้งานโปรแกรมเกี่ยวกับการจัดการ partition ให้ทดลองให้แน่ใจว่าขั้นตอนเป็นอย่างไร ก่อนจะใช้งานกับข้อมูลที่สำคัญ ในการ backup ควรใช้ external harddisk ในการ backup drive เมื่อมีปัญหาก็สามารถ restore จาก external harddisk นั้นได้ โปรแกรมประเภทนี้จำเป็นจะต้องให้เราสร้าง CD สำหรับ boot เครื่องขึ้นมาเองเวลาที่ต้องการ restore ให้สร้าง boot CD แล้วเก็บไว้ในที่ที่ปลอดภัย หรือถ้าสามารถสร้าง backup บน usb harddisk แล้วกำหนดให้เครื่อง boot จาก usb drive ได้เลยก็จะสะดวก

# ไวรัสคอมพิวเตอร์

ไวรัสคอมพิวเตอร์

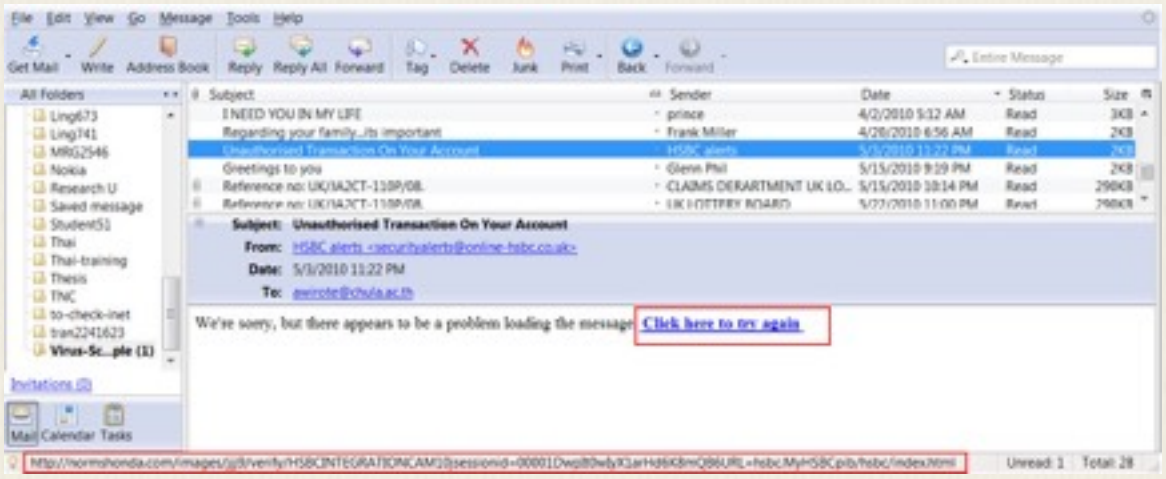
ผู้วิจัยควรมีโปรแกรม Antivirus ติดตั้งและต้องตั้งค่าให้ update ตัวเองอยู่ทุกวัน เครื่องที่ใช้จึงควรต่ออินเทอร์เน็ตอยู่เสมอเพื่อให้สามารถนำข้อมูลใหม่มา update ได้ โปรแกรมป้องกันไวรัสที่ได้แถมมากับเครื่องมักมีระยะเวลาให้ใช้ได้ช่วงหนึ่ง เช่น Norton, McAffee เมื่อพ้นกำหนดแล้วถ้าไม่จ่ายเงิน จะไม่ได้ update ตัวเอง จึงเป็นอันตรายที่ไวรัสใหม่ๆ จะเข้ามาได้ง่ายๆ หากเราไม่รู้และไม่ได้จ่ายเงินรายปีต่อ ทางที่ดีคือให้ใช้โปรแกรม Anti virus ที่มี free version ให้ใช้ อย่างน้อยหนึ่งตัว เช่น AVG (http://free.avg.com/), Avast (http://www.avast.com/free-antivirus-download), Avira (http://www.free-av.com/) และที่สำคัญ จะต้องคอย update โปรแกรมต่างๆ บนเครื่อง ไม่ว่าจะเป็น windows, Internet Explorer, Firefox, etc. เพราะโปรแกรมเก่าจะมีช่องว่างให้ไวรัสหลุดเข้ามาได้ง่าย

ไวรัสคอมพิวเตอร์เป็นโปรแกรมที่เขียนขึ้นมาเพื่อแฝงตัวในคอมพิวเตอร์และแพร่กระจายไปคอมพิวเตอร์เครื่องอื่นๆ ไวรัสสามารถแพร่ผ่านสื่อต่างๆ diskette, flash drive, CD, internet ผลกระทบที่เกิดจากการติดไวรัสมีความแตกต่างกันไปหลายแต่ว่าไวรัสนั้นถูกเขียนมาเพื่อให้ทำอะไร

โปรแกรมไวรัสมีหลายประเภท มีที่เป็นแบบไวรัสคือต้องอาศัยไฟลือื่นเป็น host ในการแพร่ ที่เป็นแบบ Trojan คือแฝงมาในโปรแกรมที่คิดว่านำมาใช้งานหนึ่งแต่แอบซ่อนโปรแกรมร้ายไว้ ที่เป็นแบบ worm คือกระจายตัวได้เองพยายามแพร่ตัวเองผ่านเตรือข่ายจนระบบล่มไป

ปัจจุบันไวรัสมาในรูปแบบต่างๆ ส่งผ่านอีเมล์หรือเว็บ เพื่อหลอกผู้ใช้ให้เปิดดู เมื่อเปิดก็จะถูกจู่โจมและติดทันที บางครั้งมาในรูปการหลอกว่าเครื่องเราได้ติดไวรัส หากต้องการ remove หรือป้องกันให้ install โปรกแกรมป้องกันที่ให้โดยด่วน วิธีป้องกันคือไม่เปิดไฟล์ที่ไม่ใช่โปรแกรมที่เรารู้จักและได้ติดตั้งเองในเครื่อง ไม่เปิดไฟล์คนแปลกหน้า หรือแม้แต่ไฟล์ที่ส่งมาจากคนรู้จัก หากเนื้อความ

ในจดหมายไม่มีข้อความอะไรเป็นสาระมากพอให้สงสัยว่า อีเมล์นั้นถูกส่งมาด้วยโปรแกรมไวรัสโดยอาศัยรายชื่ออีเมล์ที่คนนั้นเก็บไว้ในเครื่อง




อีเมล์หลอกมักจะมี link ประหลาดให้กด ให้ตรวจสอบดู address ที่เห็นให้ดีก่อนที่จะกด link ไป เพราะหากกดไปแล้ว บางครั้งโปรแกรม Antivirus ที่มีอยู่ก็ไม่สามารถกำจัดไวรัสนั้นออกไปได้ บางกรณี หากสังเกตุจากเนื้อความก็จะเห็นว่าเป็นจดหมายหลอกดังเช่นข้อความข้างล่างนี้ที่อ่านดูจะเห็นว่าเป็นภาษาไทยที่ได้จากการใช้เครื่องแปลหลอกให้ผู้ใช้ให้ข้อมูล user name และ password

*เราอยากจะแจ้งให้ทราบว่าขณะนี้เรากำลังดำเนินการบำรุงรักษาที่กำหนดและอัพเกรดการให้บริการเว็บของเราและเป็นผลจากการนี้ HTK4S ไวรัสได้รับการตรวจพบในโฟลเดอร์ที่บัญชีของคุณและบัญชีของคุณจะต้องมีการอัพเกรดใหม่ของเรา F-Secure รุ่น HTK4S anti-virus/anti-Spam 2013 เพื่อป้องกันไม่ให้เกิดความเสียหายต่อไฟล์สำคัญของคุณ เติมคอลัมน์ด้านล่างและส่งกลับมาหรือบัญชีอีเมลของคุณจะถูกระงับชั่วคราวจากการให้บริการของเรา*

**********

*ชื่อผู้ใช้: .....*

*รหัสผ่าน: .....*

*วันเดือนปีเกิด: .....*

********** *

*ล้มเหลวที่จะทำเช่นนี้ภายใน 24 ชั่วโมงทันทีจะทำให้บัญชีอีเมลของคุณปิดการใช้งานจาก chula.ac.th ฐานข้อมูลของเรา,*

*ลิขสิทธิ์ 2013 chula.ac.th*

*(ค) สิทธิเครือข่ายทั้งหมด*

อย่างไรก็ตาม การติดไวรัสอาจมาจากการใช้เครื่องสาธารณะที่มีผู้อื่นนำไฟล์ที่ติดไวรัสมาใช้ เมื่อเรานำเอา flash drive ไปใช้ก็สามารถติดไวรัสมาได้ flash drive จึงไม่ควรใช้เพื่อเก็บข้อมูลสำคัญอย่างถาวร ควรคิดเสมอว่าข้อมูลใน flash drive จะถูกลบเมื่อใดก็ได้

เมื่อติดไวรัสแล้ว ให้ลองใช้โปรแกรม Antivirus ที่มีอยู่ scan และทำลายไวรัสนั้น หากไม่ได้ผล ให้ลองใช้ windows restore ไปยังวันที่เครื่องไม่มีปัญหา หากยังไม่สามารถแก้ไขได้ ให้ใช้โปรแกรม restore image ของ hard drive กลับคืนไป

นอกจากเรื่องไวรัส การถูกจู่โจมโดย hacker ก็เป็นเรื่องที่อาจเกิดขึ้นได้ ทั้งการ hack เพื่อยึดเอา email, facebook, twitter ของเราไปใช้ ทำให้ข้อมูลต่างๆ อาจถูกทำลายได้ สิ่งสำคัญคือ ควรมีการเปลี่ยน password อยู่เสมอ อย่าใช้ password ชุดเดียวกันกับหลายๆ account อย่าตั้ง password ที่ง่ายเกินไป สามารถเดาได้ อย่าตั้งไว้สั้นเกินไป มีผู้แนะนำว่าให้ใช้ password ยาวแต่จำได้ง่าย อาจใช้คำในภาษามาผสมกันแต่ไม่เป็นวลีที่ใช้ในภาษาเพื่อช่วยให้จำง่าย เช่น “microwave someone Turkish” เป็นต้น

มีตัวอย่างนักคอมพิวเตอร์ ที่ถูก hack Google account, Twitter, Apple ID โดยอาศัยช่องโหว่ของระบบการตรวจสอบ โดยได้โทรศัพท์ไปที่ Amazon ขอเพิ่มเลขบัตรเครดิตใน account โดยมีข้อมูลชื่อ ที่อยู่ และ email เท่านั้น หลังจากนั้น โทรไปหา Amazon ว่าลืม password และใช้ข้อมูลชื่อ email และหมายเลขบัตรเครดิตที่เพิ่งเพิ่มเข้าไป เมื่อสามารถเข้า account Amazon ก็สามารถเห็นเลขสี่ตัวสุดท้ายบัตรเครดิตแท้จริง แล้วใช้เลขสี่หลักนี้โทรไป Apple เพื่ออ้างเป็นเจ้าของ AppleID และขอ password ใหม่ เมื่อยึด AppleID account ได้ก็สามารถใช้โปรแกรม Find myiPhone เข้าไปลบข้อมูลทั้งหมดใน iPhone, iPad และ MacNook จากนั้นใช้ข้อมูลนี้ขอ password ของ

Gmail ใหม่ และขอ password ของ Twitter ใหม่ได้ เนื่องจาก setup ให้ขอ reset password โดยส่งไปที่อีก email ที่เรากำหนดไว้ได้ (ดู http://www.wired.com/gadgetlab/2012/08/apple-amazon-mat-honan-hacking/)