# โปรแกรมและเครื่องมือ เพื่องานวิจัยภาษาศาสตร์

## วิโรจน์ อรุณมานะกุล



draft : ๒๕ ม.ค. ๒๕๖๓ ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

# บทนำ

เอกสารนี้ใช้ประกอบการสอนวิชาเทคนิคการวิจัยทางภาษาศาสตร์ ของภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย โดยเน้นที่เรื่องการจัดเก็บข้อมูล และการใช้เครื่องมือ หรือโปรแกรมต่างๆ เพื่อประโยชน์ในการวิจัย นอกจากการใช้ประกอบวิชาการวิจัย ในรายวิชาอื่น ๆ ที่เกี่ยวข้องกับการใช้คลังข้อมูลภาษาและการวิเคราะห์ ก็สามารถใช้ประโยชน์จากบทที่เกี่ยวข้องได้

เมื่อกล่าวถึงข้อมูลสำหรับงานวิจัย เราสามารถแยกเป็น ข้อมูลปฐมภูมิ (Primary sources) กับ ข้อมูลทุติยภูมิ (Secondary sources) ข้อมูลปฐมภูมิ คือข้อมูลดิบที่ นักวิจัยรวบรวมมาเพื่อใช้ในงาน วิจัย ส่วนข้อมูลทุติยภูมิ คือข้อมูลที่มีผู้เก็บรวบรวมหรือวิเคราะห์มาแล้ว ถ้าพูดถึงเอกสาร (document) ก็จะเป็น primary sources เป็นเอกสารที่ให้ข้อมูลดิบเพื่อการวิจัย ส่วน secondary sources เป็นเอกสารที่ให้ข้อมูลเกี่ยวกับ primary sources อีกทีหนึ่ง

สำหรับข้อมูลในงานวิจัยทางด้านภาษาศาสตร์อาจจะแยกข้อมูลออกเป็นข้อมูลภาษา กับ ข้อมูล งานวิจัย ข้อมูลภาษาอาจแยกเป็นข้อมูลประเภทต่างๆ เช่น ข้อมูลภาษาเขียน ภาษาพูด, ภาษาปัจจุบัน ภาษาเก่า, ภาษามาตรฐาน ภาษาถิ่น, เป็นต้น ข้อมูลภาษามีความสำคัญมากเพราะเป็นตัวแทนของ ปรากฏการณ์ทางภาษาที่เราต้องการศึกษา จึงต้องมีปริมาณที่มากพอจะศึกษาได้และเป็นตัวแทนของ ภาษาที่ต้องการศึกษาอย่างแท้จริง งานวิจัยจำนวนมากจึงอาศัยการวิเคราะห์ข้อมูลจากคลังข้อมูล ภาษาที่เป็นตัวแทนของภาษาที่ต้องการศึกษา

ข้อมูลอาจเก็บจากเอกสารต่างๆ หรือในบางกรณีก็เป็นการเก็บข้อมูลภาคสนาม สิ่งที่ต้องคำนึง คือทำอย่างไรให้เก็บข้อมูลที่ต้องการให้ได้ตามจำนวนที่เพียงพอ ไม่มีผลกระทบจากปัจจัยเกี่ยวกับผู้ เก็บข้อมูลและวิธีการเก็บข้อมูล การควบคุมตัวแปรต่างๆ หากมีข้อมูลที่คนอื่นเก็บมาหรือสร้างไว้แล้ว เราต้องพิจารณาว่าข้อมูลนั้นตรงกับที่เราต้อการหรือไม่ มีรายละเอียดที่มาครบถ้วน และสามารถนำ มาใช้ได้โดยตรงหรือไม่ การทำแบบสอบถามผ่านทางอินเทอร์เน็ตก็เป็นวิธีการหนึ่งที่เราสามารถใช้ได้ หากต้องการสำรวจข้อมูลต่างๆ เพื่อใช้ในงานวิจัย

นอกจากข้อมูลภาษา ข้อมูลงานวิจัยก็เป็นสิ่งสำคัญ เป็นส่วนของการทบทวนวรรณกรรม ใน ระหว่างการสำรวจงานต่างๆ ผู้วิจัยจึงต้องเรียนรู้ที่จะเก็บข้อมูลงานวิจัยต่างๆ รู้จักการจัดการข้อมูล บรรณานุกรม การใช้โปรแกรม Endnote เพื่อทำบรรณานุกรม หากไม่มีโปรแกรม Endnote การใช้ โปรแกรม Zotero ร่วมกับ web browser ต่างๆ ก็เป็นเครื่องมือหนึ่งที่ช่วยในการเก็บระเบียนข้อมูล ระหว่างการค้นคว้าได้ดี ในเอกสารนี้จะกล่าวถึงการสร้างรายการบรรณานุกรมใหม่, การโหลดข้อมูล จากห้องสมุด, การเลือกรูปแบบ, การคัดลอกบรรณานุกรมใส่รายงาน โดยใช้โปรแกรม Endnote

การค้นข้อมูลงานวิจัย ปัจจุบันมีฐานข้อมูลจำนวนมากที่มหาวิทยาลัยรับเป็นสมาชิก ในจุฬาฯ หากเข้าไปที่เว็บห้องสมุดจะเห็น CU-reference database ที่เชื่อมไปยังฐานข้อมูลต่างๆ ทำให้ สามารถค้น บทคัดย่อ บทความวิจัย วารสาร หนังสือออนไลน์ นอกจากนี้ ยังมีแหล่งข้อมูลบน อินเทอร์เน็ตอีกเป็นจำนวนมากที่สามารถใช้ประโยชน์ค้นข้อมูลวิจัย เช่น <u>www.scribd.com</u>, google search, scholar.google.co.th,

นอกจากโปรแกรมที่ใช้ในการจัดการข้อมูลต่างๆ แล้ว โปรแกรมสำหรับใช้ในงานทาง ภาษาศาสตร์นั้นมีมากมาย ได้แก่ โปรแกรมวาดต้นไม้ โปรแกรมวิเคราะห์เสียง โปรแกรมสำหรับ ทำการทดลองทางจิตวิทยา โปรแกรมกำกับข้อมูลภาษา เป็นต้น การทบทวนวรรณกรรมจะช่วยให้เรา เห็นว่ามีเครื่องมืออะไรบ้างที่สามารถใช้ได้ และเครื่องมือไหนเป็นที่ยอมรับใช้กันแพร่หลายในสาขาที่ เราสนใจ โปรแกรมที่ไม่ได้ครอบคลุมในหนังสือเล่มนี้ และมีการใช้งานแพร่หลาย คือโปรแกรม Concordance เช่น AntConc ซึ่งมีกล่าวถึงไว้อยู่ในหนังสือ ภาษาศาสตร์คลังข้อมูล

นอกเหนือจากโปรแกรมสำเร็จรูปที่มีคนทำให้เราได้ใช้ในเรื่องต่าง ๆ บางครั้ง เราก็จำเป็นที่จะ ต้องเขียนโปรแกรมขึ้นมาเพื่อประมวลผลภาษาเอง ในหนังสือแยกต่างหากอีกเล่ม คือ R ในงานสถิติ

ii

และการทำเหมืองข้อความ ก็เป็นอีกคู่มือหนึ่งสำหรับผู้สนใจที่ต้องการเขียน code หรือใช้ code เพื่อ งานทางวิจัยได้ โดยเฉพาะในการคำนวณสถิติ การสร้างกราฟแสดงผล และการทำเหมืองข้อความ นอกจากนี้ สิ่งสำคัญในการวิจัยคือการเก็บรักษาข้อมูลให้ปลอดภัย การสำรองข้อมูล การรักษา ความปลอดภัยข้อมูลจึงเป็นสิ่งสำคัญและจำเป็น เพราะเราไม่รู้ว่าพรุ่งนี้จะเปิดคอมพิวเตอร์มาแล้ว ใช้ได้ไหม จึงต้องมีการสำรองข้อมูลที่ดี หากข้อมูลสูญหายไม่ว่าด้วยเหตุใด งานวิจัยนั้นก็ล้มเหลวก่อน จะถึงปลายทาง

# โปรแกรม Endnote

โปรแกรม Endnote เป็นโปรแกรมสำหรับช่วยทำบรรณานุกรม ข้อดีของโปรแกรมนี้คือสามารถ จัดรูปแบบบรรณานุกรมและการอ้างอิงตามที่แต่ละวารสาร หรือแต่ละสมาคมต้องการได้ โดยที่ผู้วิจัย ไม่ต้องนั่งพิมพ์หรือจัดรูปแบบใหม่ทุกครั้งที่ต้องการใช้ โปรแกรมยังสามารถคัดลอกข้อมูล บรรณานุกรมจากภายนอก เช่น ห้องสมุด ฐานข้อมูลงานวิจัยต่างๆ หรือจากโปรแกรมอื่นๆ ได้ โดยไม่ ต้องเสียเวลาพิมพ์รายการบรรณานุกรมเข้าไปใหม่

# การติดตั้งโปรแกรม EndNote

### การติดตั้งโปรแกรม EndNote

ให้เข้าไปโหลดโปรแกรมได้ที่ https://www.car.chula.ac.th/endnote.php เนื่องจากจุฬาฯ ได้ ซื้อสิทธิ์การใช้โปรแกรม Endnote X9 มาให้อาจารย์และนิสิตได้ใช้ จึงต้องขอ password จากเจ้า หน้าที่ตามที่ระบุไว้ในหน้าเว็บ (โปรแกรมนี้ให้ใช้สำหรับอาจารย์และนิสิตจุฬาเท่านั้น ห้ามมิให้เผยแพร่ แก่บุคคลภายนอก และ password จะเปลี่ยนทุกภาคการศึกษา) โปรแกรมมีทั้งที่เป็นเวอร์ชั่นสำหรับ Windows หรือ Mac SECTION 2

# การตั้งฐานข้อมูลใหม่

การตั้งฐานข้อมูลใหม่

ให้เลือก File – New ตั้งชื่อตามที่ตัวเองต้องการ เช่น research การสร้างข้อมูลบรรณานุกรม ใหม่

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



กรอกข้อมูลต่างๆ ใน field ที่เห็น หากเลือกประเภทรายการนั้นผิด ก็สามารถเลือกเปลี่ยน ประเภทใหม่ได้ภายหลัง ข้อมูลใน field ต่างๆจะไปอยู่ใน field เดียวกันของประเภทรายการใหม่

## การคัดลอกข้อมูลบรรณานุกรมจากห้องสมุดออนไลน์

### การคัดลอกข้อมูลบรรณานุกรมจากห้องสมุดออนไลน์

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

Name	Information Provider	
U Edinburgh	Library Catalogs	
U Georgia	Library Catalogs	
U Helsinki	Library Catalogs	
U London SAS	Library Catalogs	
U Michigan-Ann Arbor	Library Catalogs	
U NC-Chapel Hill	Library Catalogs	
U Oxford	Library Catalogs	
U Pennsylvania	Library Catalogs	
U Pittsburgh	Library Catalogs	
U Southern Calif	Library Catalogs	-
U Toronto	Library Catalogs	=
US Geological Survey	Library Catalogs	
Villanova U	Library Catalogs	-
Duick Search 👻		Find by •
★ Less Info:	Cancel	Choose
File Name: U Pennsylvania.enz Created: Friday, July 16, 2010,	2:10:27 PM	
Modified: Tuesday, October 28,	2008, 5:53:00 AM	
Based On: Voyager		
Based On: Voyager Category: Library Catalogs		

เมื่อเชื่อมต่อได้แล้ว ให้เลือกค้นรายการตามปกติ จะค้นตามชื่อผู้แต่ง ชื่อเรื่อง หรือคำสำคัญก็ได้ หากค้นพบรายการที่ต้องการจะมีหน้าจอบอกจำนวนรายการที่ค้นได้ออกมา

Confirm Online Search

รายการที่ค้นมาได้จะถูกเพิ่มเข้าไปในฐานข้อมูลปัจจุบันที่ใช้อยู่

🚱 💗 🌑 APA Sh	🔜 🗟 🔕 🔮 🚱 🥞 🦉 🐷 🍯 🚱 Ouick Search	
My Library ^	8 Author Year Title	
All References (52) Ounfiled (52) Trash (0) My Groups Online Search QWeb of Scien (0)	Verhaar1995Toward a reference grammar of Tok HSvartvik1992Directions in corpus linguistics :Aijmer1991English corpus linguistics : studieAarts1990Theory and practice in Corpus linguKytö1988Corpus linguistics, hard and soft :Meijs1987Corpus linguistics and beyond : proAarts1986Corpus linguistics II : new studies	p p p p p p p p
Q PubMed (NLM) (0) Q LISTA (EBSCO) (0) Q LISTA (EBSCO) (0) D LIBRARY OF C (0) more	Frevenew     Online Search - Library Catalog at University of Pennsylvania       Search     Options +       Author        • Contains       •        And     • Year	

สามารถลบรายการที่ไม่ต้องการจากฐานทิ้งด้วยการเลือกย้ายรายการนั้นลงขยะ หรือกดปุ่ม Ctrl+D

หรือ DEL

research.eni		
🚱 😺 🌑 🔤 APA Sh		Duick Search
My Library ^ All References (52) Dufiled (52) Trash (0)	0         Author         Year         Title           Morley         2009         Corpus-assisted discourse :           Balasubrama         2009         Register variation in India           Lindquist         2009         Corpus linguistics and the	studies on an English descripti
<ul> <li>My Groups</li> <li>Online Search</li> <li>QWeb of Scien (0)</li> <li>QU Pennsylvania (52)</li> </ul>	Quadlio     Record Summary_     prover the structure       Romero-Trillo     New Reference     prover ling       Dash     Edit References     ts : an int       Hundt     Move References to Trash     ts and the       *     Add References To     *	sitcom Fri uistics : troduction web
Q PubMed (NLM) (0) Q LISTA (EBSCO) (0) Q Library of C (0) more	Preview         Online Search - I         Copy References To         *           Search         Optio         Copy         mote Library           Author         Paste	• •
transfer Find Full Text howing 52 of 52 references in Group.	And Vear Show All References And Title Hide Selected References Hide Selected References Hide Tab Pane	istics Hide Tab Pane

สามารถจัดเรียงข้อมูลตาม field ต่างๆได้ โดยการกดที่ชื่อ field นั้น เช่น กด title จะเป็นการเรียงตาม ชื่อเรื่อง

0	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
	Dach	2008	Cornel linguistics . an introduction		Raak

หากต้องการค้นหนังสือเราสามารถเลือก Tools - Online Search แล้วเลือกห้องสมุดต่างๆ ที่เปิดให้ บุคคลภายนอกค้นผ่านอินเทอร์เน็ตได้ แต่หากต้องการค้นบทความในวารสารให้ไปค้นจากฐานข้อมูลที่ มหาวิทยาลัยบอกรับไว้ เช่น Web of Sciences (ISI) ก็จะได้รายการบทความต่างๆ มาเก็บใน End-Note ทันที

a research.eni		
🚱 😝 🌒 🗛Sh		
Ny Library All Refer (154) ByUnfiled (154) Trash (1) Ny Groups Coline Search Callo Search (103)	# Author Year Title Jo Tambovtsev 1995 Corpus Linguistics and the Automatic Analysi Wo Futrelle 1995 Corpus linguistics for establishing the natu Di Nettemann 1994 Directions in Corpus Linguistics - Proceedin An Oostdijk 1994 Directions in Corpus Linguistics - Proceedin St Biber 1994 Corpus-Based Approaches to Issues in Applied An Mair 1993 English Corpus Linguistics - Aijmer.K. Alten An Jager 1993 Corpus Linguistics and the Automatic-Analysi En	arnal Ref Ty rd-J Journs gita Journs glia Journs udie Journs olio Journs glis Journs glis Journs
Q(U Pennsy (S1) Q(PubMed (NLM) (0) Q(LISTA (E8 (0) Q(Library o (0) more EndNote Web transfer Find Full Text	Prevery Define Search - Bil Chance Indexes at web of Science (SQ) Biber, D., Conrad, S., & Reppen, R. (1994). Corpus-Based Approaches to Issu Linguistics. Applied Linguistics, 15(2), 169-189.	es in Applied
howing 103 of 103 references in	Group. (All References: 154)	# Hide Tab Pane

หาก click เข้าไปดูในแต่ละรายการ เราจะเห็นว่ามี link ที่จะโยงเราไปยังเว็บ ISI ได้เลย

Gavioli,	2008 #69								0	۲	*
00	Reference T	ype [	Journal Article	-				*	fide En	pty Fie	olds
9 🐂		3	Plain Font	▼ Ptain Size	٠	в	1	U	P	$\mathbf{A}^{i}$	Α,
											*
Research	Notes										
URL											
<00 to 1	tachments	026283	8900011								
File Act											
Author J	Iddress										
Author 3	lddress										1
Author J Figure	kddress										and the second s
Author A Figure Caption	Address										



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





ฐานข้อมูลต่างๆ ตลอดจนวารสารต่างๆ ที่สามารถสืบค้นได้ มักจะมีทางเลือกให้เรา export ข้อมูล

บรรณานุกรมออกมาเข้า EndNote ได้ ตัวอย่างเช่น





กรณีที่ export เข้า EndNote แบบอัตโนมัติไม่สำเร็จ ให้เลือก save เป็นไฟล์ไว้ ไฟล์ที่ได้ควรมี นามสกุล .enw ถ้าไม่ได้ไฟล์แบบนี้ให้ลองเปลี่ยนนามสกุลเอาเองเป็น .enw เมื่อเปิดไฟล์นี้ ข้อมูลจะ ถูกนำเข้า EndNote ให้เอง

## การเลือกแบบบรรณานุกรม

#### การเลือกแบบบรรณานุกรม

เราสามารถเลือกแบบบรรณานุกรมที่เราต้องการได้ ขึ้นอยู่กับว่าต้องการทำรายการอ้างอิงเพื่อตี พิมพ์ในวารสารใด เช่น Language, IEEE Proceedings ฯลฯ ให้เลือก select another style แล้ว เลือกรายการแบบที่ต้องการ กดปุ่ม Style Info/Preview เพื่อดูตัวอย่างที่ออกมาว่าใช่แบบที่ต้องการ หรือไม่



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 ซึ่งอาจถูกลบทิ้ง ไปหลังการติดตั้ง หากเป็นเช่นนั้น ให้ install โปรแกรม EndNote ใหม่ และในระหว่างติดตั้งนี้ ให้ดูว่า temporary folder อยู่ที่ไหน ให้เข้าไป copy ไฟล์นี้ออกมาก่อนที่จะถูกลบทิ้ง จากนั้นจึงเริ่มขั้นตอน การ modify โปรแกรมใหม่

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

		EndNote X	5 Customi	izer
			Compo	
	CONTRACTORY CONTRA	Component Cite While You Write Services Spotlight ► Connections ► Import Filters ■ Output Styles	Installed	Info Cite While You Write addin for Microso Services for Mac OS X Spotlight Indexing Additional Connection Files Additional Import Filters Additional Bibliographical Output Style
(	? Cancel	Uninstall		Back

สไตล์ต่างๆ ที่เลือกใช้สามารถ preview ดูได้ว่าใช่แบบการอ้างอิงที่ต้องการหรือไม่ โดยเลือกแท็บ preview จะเห็นรายการสิ่งที่ถูกเลือกไว้

Cargage Coups  All References (8  Trash (3  References Coups	0 ( 0) 3)	<b>0 0</b>	Author Jones	Year 2002	Cuck Search
Groups All References (8 Trash (3 References Creations)	(0) 3)	14° 0	Author Jones	Year 2002	Title -
All References (8 Trash (3	90) 3)		Jones	2002	A set a second set of a second set of a second set of the
- Smart Groups			Renoul Nesselhauf Cresti Adel Aston Lindquist	2004 2006 2005 2005 2008 2004 2004	Antonymy : a corpus-based perspective Applied corpus linguistics : a multidimensional perspective The charging face of corpus linguistics Colocations in a learner corpus Corpora and language reference corporator spoken Romance I. Corpora and liscourse : the challenges of different settings Corpora and language learners Corpus approaches to grammaticalization in English
Online Search     Library of Congr (I     LISTA (EBSCO) (I     PubMed (NLM) (I     U Pennsylvania (8)	0) 0) 0) 30)	Revie Nos	search sehauf, Nadja, 2 Pub. Co.	005.Col	Iocations in a learner corpus Amsterdam ; Philadelphia: J. Benjamins

# 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 ทำแบบนี้กับทุก รายการที่ต้องการใช้สำหรับ paper1 (สามารถ add หลายรายการพร้อมกันก็ได้ ให้เลือกทีเดียวหลาย รายการ กด Ctrl+ mouse click)

My Librafy          • Arash          My Librafy          • Author       Year         Trash       (0)         My Groups          • Author       Year         Danchoz-St.       2007       Compounding construction in Thair         Perreira       2007       Conceptual metaphor and corpus lip         My Groups          • Author       2010       Corpus and soci         Panchoz-St.       2007       Corpus linguist       Record Summary         Nurphy       2010       Corpus linguist       Record Summary         Nurphy       2010       Corpus linguist       New Reference         Colucker       2007       Corpus linguist       Record Summary         New Beferences       To Add References To       Copy References To       Cot         Queb of (0)       Sanchez-Stockhammer, C. (2009). Corport       Copy References To       Cut         Quitstrat ( (0)         Sanchez-Stockhammer, C. (2009). Corport       Cut       Copy         Copy Copy Formatted       Patte       Show All References       Show All References       Show Statested Before		x
My Librafy  All Ref (156) All Ref (156) All Ref (156) All Ref (156) Trash (0) My Groups Dapaper1 (0) Conceptual metaphor and corpus linguist Conceptual metaphor and corpus linguist Corpus based te Schlucker 2007 Corpus based te Schlucker 2007 Corpus linguist Copy References To Copy References Copy Ref	ok Search	
Trash       (0)         My Groups       (0)         Murphy       2010       Corpus and soci         Teich       2007       Corpus based te         Schlucker       2007       Corpus linguist         Move References to B       Move References to B         Output       Based Approaches to Syntax and Le         Output       Gate Approaches to Syntax and Le         Output       434-448.	Its contri nguistics	JC * Et E
Q.Web of (0)       Preview Search       Copy References To         Q.U Penns (0)       Sanchez-Stockhammer, C. (2009). Corport       Cut         Q.LISTA ( (0)       Based Approaches to Syntax and Lt       Copy         Q.LIbrary (0)       Based.       Sanchez-Stockhammer, C. (2009). Corport       Cut         Based.       Add References To       Cut       Copy References To         Based.       Add References to Syntax and Lt       Copy Copy Formatted         Based.       Show All References       Show All References	g of	St Ze Ze
QU Penns (0)       Sanchez-Stockhammer, C. (2009). Corpora       Cut         QPubMed (0)       Based Approaches to Syntax and Le       Copy         QLISTA ( (0)       434-448.       Copy formatted         Patte       Show All References       Show All References	• Cre	Genues
BODE Show All References	P	aper1
EndNote Web  Transfer  Hidd Selected Referen	ices	

รายการที่เลือก add เข้า group ก็จะยังคงอยู่ใน All references เราสามารถลบรายการที่ไม่ ต้องการใน group ออกได้ โดยไม่มีผลไปการลบรายการออกจากฐานข้อมูลหลัก ถ้าต้องการลบจาก ฐานข้อมูลหลักต้องเข้าไปที่ All references แล้วเลือกรายการที่ต้องการลบ Group ใดที่เมื่อเลิกใช้แล้ว ก็สามารถลบทิ้งได้ เพราะจะไม่มีผลต่อรายการที่เก็บในฐานข้อมูล หลัก (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 ที่ถูก ต้องทันที

## 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 หลังจาก เปิดทำงาน 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

				_					
0	0	2	6	311	0	0	Quick Search	*	
64	0	Author		Ye	ar	Title	*		
		Aiimer		200	04	Advar	nces in corpus li	nauistics :	papers from the 2

จะได้หน้าจอที่ให้เราเลือกไฟล์ Word ที่เราต้องการทำงาน และเลือกแบบบรรณานุกรมที่ต้องการ

ormat bibliography	Layout	Instant Formatting	Libraries Used
Format document:	paper-1	.doc	
With output style:	Langua	ge in Society	Browse
Temporary citation	n delimiter	\$	
Left: {	Right:	}	

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

Citation



Citation และ reference ตาม format style ที่เราเลือกจะถูก paste ลงในไฟล์ Word ที่เราระบุไว้

#### ใน Format Bibliography



เมื่อพิมพ์งานต่อไปแล้วมีการอ้างอิงรายการเพิ่มก็ทำต่อได้เรื่อยๆ จนจบบทความ รายการอ้างอิงต่างๆ ก็จะถูกนำมาใส่ในบทความนั้น ตรงตาม format ที่เราได้เลือกไว้

Thal National Corpus (TNC) is a general corpus of the standard Thal 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)

## การย้ายข้อมูลระหว่างฐานข้อมูล

### การย้ายข้อมูลระหว่างฐานข้อมูล

เราสามารถเปิดฐานข้อมูลมากกว่าหนึ่งฐานขึ้นมาใช้ได้ และสามารถ copy รายการจากฐานหนึ่ง ไปยังอีกฐานหนึ่งได้ ให้เลือกรายการที่ต้องการ แล้วกดลากไปยังอีกฐานหนึ่ง



# โปรแกรม Zotero

โปรแกรม Zotero เป็นโปรแกรมช่วยจัดเก็บข้อมูลบนเว็บที่พบ เดิมใช้กับ Firefox ทำให้ สามารถคัดลอกข้อมูลบรรณานุกรมจากเว็บได้โดยตรง 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 <u>http://www.zotero.org</u>/ เพื่อ โหลดโปรแกรม addon Zotero นี้

Zotero สามารถทำสิ่งต่างๆ เหล่านี้ได้ขณะที่ browse internet : ดึงเอา citation ออกมาโดย อัตโนมัติ เก็บไฟล์ pdf, web และรูปภาพ สร้าง library ของตัวเองโดยการเก็บข้อมูลและงานต่างๆ ที่ พบระหว่างการใช้อินเทอร์เน็ต

## การติดตั้งโปรแกรม Zotero Standalone

### การติดตั้งโปรแกรม Zotero Standalone

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

 Open control
 Zotero

 Open control
 Open control</

## 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 จะสังเกตุ ว่าที่ address bar ด้านขวาจะมีไอค่อนรูปหนังสือ หมายความว่าสามารถนำข้อมูลบรรณานุกรมมา เก็บใน

Zotero ได้ ให้กดที่ไอค่อนนี้

Interview       + Pocket       WebMail - CU       M Gmail       Facebook       User Ident       Googie Ngram         Main Database Chulaiongkorn University Library and Information Network       Image: Chulaiongkorn University Library and Information Network       Image: Chulaiongkorn University Library and Information Network         Image: Chulaiongkorn University Library and Information Network       Image: Chulaiongkorn University Library and Information Network       Image: Chulaiongkorn University Library and Information Network	
Main Database Chulalongkom University Library and Information Network	
Start Over Add to My Lists Save Records UKAC Diplay Another Search	
RECORD # b1618707 Search Limit search to available items	
Author Kennedy, Graeme D Title An introduction to corpus linguistics / Graeme Kennedy	
Imprint London and New York : Longman, 1999, c1998	
Arts Library P98 K35I CHECK SH	ELVES

เมื่อกดที่ไอค่อน Zotero ตรงมุมขวาของ Firefox จะเห็นหน้าจอ Zotero ขึ้นมาในอีกแท็บ และ รายการหนังสือนี้จะถูกเก็บใน Zotero แล้ว



เราควรสร้าง collection เฉพาะสำหรับงานวิจัยแต่ละเรื่อง ที่มุมซ้ายบนของ Zotero มีไอค่อน New Collection ให้ตั้งชื่อ collection ของงานตนเอง แล้วย้ายรายการข้อมูลที่ต้องการไปยัง collection ที่ สร้างขึ้น

My Library จะเก็บรายการข้อมูลทั้งหมด item เดียวกันสามารถไปอยู่ในหลายๆ collection ได้ collection จึงไม่ได้เก็บข้อมูลรายการจริง แต่คล้ายๆ กับ playlist ใน media player ที่เราเลือกเฉพาะ ชื่อรายการมาจัดเป็น collection หนึ่งๆ แต่ตัวรายการจริงเก็บไว้อีกที่ เวลาที่ต้องการให้ข้อมูลไปอยู่ใน collection ไหน ก็ให้เปิด Zotero ค้างไว้ที่หน้า collection นั้น เมื่อมี การเก็บรายการข้อมูลใหม่ ชื่อรายการนั้นจะไปอยู่ใน collection ที่เราเลือกค้างไว้เอง



เว็บที่เกี่ยวข้องกับรายการหนังสือ วารสาร เช่น Amazon เว็บห้องสมุดต่างๆ เมื่อเราค้นเจอรายการที่ ต้องการ โดยปกติเราจะเห็นไอค่อนรูปหนังสือให้ save รายการเก็บไว้ได้ จึงค่อนข้างสะดวกต่อการเก็บ รายการอ้างอิงต่างๆ ไว้ใช้ภายหลัง

บางเว็บไซ้ต์ เราจะเห็นไอค่อนให้ save to Zotero เป็นรูปอื่นไม่ใช่รูปหนังสือ หมายความว่า เรา สามารถเก็บข้อมูลเว็บนั้นลงใน 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 ติด อยู่ด้วย ข้อมูลที่เก็บแบบ snapshot ไว้จะสามารถมาเปิดดูภายหลังได้ แม้ว่าไม่ได้ต่ออินเทอร์เน็ตอยู่



Title	^	Creator	@  E	Ę
	An introduction to corpus linguistics	Kennedy		
►	Data Scientist: The Sexiest Job of the 21st Cent		•	
►	IBM - What is a Data Scientist? – Bringing big d		•	
	IBM - What is a Data Scientist? – Bringing big d			

### ในโหมดที่ดู Sanpshot เราจะเห็นว่าข้อมูลนั้นเก็บที่เครื่องเรา ดูได้จาก address จะเป็น



้ตัวอย่างสาธิตการใช้งาน Zotero เพื่อเติมรายการข้อมูลต่างๆ ที่พบระหว่างการค้นผ่าน Firefox

## การย<sup>้</sup>ายข้อมูล 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

/tab.xul	Format: RIS		le
🛐 BKKnews + 🦳 + Pocket 🔮 W	Translator Options		t <u>8</u> Goog
ی کی دی	Export Notes		0
Title	Export Files		0 E
A Glossary of Corp			
Animacy and Refer	Canc	el OK	
An Introduction to (			
An introduction to corr	ous linguistics	Kennedy	
<ul> <li>Automatic corpus-bas</li> </ul>	ed Thai word extraction w	Sornlertlamvanich et al.	•
ACM Full Text PDF	-		•
<ul> <li>Contemporary translat</li> </ul>	tionese in Japanese popul	Meldrum	
One of the main ai	ms of this thesis is to exa		
Corpora and Discours	e: The Challenges of Diff	Adel and Reppen	
Corpora for Theory an	d Practice	Barlow	
Corpora in Cognitive L	inguistics: Corpus-Based	Gries and Stefanowitsch	
Corpus and Context: I	nvestigating Pragmatic F	Adolphs	
Corpus-Based Analys	es of the Problem-Solutio	Flowerdew	
Corpus-Based and Co	omputational Approaches t	Botley and McEnery	
V G Corpus-based Approa	ches to Contrastive Lingu	Lerot	
Amazon.com Link			
			TO DE CONTRACTOR

เมื่อ save ไฟล์ RIS แล้ว ให้เปิดไฟล์นั้น ข้อมูลต่างๆจะถูกนำเข้าในฐานข้อมูลของโปรแกรม End-Note ที่กำลังใช้งานอยู่ทันที

ข้อมูลที่ import มาจะแสดงแยกให้เห็น ข้อมูลหนังสือ หรือแม้แต่เว็บเพจ ก็จะถูกเก็บเข้าใน End-Note และนำไปใช้งานต่อได้



# 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

Preadlines - Nation - BEX.eev      December 2012     December	ro.org/styles ws + Pocket Web sitory r use with Zotero 2.1 (or you can create your own citation sty	Confirm Install style "Journal of Pragmatics" from http://www.zotero.org/styles/journal- of-pragmatics?install=1? Cancel Instal	y for styles compatib
Style Search Title Search Show only unique styles	Format: author-date T Fields: humanities	nguistics philosophy	
3 styles found: • Journal of Pragmatics (Instal • Unified Style Sheet for Lingu • University of Bologna - Liber	l] (2012-04-03 09:46:03) istics Journals [Install] (2012-04-0 al Arts College (Università di Bologn	3 09:46:03) a - Facoltà di Lettere e Filosofia) (Italian) (Install	(2012-05-19 00:46:07)

์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

General Sync Search Export Styles Proxies Storage Location © Use Firefox confile directory	a Shortout Keys	K
Storage Location	Shortout Keys	the second
Use Firefox profile directory		Advance
O Quatom:		hoose
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/

# โปรแกรม RapidMiner

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

การใช้งานคือการนำ Operator มาประกอบกันเป็นลำดับขั้นตอนที่ควรเป็น และเชื่อมโยง output 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 และให้นำ ข้อมูลออกที่ได้ไปใช้ต่อในกระบวนการถัดไป Operator พื้นฐานจะมาพร้อมโปรแกรมแล้ว แต่หาก ต้องการติดตั้งเพิ่ม ให้ไปที่ Extensions - Market Place เพื่อค้นหา operator เพิ่มเติม ตัวที่ควร เพิ่มเติมเพื่อใช้ในการประมวลผลข้อความ 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" แล้วระบุชื่อไฟล์ที่ ต้องการให้อ่านเข้ามา หรือถ้าต้องการอ่านไฟล์ทั้งหมดใน folder ให้ใช้ "Process Document from Files" แล้วระบุ folder และประเภทไฟล์ที่ต้องการให้อ่าน ตั้งชื่อ class สำหรับ folder นั้นๆ จาก นั้นภายใน "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 ที่ใช้สำหรับอ่าน ไฟล์เหล่านั้น เช่น "Read CSV", "Read Excel", "Read XML", "Read URL"

# การสร้างรายการคำและความถื่

Operator หลักที่ใช้ คือ "Read Document" กับ "Process Document" ให้ค้นจากรายการ เครื่องมือ แล้วลากกล่องนั้นมาที่หน้าจอออกแบบ ใน "Read Document" ให้เลือกไฟล์ที่ต้องการ ให้อ่านข้อมูล แล้วลาก output จาก "Read Document" ซึ่งเป็น doc ไปที่ input ของ "Process Document" ซึ่งเป็น doc แล้วเลือก output wor ซึ่งคือ wordlist ไปที่ res หรือ result ด้านขวา สุด 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)" โดย เลือกจำนวน n-gram ที่ต้องการ กระบวนการนี้จะสร้างข้อมูล n-gram ที่กำหนด รวมถึงข้อมูลย่อย ของ n-gram ที่ต่ำลงไปจนถึงคำเดี่ยวด้วย เช่น ถ้าให้สร้าง 3-gram จะได้ข้อมูลรายการ 3-gram, 2-gram และคำเดี่ยว โดยมีเครื่องหมาย "\_" ใช้เป็นตัวเชื่อมระหว่างคำใน n-gram

Process Documents			
	Tokenize	Generate n-Grams (Terms)	
doc	doc doc	doc doc	doc doc

หากต้องการสกัดเฉพาะ n-gram ที่ต้องการ ก็จะต้องกำหนดตัวกรองหรือ filter เอง เช่น ถ้า ต้องการ 3-gram ให้เติม operator "Filter Tokens (by Content)" ภายใน process นี้ให้ กำหนดว่าดึงเฉพาะรายการ token ที่ "contain match" ตรงตามค้นแบบ "\_.+\_" ซึ่งเป็น regular expression ที่จะ match กับรายการที่มีเครื่องหมาย \_ สองครั้ง



#### การทำ simple concordance

หากต้องการดึงตัวอย่างประโยคที่มีคำที่ต้องการจาก text สามาถทำได้โดยการใช้ ."Tokenize" โดย เลือก token แบบ "linguistic sentences" (ภาษาอังกฤษมีตัวเลือกนี้ให้ใช้) แล้วจึง "Filter Token (by Content)" โดยระบุคำที่ต้องการค้นเพื่อให้เหลือเฉพาะตัวอย่างประโยคที่มีคำนั้น หาก ต้องการนำผลที่ได้ไปเปิดใน Excel เพื่อใช้งานต่อ ก็อาจทำต่อโดยใช้ "Replace Tokens" และ แทนที่คำค้นด้วยคำเดิมแต่มีเครื่องหมายเฉพาะ เช่น | อยู่หน้าหลังเพื่อให้ไปเปิดใน Excel เป็นสาม คอลัมน์ได้ เช่น ถ้าต้องการค้นคำว่า digital ที่อักษรแรกเป็นตัวพิมพ์ใหญ่หรือเล็กก้ได้ ก็ใส่ replace what ด้วย "\s([d|D]igital)\s" และใส่ replace by ด้วย " \|\$1\| "



# การเก็บข้อมูลลงไฟล์

หลังจากสร้างรายการคำหรือรายการ n-gram แล้ว หากต้องการเก็บผลที่ได้เป็นไฟล์ สามารถทำได้ โดยการเรียก 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 ตัวอย่างข้างล่าง แสดงการเพิ่ม 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) ไป ยัง 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 ไม่ สามารถแสดงผลออกาหน้าจอได้ จึงต้องให้เก็บเป็นไฟล์ png ส่วนคำสั่ง dev.off บอกให้จบการนำ output ลงในไฟล์ ส่วนคำสั่ง wordcloud เป็นคำสั่งสำหรับสร้างภาพคำตามขนาดความถี่ที่พบ รายละเอียดการใช้งานจริง ให้ค้นคว้าต่อในเรื่องการเขียนโปรแกรมด้วยภาษา R ภาพที่ได้จากการใช้ คำสั่งนี้จะเก็บเป็นไฟล์ test.png และมีรายละเอียด ดังนี้



#### การใช้ Python module

RapidMiner มี operator Execute Python ที่ช่วยให้่เราสามารถเขียนโปรแกรมภาษา Python ประกอบการใช้งาน RapidMiner ได้ ตัวอย่างข้างล่างเป็นการใช้โมดูลเพื่อทำ pos tagging ข้อมูล ภาษาไทย โดยข้อมูลภาษาไทยเตรียมไว้เป็นไฟล์ csv มีหนึ่งคอลัมน์ชื่อ "content" แต่ละแถวเป็น ข้อความที่ต้องการตัดคำและใส่ 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 dataframe "data" นี้ ซึ่งในที่นี้มีคอลัมน์เดียวชื่อ "content" ส่วน data2 เป็น dataframe ที่สร้าง ขึ้นโดยนำข้อมูลจาก 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

```
1 import pandas as pd
 2 import tltk
 3
4 # rm_main is a mandatory function,
 5 # the number of arguments has to be the number of input ports (can be none)
 6 def rm_main(data):
7
8
      data2 = pd.DataFrame(data)
      data2['postag'] = ""
9
10
11
      for idx, row in data.iterrows():
           intxt = row['content']
12
           intxt = intxt.replace('"','')
13
           out = tltk.nlp.word_segment(intxt)
14 #
           lstx = tltk.nlp.pos_tag(intxt)
15
           out = ''
16
17
           for lst in lstx:
               for (w, pos) in lst:
18
                   out += w+'/'+pos+'|'
19
20
21
           data2.loc[idx].postag = out
22
       return data2
23
```

#### 1 "content", "postag"

- 2 "ลดน้ำหนักวิธีใหม่", "ลด/VERB|น้ำหนัก/NOUN|วิธี/NOUN|ใหม่/ADJ|<s/>/PUNCT|"
- 3 "คนจำนวนมากต้องการลดน้ำหนักโดยไม่เข้าใจศาสตร์ในเรื่องนี้อย่างลึกซึ่ง", "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|"
- 4 "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 หรือจัดกลุ่มข้อมูลที่มีความ คล้ายคลึงกัน ตัวอย่างข้างล่างเป็นการใช้ operator "K-mean" สำหรับทำ Clustering" ซึ่งเป็นวิธี การจัดกลุ่มที่นิยมใช้กันวิธีหนึ่ง โดยเราต้องกำหนดจำหนวนกลุ่มที่ต้องการ และกำหนดจำนวน รอบที่เครื่องทดลองการขัดกลุ่ม ผลที่ได้จะเป็นกลุ่มตามจำนวนที่กำหนดให้ โดยมีเอกสารที่ คล้ายกันอยู่ในกลุ่มนั้น ๆ และมีตารางแสดงรายการคำที่บอกถึงน้ำหนักความเกี่ยวข้องกับแต่ละ cluster ตัวอย่างที่เห็นเป็นการ process documents ซึ่งแต่ละ document เป็นคำปราศรัยการรับ ตำแหน่งของประธานาธิบดีแต่ละคน กำหนดให้จัดกลุ่มเป็น 8 กลุ่ม คำปราศรัยแต่ละอันจะถูกจัด เข้ากลุ่มตามความคล้ายคลึงของเนื้อหา (ภายใน "Process Documents" นอกจากการ "Tokenize" คำ แปลงเป็นตัวพิมพ์เล็ก "Transform Cases" ตัดคำไวยากรณ์ออก "Filter Stopwords (English)" และแปลงเป็นรากคำ "Stem (Snowball")

Process ) inp	Process Documents	Clustering (2) exa state clu clu			res	
Process doc d	Documents from File Tokenize (2) T oc doc	es ransform Cases doc doc	Filter Stopwor	doc	doc doc	doc doc



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 Filename 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 attribFi								
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 จากนั้นจะต้องบอกว่าให้ดู แท็กชื่ออะไรในการสกัดข้อมูลออกมาโดยการระบุ xpath วิธีที่สะดวกสุดคือใช้ Import Configuration Wizard ในการค้นหาแท็ก จนกระทั่งเห็นข้อมูล text ที่ต้องการปรากฏ ตัวอย่างข้างล่างเป็น ไฟล์ xml ที่ได้มาจาก Oxford Text Archive ซึ่งข้อมูลที่ต้องการคือตัวข้อความในหนังสือ อยู่ภายใต้ แท็ก และใช้ function text() เพื่อดึงตัวข้อความออกมา รูปข้างล่างแสดงการตั้งค่าและ ตัวอย่างข้อมูลที่สกัดออกมาได้จากการใช้ Read XML

<ul> <li></li> <li><th>be used as features in elect multiple nodes). Ir own. Select zero or mo ents with the same</th></li></ul>	be used as features in elect multiple nodes). Ir own. Select zero or mo ents with the same
element       attribute       value         //default:       type       chapter         //default:       null:id       c1         //default:       c1         Apply Selection	s no lady or gentlemar ly without being it ct line from Adam ed with the and malicious y, displayed an II be considered iority of the house taken into
//default:     type     chapter       //default:     null:id     c1         indefault:     null:id     c1         Apply Selection          Path matches 7.001 elements.	that as there was, in th
//default:     null:id     c1       of character. Indeed, it may be laid down as a general princing more extended the ancestry, the greater the amount of violer vagabondism; for in ancient days those two amusements, con excitement with a promising means of repairing shattered for        Apply Selection	e same phase
Apply Selection vagabondism; for in ancient days those two amusements, con excitement with a promising means of repairing shattered for Path matches 7.001 elements.	ciple, that the
Apply Selection <	mbining a wholesome ortunes, were at
Path matches 7.001 elements.	>
//default:TEI/default:text/default:body/default:div/default:p	
	8 million 100 million

•		Data im	port wizard - Step 4 of 5	
7	This wizard guide Step 4: In this ste through the eleme	es you to import your data. p you can select the attribu ents matching the XPath fro	tes which will be used as feature m the previous step using the b	es in the final example set. Navigate uttons at the top.
← P	revious Match	→ Ngxt Match	XPaths for attribu	tes:
Current ele	ement structure:		XPath	current value
default:p	[1]>		text()	A, with any claims to polit
<detau< td=""><td>icn(1)&gt;</td><td></td><td></td><td></td></detau<>	icn(1)>			
			0	
Vew attrib	utes in currently selv	ected node:	•	
iew attrib attribute	utes in currently selve	ected node: value	•	
iew attrib attribute	utes in currently selve	ected node: value	0	
iew attrib attribute	utes in currently selve	ected node: value	0	
iew attrib attribute	utes in currently selve	ected node: value	0	
iew attrib attribute	utes in currently selve	ected node: value	0	
iew attrib attribute	utes in currently sel	ected node: value		

หากต้องการอ่านหลาย ๆ ไฟล์ที่มีโครงสร้างแบบเดียวกันนี้ ก็ทำได้โดยการใช้ operator "Loop Files" ตามด้วย "Append" เพื่อนำผมที่ได้แต่ละไฟล์ไปต่อกันเป็นข้อมูลตารางเดียว แล้วให้ "Read XML" อยู้ภายใน "Loop Files"



ExampleSet (16	a954 examples, 0 special attributes, 1 regular att Filter (16,954 / 16,954 examples): a
Row No.	text()
1	S to the of this direful Scourge of Mankind call'd the I find the moft learn
2	I muft here indeed freely own that I am much pleafed with the and of the learned
3	But it may not be improper to add fomething under this from that great and (while
4	What the Learned Dr. advances, about the Beginning of 4. of his Difcourfe, in thefe
5	In 7. the briefly mentions that most terrible which carried over the greatest Part of
6	Befides what I have quoted from Mr. I fhall entertain the Reader with a more Circu
7	If that Part of the Story be Fact concerning the Worms, Snakes, which ar
8	This molt fearful is allo taken Notice of by many other relates that at it carried off
9	Many other deftructive before and after this are mentioned by Hiftorians, and feveral
10	In 11. towards the End, the juft mentions the Opinion of fome who have thought th
11	This Notion feems to be much a-kin to that which was advanced by (that famous i
12	What the Learned fuggefts in Page 22, 23, againft the old but unhappy Cuftom of

#### การสกัดข้อมูลจากเว็บ

Operator "Crawl Web" ใช้ในการดึงข้อมูลจากเว็บ โดยระบุ url และระดับความลึกของการ กดเข้าไปดูข้อมูล 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 เมื่อเก็บข้อมูลลงในเครื่อง แล้วจึงมา 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

	ant who has last a shild this mavie shade a new light on the
As a par	ent who has lost a child, this movie sneds a new light on the
hristian Fa	uteux 17 November 2016
m not or	te to flat out break down in a movie theater but it just happened. My daughter
ad an ex	tremely rare incurable disease that took her from us a little over a year ago.
he grief	and torment my wife and I have endured cannot be expressed or understood by
nost peop	plethen this movie happened. It shed a new light on the beauty of life and
ow if per	The parties the elegendation of the music (NY COD TUE MUSIC), the
rilliance.	The pacing, the cinematography, the music (MY GOD THE MUSIC)the
rief is pa	and Amy Adams. My god she deserves all the praise she's been getting. Her
ner is pa	thinking of the little things and memories of my daughter that I'll cherich
nrever I	don't know how they accomplished this feat of a movie, but I am so glad that
or once	since my daughter has passed away. I don't feel so, alone
91 out of 3	367 found this helpful. Was this review helpful? Sign in to vote.
ermalink	
▲ 9/10	
A 91.00	
lo CGI (	overkill, just some fine acting and directing
lo CGI ( ladsen7 1	3 February 2017
lo CGI ( ladsen7 1	3 February 2017
lo CGI ( ladsen7 1	Diverkill, just some fine acting and directing I3 February 2017 Eneuve is without a doubt an upcoming director and I can't wait to see Blade 149. Prisoners, Sicario and now Arrival (bayen't seen Enemy yet, or some of his
lo CGI ( ladsen7 ) Denis Ville tunner 20	aneuve is without a doubt an upcoming director and I can't wait to see Blade Alexandree Design and now Arrival (haven't seen Enemy yet, or some of his rk). Denis knows how to capture the tension. It is almost the strongest point of
lo CGI ( ladsen7 1 )enis Ville tunner 20 arlier wo	eneuve is without a doubt an upcoming director and I can't wait to see Blade 049. Prisoners, Sicario and now Arrival (haven't seen Enemy yet, or some of his rk). Denis knows how to capture the tension. It is almost the strongest point of
ladsen7 1 Denis Ville Lunner 20 arlier wo III of his f	eneuve is without a doubt an upcoming director and I can't wait to see Blade 049. Prisoners, Sicario and now Arrival (haven't seen Enemy yet, or some of his rk). Denis knows how to capture the tension. It is almost the strongest point of ilms. Minimal use of computer generated images, and main focus on story, s. acting and thrilling scenes. Back on IMDb board. I have noticed a lot of
No CGI ( ladsen7 1 Denis Ville tunner 20 arlier wo ill of his f haracters people ca	eneuve is without a doubt an upcoming director and I can't wait to see Blade 049. Prisoners, Sicario and now Arrival (haven't seen Enemy yet, or some of his rk). Denis knows how to capture the tension. It is almost the strongest point of ilms. Minimal use of computer generated images, and main focus on story, s, acting and thrilling scenes. Back on IMDb board, I have noticed a lot of ling Arrival a boring movie. So many hypocrites these days. People trying to

เราต้องการสกัดตัวบทวิจารณ์ที่เห็นของแต่ละคนมาเก็บรวบรวมเป็นคลังข้อมูล Get Page จะดึง ข้อมูลในหน้าเว็บนั้นออกมาเป็น document จากนั้นจึงใช้ operator Cut Document ทำหน้าที่สกัด เฉพาะส่วนของข้อความที่ต้องการ โดยสามารถระบุวิธีการสกัดได้ ในตัวอย่างนี้นี้จะใช้ regular region โดยระบุ tag เริ่มต้นที่ต้องการและ tag ปิด ในตัวอย่างนี้ บทวิจารณ์แต่ละคนจะอยู่ภายใน แท็ก <div class="text show-more\_control"> ... <.div> หากเราต้องการข้อมูลส่วนอื่นก็ สามารถเติม field อื่น ๆ เพิ่มได้

1 111 110	Coment Cocoments to Gata			
• •	Edit Parameter List: r	regular region queries		
1	Edit Parameter List: <b>regular region queries</b> Specifies a list of attribute names and their cor might be specified in order to define the start matches will be delivered as result.	rresponding regular expre and the end of a region.	essions. Two regular ex Everything in between t	pressions he two
attribute	name	region delimiter		
rev		<div class="text show</th> <th>📩 </th> <th></th>	📩	
	Add	Entry	ntry Apply	Cance

จากนั้นจึงส่งผลที่สกัดเป็น document ได้ส่งต่อ operator Document to Data เพื่อแปลง ข้อมูลที่ได้เป็น example set (exa) ตัวอย่างผลที่สกัดได้เป็นตารางข้อมูลที่มี field "rev" ตามที่ ต้องการ

-			-			
Row No.	rev	URL	Response	Response	Content-T	Content-L
1	<div class="&lt;/td"><td>https://www</td><td>200</td><td>OK</td><td>text/html;ch</td><td>?</td></div>	https://www	200	OK	text/html;ch	?
2	<div class="&lt;/td"><td>https://www</td><td>200</td><td>ОК</td><td>text/html;ch</td><td>?</td></div>	https://www	200	ОК	text/html;ch	?
3	<div class="&lt;/td"><td>https://www</td><td>200</td><td>ОК</td><td>text/html;ch</td><td>?</td></div>	https://www	200	ОК	text/html;ch	?
4	<div class="&lt;/td"><td>https://www</td><td>200</td><td>ОК</td><td>text/html;ch</td><td>?</td></div>	https://www	200	ОК	text/html;ch	?
5	< <div class="t&lt;/td&gt;&lt;td&gt;ext show-more&lt;/td&gt;&lt;td&gt;control">Sometir</div>	nes I can get very	^ch	?		
5	< for giving too	much away (case	in point, "I	Room" and	ch	?
7	< Sometimes I o	aquot;Passenger	ed by a really go	od teaser trailer	ch	?
8	< Lane").	But most of the tir	ne a "ho h	um" trailer	ch	?
9	< of a "ho	hum" film:	"Jack Reac	her: Never Look	ch	?
10	< recent examp	eing a good ble. Then there is	"Arrival&qu	;tot;	ch	?
11	< the trailer for	/>Because "Arrival&qu	ot; belies absolu	tely nothing about	ch	?
12	of the film. At	face value, it lool	ks like a dubious	"Close	ch	?
13	a threat of me	uot; wannabe, wi ovement towards	th the likes of &quo	t;Independence	ch	?
14	day" an ⊲ Wave"	d "The 5th Actually what you	get is a film that	approaches the	ch	?
15	Press "F3" fo	r focus.	nters&auot:		ch	?

หากต้องการดึงข้อมูลจากเว็บหลาย ๆ page กสามารถใช้ operator Get Pages ได้โดยเอา รายการ url ทั้งหมดไปเก็บไว้ใน Excel ไฟล์ก่อน ใช้ operator Read Excel เพื่ออ่านข้อมูลมาให้ กับ operator Get Pages



#### การสร้าง model สำหรับ multiple regression

multiple regression model เป็นการสร้างโมเดลสำหรับทำนายค่าตัวแปรตามโดยที่ตัวแปรค้น ที่เกี่ยวข้องอาจมีหลาย ๆ ตัวแปร ตัวอย่างต่อไปนี้ ข้อมูลเก็บไว้อยู่ใน Excel คอลัมน์ที่เป็นตัวแปรตาม คือ Y ที่เหลือเป็นตัวแปรที่อาจเกี่ยวข้องกับโมเดลได้ .ให้ใช้ import wizard เพื่อตรวจสอบข้อมูล ้ว่านำเข้าได้ถูกประเภท และจะได้กำหนด role ของคอลัมน์สุดท้ายซึ่งเป็นข้อมูลตัวแปรตามให้เป็น label ตามตัวอย่างในรูป ซึ่งเมื่อเปลี่ยน role แล้วจะเห็นเป็นแถบสีเขียวขึ้น

•					Import Dat	a - Form	at your colu	umns.					-
					Format	your	column	s.				_	
	<u>D</u> ate format	MMM d,	yyyy h:mm:	ss a z	٣		Replace	e errors w	ith missing va	lues 🛈			nrg-
	FTES	0 <b>-</b>	SEC	۰.	LEC	۰.	LAB	۰.	STUDENT	۰.	Y	0 -	
	real		integer		integer		integer		integer		real	Cha	nge Typ
1	103.250		40		41		32		702		20.500	Cha	nge Rol
2	39.680		40		43		36		285		20.333	RIC	pens a
3	44.330		33		39		34		344		17.000	Exc	lude col
4	109.670		33		42		29		743		15.500		
5	231.890		72		73		12		1391		12.333		
	226 970		70		70		0		1261		11 333		

• (					Import Data - For	mat your columns.		
					Format you	r columns.		
	<u>D</u> ate format	MMM d,	yyyy h:mm:ss	a z	¥	Replace errors v	with missing values $ \mathbb{O} $	
	FTES real	¢ •	SEC integer	0 -	LEC O	r LAB O v	STUDENT & -	Y Ø real label
1	103.250		40		41	32	702	20.500
2	39.680		40		43	36	285	20.333
	44.330		33		39	34	344	17.000
3								

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 มาใช้ เพื่อทดสอบโมเดลที่สร้างและรายงานความถูกต้อง ผลที่ได้เป็นค่าสัมประสิทธิ์ของตัวแปรต่างๆ ที่จะ เกี่ยวข้อง

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 ก็สามารถทำลักษะแบบ เดียวกันนี้ได้ อย่างไรก็ดี ตัวโมเดลนี้โปรแกรม RapidMiner สร้างให้โดยขึ้นกับข้อมูลที่นำเข้า ส่วน ผลที่ได้ว่าโมเดลบอกถึงความสัมพันธ์ระหว่างตัวแปรต่าง ๆ จริงหรือไม่ เป็นเรื่องที่ผู้วิจัยต้อง verify ข้อมูลต่าง ๆ ก่อนว่าเข้าเงื่อนไขที่จะใช้โมเดลนั้น ๆ ได้หรือไม่

#### การสร้าง model แบบง่ายสำหรับ sentiment analysis

sentiment analysis เป็นการวิเคราะห์ข้อความเพื่อจัดประเภทว่าข้อความนั้นว่าบอกลักษณะ อารมณ์แบบใด วิธีแบบง่าย ๆ ที่ทำกันคือแยกเป็นสองขั้วอารมณ์คืออารมณ์บวกหรือลบ โดยมีข้อมูล ชุดหนึ่งเป็นข้อมูลที่มีการวิเคราะห์ให้ label แล้วว่าข้อความนั้น ๆ บอกอารมณ์บวกหรือลบ และใช้ เป็นข้อมูลฝึกสอนเพื่อสร้างโมเดลสำหรับไว้ตัดสินข้อความอื่น ๆ ว่าบอกอารมณ์ไปในทางใด ตัวอย่าง ข้างล่างเป็นตัวอย่างจาก forum ของผู้ใช้ RapidMiner ที่อธิบายการสร้าง SVM model สำหรับ sentiment analysis

(https://community.rapidminer.com/discussion/31827/sentiment-analysis-as-a-supervise d-learning-problem)



ตัวอย่างข้อมูลในไฟล์ sentiment\_training เป็น Excel ไฟล์มีสองคอลัมน์ คือ Text กับ Sentiment เป็นข้อมูลที่กำกับขั้วอารมณ์ไว้แล้ว operator ต่าง ๆ ตามรูปทำหน้าที่ตามลำดับ คือ อ่าน ไฟล์ 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 และรายงานผลมา เป็น %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 และจะเห็น ค่า confidence (negative) กับ confidence (positive) ที่คำนวณออกมาได้ของแต่ละ text

จะเห็นได้ว่าวิธีการนี้เป็นวิธีการแบบง่าย ๆ คือมองที่คำเป็นหลักในการสร้างโมเดลมา ทำนายต่อ ไม่มีการใช้ข้อมูลทางภาษาระดับอื่นมาใช้ด้วย เป็นวิธีการมองภาษาแบบที่เรียกว่าเป็นถุง คำ (bag of words) ลำดับการปรากฏของคำในประโยคไม่ได้ถูกนำมาใช้ด้วย หากต้องการทำให้ดีขึ้น อาจใช้ n-gram ของคำมาเป็น token ในการสร้าง vector แทนที่จะใช้แค่คำเดียวก็ได้ กรณีนี้ทำได้ โดยการปรับ Process Documents from Data ให้เป็นดังรูป คือ มีการ "Generate n-gram" มาใช้ แทนคำ

58

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 I
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

 $\square$ 

# โปรแกรมเพื่องานภาษาศาสตร์

ส่วนนี้จะกล่าวถึงโปรแกรมต่างๆ ที่สามารถนำมาใช้เพื่องานวิจัยทางภาษาศาสตร์ได้ ได้แก่

- คารทำแบบสอบถามออนไลน์
- Sensiver of the sensitive of the sensit
- 🥰โปรแกรมกำกับข้อมูล POS/Semantic Tagger
- 🗳การวิเคราะห์เสียงด้วยโปรแกรม Praat
- Sentence (Construction) มี Wordcloud จากข้อมูลภาษา
- คารสร้างภาพความสัมพันธ์เครือข่าย
- การสกัดข้อมูลจากเว็บ
- Sentension Reduction ของข้อความ
- 🥰การทำ Topic Modeling จากคลังข้อมูล
- 🗳 การกำกับข้อมูลด้วยโปรแกรม Corpus Annotator

# การทำแบบสอบถามผ่านคอมพิวเตอร์

#### การทำแบบสอบถามผ่านคอมพิวเตอร์

มีเครื่องมือและเว็บไซ้ต์จำนวนมากที่เปิดให้บริการสร้างแบบสอบถามและจัดส่ง link ทางอีเมล์ ให้ทำแบบสอบถาม และสรุปผลให้ ส่วนมากเป็นบริการที่คิดเงิน มีให้ใช้ฟรีได้บ้างแต่มักจะมีข้อจำกัด เช่น จำนวนข้อที่ถามไม่เกิน 20 ข้อ จำนวนคนที่ตอบแบบสอบถามไม่เกิน 200 คน ตัวอย่างเช่น <u>http://www.survs.com/ ใช้</u>บริการฟรีได้ไม่เกิน 200 response แต่บางที่ยอมให้ใช้ได้เป็นปริมาณ มาก เช่น <u>http://www.kwiksurveys.com/</u> แต่จะคิดค่าบริการจากการให้คำปรึกษาแทน Kwik surveys <u>http://www.kwiksurveys.com/</u>

สามารถแสดงคำถามเป็น text หรือเป็นเสียง mp3 หรือ flash วีดีโอได้



สามารถเลือกสร้างรูปแบบคำตอบได้หลายแบบ ตัวอย่างเช่น



	Excellent	Good	Neutral	Fair	Poor	N//
Room Service	0	0	0	0	0	0
Restaurant	•	0	0	0	0	0
Laundry service	0	0	0	0	0	0
Booking process	•	•	•	•	•	•
ease Enter Your st Name	Contact Det	ails				
ease Enter Your	Contact Det	ails				
lease Enter Your rst Name ist Name	Contact Det	ails				

Rank the following items of importance when going to the beach:						
Scale:						
<ul> <li>1 Must Have</li> <li>5 Do Not Need.</li> </ul>						
las Carros	1	z	3	•	5	
Ice Cream	0	0	0	0	0	
Sun Cream	0	0	0	0	0	
Deck Chair	0	0	0	0	0	
Towel	0	0	0	0	•	
Sun Hat	0	0	0	0	0	

Google doc เป็นอีกวิธีหนึ่งที่สามารถใช้สร้างแบบสอบถาม online ได้ฟรี เพียงแค่มี Google account แล้วสร้าง form ขึ้นมาใน Google doc รายละเอียดสามารถอ่านได้จาก tutorial ต่อไปนี้ <u>https://support.google.com/docs/answer/87809?hl=en</u> <u>http://teacherlink.ed.usu.edu/tlresources/training2/Google/GoogleForms.pdf</u>

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 โปรแกรมจะช่วยอำนวยความสะดวกใน การสร้าง 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 ครอบเพื่อ ให้เรียกใช้งานโปรแกรม tagger ได้ง่ายขึ้น โปรแกรม TagAnt สามารถเลือก tag POS ของภาษา เหล่านี้ได้ 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 ให้แท็ก ข้อมูล ผลลัพธ์ที่ได้มีสองแบบ แบบที่สามารถนำไปใช้กับโปรแกรม AntConc ต่อได้คือแบบแนว นอนที่ 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 สามารถใช้กับข้อมูลแท็ก แบบนี้หรือแท็กตามแบบมาตรฐาน TEI ได้

65

•••	TagAnt 1.2.0
O Input Text Clear	Results
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	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_PPS conduct_NN will_MD regulate_VV the_DT treatment_NN and_CC fare_VVP he_PP will_MD receive_VVSENT 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_NNSSENT
	• Horizontal Vertical Clear
Language English ᅌ Start S	itop

•••	Global Settings				
Category Character Encoding Colors Files Fonts	Tag Settings       Show tags     Hide tags     Hide tags (Search in Conc/Plot/File View)       Hide non-embedded tags     Find tag				
Tags	Start tag				
Token Definition Wildcards	<ul> <li>Hide embedded tags</li> <li>Tag marker</li> <li>Hide header tags</li> </ul>				
	<telheader start="" tag<="" td=""></telheader>				
	/teiHeader> End tag				

โปรแกรม POS Tagger ที่เป็นที่รู้จักแพร่หลายอีกตัวคือ CLAW Tagger ซึ่งเป็น POS Tagger สำหรับภาษาอังกฤษ เป็นบริการมีค่าใช้จ่าย แต่เปิดให้ทดลองใช้กับข้อมูลไม่เกิน 100,000 คำได้ที่ <u>http://ucrel.lancs.ac.uk/claws/trial.html</u> Tag set ที่ใช้จะเป็นอีกชุดหนึ่งต่างจาก Penn Treebank POS นอกจาก CLAW tagger ก็สามารถใช้ตัวอื่น เช่น Stanford POS Tagger เพียงแต่วิธีใช้ จะต้องดาวน์โหลดโปรแกรมไปและเรียนใช้ผ่านคำสั่งที่กำหนดเอาเอง POS Tagger แต่ละตัวก็อาจใช้ POS tagset ที่แตกต่างกันได้ เวลาใช้งานจึงต้องดู POS Guideline ประกอบและเลือกตามความ เหมาะสม บางเว็บไซต์อาจมีบริการให้แท็กข้อมูลแแน่ไลน์ได้ เช่น ที่ parts-of-speech.info ให้ แท็กข้อมูลและให้ POS ออกมาเป็นแบบที่ผู้ใช้ทั่วไปรู้จัก ดังตัวอย่าง



แต่หากคนที่สามารถเขียนโปรแกรมได้เอง จะพบว่ามี package ภาษาต่าง ๆ ให้เลือกใช้เองได้ เช่น ในภาษา Python จะมี package NLTK สำหรับงานประมวลผลภาษาอังกฤษเรื่องต่าง ๆ ถ้า ต้องการใช้ POS Tagging ภาษาไทยก็สามารถเลือกใช้ TLTK ที่มีฟังก์ชั่น pos\_tag ให้ใช้ได้

#### โปรแกรม Parser

โปรแกรม parser เป็นโปรแกรมที่นอกจากจะวิเคราะห์ POS Tag ในข้อมูลแล้วยังวิเคราะห์ความ สัมพันธ์ทางวากยสัมพันธ์เพื่อสร้างโครงสร้างต้นไม้แบบต่าง ๆ อาจจะให้ผลของมา Phrase Structure Tree หรือ Dependency structure ก็ได้ ตัวอย่างโปรแกรมที่มีให้ใช้ คือ Stanford Statistical Parser (<u>https://nlp.stanford.edu/software/lex-parser.shtml</u>) ซึ่งสามารถดาวน์โหลด โปรแกรมและโมเดลสำหรับการ parse มาใช้ได้ ถ้าใช้บน Windows หลังจาก unzip ไฟล์ออกมา แล้วจะเห็นไฟล์ 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 แล้วค่อย ระบุโมเดล parser ที่ต้องการใช้ โปรแกรม gui ช่วยให้เรียกโมเดลและเห็นต้นไม้ได้ชัด แต่ไม่สามารถ save output ได้ จะต้องเรียกใช้ผ่าน command line

🕌 Jar File Chooser 🛛 🗡						
edu/stanford/nlp/models/lexparser/arabicFactor	ed.ser.gz					
edu/stanford/nlp/models/lexparser/chineseFact	ored.ser.gz					
edu/stanford/nlp/models/lexparser/chinesePCFG.ser.gz						
edu/stanford/nlp/models/lexparser/englishFactored.ser.gz						
edu/stanford/nlp/models/lexparser/englishPCFG.caseless.ser.gz						
edu/stanford/nlp/models/lexparser/englishPCFG.ser.gz						
edu/stanford/nlp/models/lexparser/englishRNN.	edu/stanford/nlp/models/lexparser/englishRNN.ser.gz					
edu/stanford/nlp/models/lexparser/frenchFactored.ser.gz						
edu/stanford/nlp/models/lexparser/germanFactored.ser.gz						
edu/stanford/nlp/models/lexparser/germanPCF0	i.ser.gz					
edu/stanford/nlp/models/lexparser/spanishPCF0	6.ser.gz					
edu/stanford/nlp/models/lexparser/wsjFactored.ser.gz						
edu/stanford/nlp/models/lexparser/wsjPCFG.ser.gz						
edu/stanford/nlp/models/lexparser/wsjRNN.ser.gz						
edu/stanford/nlp/models/lexparser/xinhuaFactored.ser.gz						
edu/stanford/nlp/models/lexparser/xinhuaFactoredSegmenting.ser.gz						
edu/stanford/nlp/models/lexparser/xinhuaPCFG.ser.gz						
Okay Cancel						



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 headdependent 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 <u>http://nlp.stanford.edu:8080/parser/</u> ที่จะแสดงผลทั้ง POS Tagger, PS Tree, และ dependency tree สะดวกสำหรับการดูผลการวิเคราะห์ประโยคที่ต้องการได้ทันที

Stanford Parser	
Please enter a sentence to be parsed:	
My dog also likes eating sausage.	
	1.
Language: English   Sample Sentence	Parse
Your query	
My dog also likes eating sausage.	
Tagging	
My/PRP\$ dog/NN also/RB likes/VBZ eating/VBG	sausage/NN ./.

```
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) อาศัยการจัดกลุ่ม คำในพจนานุกรม Longman ที่จัดกลุ่มคำเป็น 21 กลุ่มความหมายใหญ่ (discourse field) แล้วจึง แตกออกมาเป็น 232 ความหมายย่อย มีบริการออนไลน์ให้ใช้ครั้งละไม่เกิน 100,000 คำ ที่ <u>http://ucrel.lancs.ac.uk/usas/tagger.html</u> 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 <u>introduction to</u> the USAS category system (PDF file) with examples

A general and abstract	B the body and the	C arts and crafts	E emotion
F food and farming	G government and	H architecture,	I money and
	public	housing and the home	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

of prototypical words and multi-word units in each semantic field.

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

#### 296 words tagged

Tracing Al0+ 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 28 should S6+ be A3+ the 25 scope A1.7- of 25 the 25 area M7 ? PUNC Is A3+ there 25 overlap N5.2+ with 25 related A2.2 areas M7 , PUNC which 28 has\_Z5 impacted\_A1.1.1 on\_Z5 the\_Z5 development\_A2.1+ of\_Z5 the\_Z5 activity\_A1.1.1 ?\_PUNC What\_28 has\_25 been\_A3+ the\_25 impact\_A2.2 on\_25 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\_25 the\_25 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\_A10avenues\_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 25 context 04.1/A3+ available A3+ here M6 , PUNC the 25 approach X4.2 taken A9+ is A3+ to 25 present A9- a 25 chronological T1/N4 account T1 which\_Z8 traces\_A10+ the\_Z5 development\_A2.1+ of\_Z5 humanities\_P1 computing\_Y2 . PUNC Within\_25 this\_28 ,\_PUNC the\_25 emphasis\_All.1+ is\_A3+ on\_25 highlighting\_X5.1+ landmarks\_M7/A11.2+ where\_M6 significant\_A11.1+

### โปรแกรม Praat

โปรแกรม PRAAT เป็นโปรแกรมที่นิยมใช้วยงานวิจัยทางด้านเสียง สามารถโหลดมาใช้ได้ฟรีจาก <u>HTTP://www.fon.hum.uva.nl/praat/ มี</u>ทั้งที่ทำงานบน windows และ Mac สามารถใช้ในงาน speech ANALYSIS, SPEECH SYNTHESIS, ทำ LABEL AND SEGMENTATION, MANIPULATE SPEECH เป็นต้น ผู้สนใจ สามารถอ่านคำแนะนำการใช้งานโปรแกรม PRAAT เพื่องานทางภาษาศาสตร์ได้ที่ <u>HTTP://SAVETHEVOWELS.ORG/PRAAT/</u> หรืออ่านคู่มือเบื้องต้นได้ที่

HTTP://WWW.STANFORD.EDU/DEPT/LINGUISTICS/CORPORA/MATERIAL/PRAAT\_WORKSHOP\_MANUAL\_V4 21.pdf และ http://savethevowels.org/praat/UsingPraatforLinguisticResearchLatest.pdf



สำหรับคนที่ใช้ภาษา Python จะมี library Parselmouth ที่สามารถใช้งาน Praat ได้เลย ราย ละเอียดสามารถศึกษาได้จาก <u>https://parselmouth.readthedocs.io</u>/

# โปรแกรมสร้างการทดลองภาษาศาสตร์จิตวิทยา

โปรแกรมสร้างการทดลองภาษาศาสตร์จิตวิทยามีหลายโปรแกรมที่สามารถใช้ได้

- DMDX (http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm)
- เป็นฟรีโปรแกรมรุ่นแรกๆ จึงทำงานบน DOS หรือ windows XP นำเสนอคำถามที่เป็นภาพและ เสียง แล้วเก็บ response กับ reaction time จากคำตอบผู้ทดลอง ผู้สนใจสามารถดู Tutorial ได้ ในเว็บของ Matt Davis (<u>http://www.mrc-cbu.cam.ac.uk/people/matt.davis/dmdx.html</u>)
- E Prime (https://www.pstnet.com/eprime.cfm)
   เป็นโปรแกรมจำหน่ายสำหรับทำการทดลองทางภาษาศาสตร์จิตวิทยาที่ได้รับความนิยมเพราะ สะดวกในการใช้งาน
- Psyscope X (<u>http://psy.ck.sissa.it</u>/)

เป็นโปรแกรมฟรีทำงานบน OS X นำเสนอคำถามที่เป็นภาพและเสียง แล้วเก็บ response กับ reaction time จากคำตอบผู้ทดลองได้ ปัจจุบันกำลังปรับปรุงให้มี GUI ที่ง่ายขึ้นต่อการออกแบบ การทดลอง แต่ไม่สามารถใช้งานบน Windows

- PXLab (<u>http://irtel.uni-mannheim.de/pxlab/index.html</u>)
   เป็นอีกโปรแกรมสำหรับงานทดลองทางจิตวิทยา เขียนด้วยภาษา Java เพื่อทำการทดลองรูปแบบ ต่างๆ ได้
- Psychopy (<u>http://psychopy.org</u>/)

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



Build your first PsychoPy experiment (Stroop task) https://youtu.be/VV6qhuQgsil สามารถใช้แทน E Prime ได้ มี Tutorial video ที่เห็นจะช่วยให้เข้าใจการใช้ Psychopy ในการ ออกแบบและสร้างการทดลองง่าย ๆที่แสดงข้อความและรอรับคำตอบจากผู้ทดลอง

# การสร้าง wordcloud

wordcloud เป็นรูปแบบของการแสดงคำที่ปรากฏในข้อมูลตามความถี่ที่พบ แต่แทนที่จะแสดงเป็น ตัวเลขก็แสดงเป็นขนาดของตัวอักษรแทน คำที่ปรากฏใหญ่กว่าจะมีความถี่การปรากฏมากกว่า เรา สามารถสร้าง wordcloud ได้ง่าย ๆ โดยใช้บริการที่มีในอินเทอร์เน็ต เช่น <u>https://www.wordclouds.com</u>/ สามารถ upload ไฟล์ขึ้นไปและแสดงผลเป็น wordcloud พร้อม ปรับแต่งรูปแบบต่าง ๆ ได้



# wordtree

เป็นการสร้าง tree จากตัวบท โดยการระบุคำที่ต้องการค้น โปรแกรมจะดูคำนั้นว่ามีอะไรตามท้าย ได้บ้างไล่ไปเรื่อยๆ จนจบประโยค สามารถใช้งานโดยการ upload text ไปที่เว็บ

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



# โปรแกรมสร้าง network

#### WORDij

เป็นโปรแกรมสำหรับอ่านตัวบทและวิเคราะห์ความสัมพันธ์ของคำภายในตัวบทนั้น สามารถ ดาวน์โหลดมาใช้เพื่อการศึกษาได้ฟรีที่ <u>http://wordij.net</u>/ หลักการเหมือนกับการหา word collocation เราสามารถกำหนดระยะห่างระหว่างคำและความถี่ขั้นต่ำ โปรแกรมจะสร้างไฟล์ output ย่อยๆ ทั้งที่เป็น ไฟล์รายการคำเดี่ยว ไฟล์รายการคู่คำ (bigram) ไฟล์ที่คำนวณความสัมพันธ์โดยใช้ค่าสถิติต่าง ๆ และไฟล์ที่ แสดงภาพเครือข่ายความสัมพันธ์ (pajek format) :ซึ่งสามารถดูได้จากในโปรแกรมนี้เองโดยดูผ่าน ViSij



ไฟล์ข้อมูลที่ต้องการวิเคราะห์ต้องเป็นไฟล์เดียว หากคลังข้อมูลประกอบด้วยไฟล์ย่อยๆ หลายไฟล์ให้ ต่อไฟล์เป็นไฟล์เดียวด้วยคำสั่ง copy เช่น copy \*.txt NewfileName.txt หรือ cat \*.txt > NewfileName.txt

WordLink เป็นแท็บที่ใช้สร้างเครือข่ายความสัมพันธ์ของคำ เพียง upload ไฟล์ข้อมูล และกำหนด ไฟล์รายการคำยกเว้นหรือคำที่ไม่ต้องการนำมาวิเคราะห์ (สามารถใช้ไฟล์ droplist.txt ที่มาพร้อมกับ โปรแกรมได้) ระบุระยะห่างของคำและความถี่ขั้นต่ำ ผลลัพธ์ที่ได้จะถูกเก็บเป็นไฟล์ชื่อเดียวกับไฟล์ ข้อมูลแต่มีนามสกุลต่าง 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 และการแสดงข้อมูลบน แผนที่

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





ตัวอย่างข้างล่างแสดงคำที่แผลงจากรูปมาตรฐานแวเงในรูป network ตามขนาดความถี่ที่พบว่ามี การเขียนในรูปแบบนั้นเท่าใด



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
ไอดิม	290	ไอศครีม
ຫີນ	212	ไอติม
โอ	1947	โอเค
เค	2299	โอเค
เคร	382	โอเค
โอเซ	29	โอเค
แอด	365	แอดมิดชั่น

#### SECTION 9

### โปรแกรมสกัดข้อมูลจากเว็บ

#### โปรแกรม Httrack

(<u>https://www.httrack.com</u>/) เป็นโปรแกรม copy website เหมาะสำหรับการคัดลอกข้อมูล บนเว็บมาเก็บไว้ในเครื่องเพื่อใช้งาน offline Httrack เป็นโปรแกรมสำหรับติดตั้งบน Windows สะดวกในการเก็บข้อมูลจากเว็บใดเว็บหนึ่ง

### ใช้บริการ boilerpipe

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

#### ใช้ Chrome Extension

ใช้ Reader View addon เพื่อแปลงหน้าเว็บให้อยู่ในรูป text based เป็นหลักก่อน addon นี้ใช้
 เพื่อให้สะดวกต่อการอ่าน ซึ่งก็ทำให้สะดวกต่อการ copy-paste เฉพาะข้อความที่ต้องการออกมาได้
 วิธีนี้เป็นการ copy หน้าเว็บเองเป็นหลัก

Extensions	Developer mode
Chrome Reader View 0.0.26 Firefox's Reader View for Google Chrome Permissions Details Allow in incognito Allow access to file URLs	✓ Enabled

 2. ติดตั้ง Advanced Web Scraper เมื่อพบเว็บที่มีข้อมูลที่ต้องการ ให้กดเรียก extension นี้ แล้ว เลือกข้อมูลที่ต้องการโดยการกด \* แล้วคลิกตำแหน่งข้อมูลที่ต้องการบนหน้าเว็บนั้น จะเห็นแถบสี ระบายข้อความที่จะถูกสกัดออกมา จากนั้นให้ตั้งชื่อ field ที่แทนข้อมูลนั้น ทำเช่นนี้กับข้อมูลส่วน อื่นจนครบทุก filed ข้อมูลที่ต้องการ จากนั้นกด preview เพื่อดูข้อมูลที่จะ extract แล้วเลือก option ว่าจะบันทึกข้อมูลเป็น csv, json โปรแกรมนี้เหมาะสำหรับการดึงเว็บที่มีข้อมูลหลาย รายการที่ดึงออกมาเป็นแต่ละ record ได้ เช่น หน้าเว็บวิจารณ์เรื่องต่างๆ ที่สามารถดึง review แต่ละคนออกมาเป็น 1 รายการได้ แต่ข้อมูลที่ดึงจะยังคงติด html tag อยู่บ้าง ต้องมาหาทางตัด ออกเองภายหลังหากไม่ต้องการ

Extensions	Developer mode
Agenty - Advanced Web Scraper 2.7 An easy, powerful web scraping app by Agenty for screen scraping using C selectors and to create agents for Agenty.com. Permissions Developer website Allow in incognito	✓ Enabled  SS
The Godfather More at IMDbPro » Write review Filter: Best  Hide Spoilers: Page 1 of 246: [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11]  Index: 2454 reviews in total	Agenty Help & Support       Help & Support       New       Done       stars       img
735 out of 922 people found the following review useful: The Godfather" is pretty much flawless, and one of the greatest films ever made Author: SJ_1 from United Kingdom 30 September 2005 This review may contain spollers ***	30 results always Title K 10 results always
Rather than concentrating on everything that is great about The Godfather, a much easier way for me to judge its qualit it. Almost every film has something that I don't like about it, but I can honestly say that I wouldn't change anything aboun othing weak about it and nothing that stands out as bad. That's why it gets ten out of ten. This is one of those films that made me wonder why I hadn't seen it earlier. The acting from everyone involved is great, across perfectly as the head of the family, and James Caan and Al Pacino are excellent as his sons. The soundtrack by memorable, bringing back memories of the film every time I hear it. The plot has to be excellent for it to get ten out of the predictable and the film is the definition of a great epic. The film is pretty shocking in the way every death occurs almost instantaneously, and as it spans ten years so many dit every minute of it is great entertainment. It's a well-made and entertaining film that is only the first part of a trilogy, but it wonderful film in its own right. If you haven't seen it, what are you waiting for? This was one acclaimed film that didn't d	review X

#### โปรแกรม Parsehub

Parsehub เป็นอีกโปรแกรมหนึ่งที่มี UI ให้ออกแบบการสกัดข้อมูลจากหน้าเว็บได้ สามารถสกัด ข้อมูลที่มีลักษณะเป็นรายการซ้ำ ๆ กัน หรือมีปรากฏในหลายหน้าได้ โปรแกรมมีทั้ง version ที่เป็น Windows, OS X, Linux สามารถดาวน์โหลดได้จาก <u>https://parsehub.com</u>/ version ที่ฟรีมีข้อ จำกัดปริมาณของการสกัดข้อมูลว่าได้ไม่เกิน 200 หน้า หากต้องการใช้สกัดข้อมูลจำนวนมากเป็น ประจำจะมีค่าใช้จ่ายรายเดือนที่ \$149

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



ขั้นแรกสุด Select Page จะขึ้นมาเอง เว็บตัวอย่างนี้เป็น review ภาพยนต์ซึ่งมีหลายคนมา เขียนวิจารณ์ เราต้องการชื่อบทวิจารณ์ที่ตั้ง (เช่น Picard is back) ตัวบทวิจาณ์ และคะแนนที่ให้ ให้เลือก Select และเลือกตัวชื่อบทวิจารณ์อันนึงก่อน จะเห็นเป็นแถบสี และเนื่องจากข้อมูลมี ลักษณะเป็นรายการหลาย ๆ อัน บทวิจารณ์อันอื่น ๆ จะเห็นเป็นสีเหลืองให้กดที่ชื่อบทวิจารณ์ อีกอันนึงเพื่อบอกโปรแกรมให้รู้ว่าเราต้องการสกัดข้อมูลที่มีหลายรายการ จะสังเกตให้จำนวนตัวเลข

### ้ที่เพิ่มขึ้นมากกว่าหนึ่ง สำรวจว่าข้อมูลชื่อคำวิจารณ์ได้ครบทั้งหน้า และจั้งชื่อ field ตามต้องการ

เช่น "title"



จากนั้นจึงเลือกข้อมูลชิ้นต่อไปที่สัมพันธ์กับชื่อ title กรณีนี้ให้เลือก Relative Select เพื่อโยงชื่อบท วิจารณ์นั้นกับชื่อ title โดยกดเลือกที่ชื่อ title ก่อนแล้วลากลูกศรไปที่บทวิจารณ์ ทำเช่นนี้กับส่วนที่ เป็นคะแนนด้วยเพื่อเก็บคะแนนที่ผู้เขียนแต่ละคนให้ พร้อมทั้งตั้งชื่อ field แต่ละอัน เช่น review, score



		TAO INCAICARD
Selec	t page +	Hide Spoilers Filter by Rating: Show All Sort by: Helpfulness
Sel E	ect title (  Extract name Extract url Empty Relative selection1 (0)	<ul> <li>9/10</li> <li>Picard is back Itorian-boettcher 18 January 2020</li> <li>Inv</li> <li>Had the opportunity to watch the first 3 episodes tonight at the premiere in Berlin. Just one word Great.</li> <li>133 out of 214 found this helpful. Was this review helpful? Sign in to vote.</li> <li>Permalink</li> </ul>
Sala		176 Reviews       Hide Spoilers Filter by Rating:     Show All     Sort by:     Helpfulness     Image:
Sele	ct page + elect title = + Extract name	176 Reviews ☐ Hide Spoilers Filter by Rating: Show All Sort by: Helpfulness ↓ Picard is back
Selle	ct page + elect title • + Extract name Extract url	176 Reviews Hide Spollers Filter by Rating: Show All Sort by: Helpfulness ● ↓ Picard is back Horian-boettcher 18 January 2020
Sele	ct page + elect title + Extract name Extract url Relative review A + Relative selection1 (24) A +	176 Reviews         Hide Spoilers Filter by Rating: Show All Sort by: Helpfulness         Hide Spoilers Filter by Rating: Show All Sort by: Helpfulness         Picard is back         Inrian-boettcher         I had the opportunity to watch the first 3 episodes tonight at the premiere in Berlin. Just one word Great.

ข้อมูลที่สกัดได้ของแต่ละบทวิจารณ์ก็จะมีทั้ง title, review และ score โดยจะสกัดทั้งหมดเป็น ข้อมูลรายการที่มี field ทั้งสามนี้ แต่ในหน้าเพจนี้จะเห็นว่ามีปุ่มให้กด Load More ซึ่งเป็นการเรียก หน้าถัดไปขึ้นมา กรณีนี้ให้เลือก Select อีก item เพิ่มและให้เลือกเป็น Click และจึงบอกว่า Click นั้นเป็นการ click แบบดึงหน้าถัดไป แล้วเลือกให้กลับไปใช้ main template ให้ทำซ้ำ ก็จะเป็นการ วนรอบทำงานได้



Click setup	×
Is Load More a next page button?	
Yes	Νο
Examples of next page buttons	Examples of non next page buttons
1 2 3 4 5 18758 next	Search Advanced
« < 1234567 >>	Director: David Leitch
Page 1 of 40	Writers, Knett Keese, Paul Wernick <u>printe creats</u> *
Prev 1 2 3 Next	Refurchance Product ratings
Prev 1 2 3 Next	Image: Constraint of the second se

	- +	Star Trek: Picard (TV Series 2 🛪 🖉 🚥 Star Trek: Picard (TV Series 2 🛪 🕂
📽 imdb.co		( https://www.imdb.com/title/tt8806524/reviews?ref_=tt_urv
Select page	Þ	I'm sure people were expecting something like Discovery but Picard is a thinking mar and not Jason Bourne and I'm hoping that the series carries on with that in mind, especially considering he pushing on 90.
Select title	<del>_</del>	3 out of 6 found this helpful. Was this review helpful? Sign in to vote. Permalink
Extract uni		
Relative review 🔺	E+	v Load More
Relative score 📤	E E	
Select more (1)	Û	See also
Click each more item		Awards   FAQ   User Ratings   External Reviews   Metacritic Reviews
and oo to main temp	ilate	

ภาพขั้นตอนการสกัดข้อมูลทั้งหมดจะเห็นเป็นขั้นตอนคำสั่งตามตัวอย่างที่เห็น จากนั้นจึงสั่ง RUN เพื่อให้โปรแกรมสกัดข้อมูลตามที่ตั้งค่าไว้นี้ ผลลัพธ์ที่ได้สามารถเก็บเป็น CSV เปิดใน Excel หรือเป็น JSON ก็ได้

our data is ready! Click on	the green buttons to down
Down	load Data
CSV/Excel • J	SON API
Report a	in issue here.
Template Name	Pages Scraped
main_template	10 🖬

	A	В	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. N	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 0	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) คือเป็นโปรแก รมที่ช่วยในการสร้าง bot สำหรับมาทำงานที่ซ้ำ ๆ เราเพียงแต่กำหนดลำดับขั้นตอนว่าต้องทำอะไร ก่อนหลังอย่างไร การสกัดข้อมูลจากเว็บจึงเป็นงานหนึ่งที่สามารถใช้ UiPath นี้ได้ โปรแกรม UiPath ที่เป็น community edition สามารถดาวน์โหลดมาใช้งานได้ (<u>https://www.uipath.com/</u> <u>developers/community-edition</u>) โปรแกรมทำงานบน Windows

หลังติดตั้งโปรแกรมแล้ว ให้เรียก UiPath Studio ที่ใช้ในการออกแบบ process automation ในการสกัดข้อมูลจากเว็บ ตัวอย่างจากการสกัดข้อมูลบทวิจารณ์ภาพยนตร์ของเว็บ IMDB ให้เลือก หนังที่ต้องการ ไปที่หน้าที่มีบท review เรียงเป็นรายการ เช่น <u>https://www.imdb.com/title/</u> <u>tt0073486/reviews?ref\_=tt\_ov\_rt</u>

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 ก็ให้ ใช้ Internet Explorer แทนโดยไม่ต้องใช้ extension



เลือก New - Sequence จะเห็นกล่อง operation ที่เราสามารถ add process ต่างๆ เข้าไปได้ใน กล่องนั้น ทางซ้ายเป็นตัวเลือกเครื่องมือต่าง ๆ ที่สามารถลากมาใช้ได้ (อยู่ภายใต้หมวด activity) ให้เลือก Sequence เข้ามาเพื่อสร้างกล่องลำดับกระบวนการ จากนั้นจึงเลือก Data Scraping เพื่อ ใช้สกัดข้อมูลหน้าเว็บ



เมื่อเลือก Data Scraping แล้วจะมีหน้าจอถามให้เราเลือก element ที่ต้องการในหน้าเว็บ ข้อมูลที่ มีลักษณะเป็นรายการซ้ำๆ กันนี้ เราจะต้องเลือก second element ของข้อมูลประเภทเดียวกันด้วย จะเห็นว่าข้อมูลส่วนที่ถูกเลือกจะมีระบายสีให้เห็น UiPath จะวิเคราะห์ดูว่าสิ่งที่ต้องการสกัดคืออะไร ให้เราตั้งชื่อ field ข้อมูล ในที่นี้ใช้ title กับ url เพื่อสกัดเอา link ของบทวิจารณ์นั้นมาด้วย กรณีที่ เราบังเอิญเลือกข้อมูลสองตัวใน field นั้นไม่ตรงกัน โปรแกรมจะฟ้องว่าสกัดไม่ได้ จึงต้องดูให้ชัดว่า ส่วนที่ระบายสีทั้งของ first และ second element เป็นข้อมูล field เดียวกัน

★ 9/10 Jack Nicholson at his f Agent10 13 August 2002	Ui Extract Wizard			×
It's tough to really judge th or Jack Nicholson's best pe director and actor are quite sad, yet uplifting ending, o parodied as many times as Well, greatness was achiev		Configure Colun The identified fields are Extract Text	n <b>ns</b> e highlighted.	
163 out of 218 found this helpfu Permalink		Text Column Name	title	
<b>±</b> 10/10		UKL Column Name	un	
A masterpiece perica-43151 20 July 2018	Help	Cance	el < Back Next	
The seventies produced som Before the era of blockbuste the Hollywood power-broker	ne of the most in ers, and ever inc	teresting and worth reasing dumbing do oneymakers, there	hy Hollywood movies. own of the cinema art by was this short but truly	

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

both uplifting and disheartening, sometimes both at or	/review/ox0000071/2ref =# up/
ouching and moving a great cinematic experience	/review/rw09900/1/:rei_=t(_urv
ouching and moving, a great cinematic experience	/review/rw0143219/?ref_=tt_urv
A great order vs. chaos tale that everyone can relate to	/review/rw1105961/?ref_=tt_urv
What an excellent movie" is all that went through my i	/review/rw0143229/?ref_=tt_urv
Poetic - Powerful - Simple: The Greatness of Cuckoo's №	/review/rw0143051/?ref_=tt_urv
ack Nicholson at his finest	/review/rw0143105/?ref_=tt_urv
A masterpiece	/review/rw4245862/?ref_=tt_urv
lest film of its era	/review/rw0143063/?ref_=tt_urv
'he spirit of freedom vs. the spirit of legal-ism	/review/rw2692389/?ref_=tt_urv
A perfect mixture of entertainment and drama.	/review/rw1143457/?ref_=tt_urv
One Flew Over One of the Best Movies Ever Madel!! (Ar	/review/rw0143010/?ref_=tt_urv
Inderwhelming	/review/rw2449184/?ref_=tt_urv
Set Mad If You Want To! This Is Very Realistic!	/review/rw1712262/?ref_=tt_urv
Great Experience	/review/rw4096157/?ref_=tt_urv
ack Nicholson's best yet!	/review/rw3036484/?ref_=tt_urv
'he Ultimate Backfire	/review/rw2007664/?ref_=tt_urv
lest movie ever	/review/rw4086029/?ref_=tt_urv
ou would have to be crazy not to watch this movie.	/review/rw3200360/?ref_=tt_urv

ทำการเลือกข้อมูลที่เกี่ยวข้องในรายการนั้นจนครบที่ต้องการ จึงกด Finish ในหน้า Preview Data

	text	review	url	title
		10/10	/review/rw0998871/?ref_=!	Both uplifting and disheart
		10/10	/review/rw0143219/?ref_=!	Touching and moving, a gr
		10/10	/review/rw1105961/?ref_=t	A great order vs. chaos tak
		10/10	/review/rw0143229/?ref_=!	"What an excellent movie"
		10/10	/review/rw0143051/?ref_=:	Poetic - Powerful - Simple:
ally judge tł	It's tough to really j	9/10	/review/rw0143105/?ref_=!	Jack Nicholson at his finest
roduced so	The seventies prod	10/10	/review/rw4245862/?ref_=!	A masterpiece
One Flew ( d on Ken Ke	Milos Forman's One	10/10	/review/rw0143063/?ref_=!	Best film of its era
		9/10	/review/rw2692389/?ref_=!	The spirit of freedom vs. th
		9/10	/review/rw1143457/?ref_=!	A perfect mixture of entert
		10/10	/review/rw0143010/?ref_=:	One Flew Over One of the
		6/10	/review/rw2449184/?ref_=!	Underwhelming
		10/10	/review/rw1712262/?ref_=:	Get Mad If You Want To! TI
nd emotior	Very touching and	9/10	/review/rw4096157/?ref_=!	Great Experience
		10/10	/review/rw3036484/?ref_=!	Jack Nicholson's best yet!
		10/10	/review/rw2007664/?ref_=!	The Ultimate Backfire
st movie th	This was the best m	10/10	/review/rw4086029/?ref_=!	Best movie ever
older Jack 1	The more I see olde	10/10	/review/rw3200360/?ref_=!	You would have to be crazy
nd e st mo	Very touching and o This was the best m The more I see olde	9/10 10/10 10/10 10/10 10/10	/review/rw4096157/?ref_=1 /review/rw3036484/?ref_=1 /review/rw2007664/?ref_=1 /review/rw4086029/?ref_=1 /review/rw3200360/?ref_=1	Great Experience Jack Nicholson's best yet! The Ultimate Backfire Best movie ever You would have to be crazy

หลังจากนั้น จะมีหน้าจอถามว่าหน้าเว็บนี้เป็นประเภทที่มีหลาย ๆ หน้าไหม คือ ยังมีให้กด Next Page หรือ More อีกไหม ถ้ามีก็จะไปกดหน้าถัดไปและไปสกัดข้อมูลที่ต้องการจนได้จำนวน รายการครบตามที่ตั้งค่าไว้ แต่ก่อนจะกดเลือกขั้นนี้ ควรกลับไปที่หน้าเว็บแล้วเลื่อนให้เห็นปุ่มสำหรับ กด Next Page ก่อน

Permalink	Ui Indicate Next Link	×				
<b>*</b> 9/10	Is data spanning multiple pages?					
An Instant Classic Gideon24 8 March 2015	Identify the element that navigates to the next pages. It could be a 'Next' button or an arrow (not a page number).					
Warning: Spoilers	Press Yes to indicate it.					
5 out of 6 found this helpful. Was this	Yes	No				
Permalink		~				

เมื่อกด Yes แล้ว ให้ไปที่หน้า Browser เพื่อบอกว่าปุ่ม Next Page หรือ More อยู่ที่ไหน เป็นอัน เสร็จสิ้นกระบวนการสกัดข้อมูล เราจะเห็น process ที่กำหนดเป็นลำดับหน้าจอและคำสั่ง เหมือนเป็น visual programming เพื่อให้เห็นว่ากระบวนการนี้มีขั้นตอนอย่างไร



ข้อมูลที่สกัดได้จะเป็น Data Table ที่มี field ต่างๆ ตามชื่อที่ตั้งไว้คือ title, url, review, text ขั้น ตอนต่อไปคือการเก็บข้อมูลลงไฟล์ ซึ่งสามารถเก็บเป็น Excel ได้ ให้พิมพ์หา Excel ในกล่องเครื่อง มือจนเจอ Excel Application Scope แล้วลากมาตอนท้ายของกล่องของ Data Scraping เพื่อบอก ว่าให้ใช้ Excel Application และในกล่อง Excel นี้ให้ใส่ชื่อไฟล์ที่ต้องการเก็บข้อมูล ในที่นี้ใส่ test.xlsx จากนั้นลาก Write Range แล้วบอกว่า Input ที่เป็น DataTable มาจากไหน ก็คือเอา มาจากข้อมูลที่สกัดจากเว็บ เนื่องจากเราไม่ได้เปลี่ยนชื่อตัวแปรใน Output ของ process ก่อน ตัว แปรที่เก็บข้อมูลคือ ExtractDataTable (กรณีที่ถ้าเราต้องการใช้ตัวแปรที่อยู่นอกกล่อง ให้ไปดูที่ตัว แปรนั้นแล้วเปลี่ยน scope ให้ครอบคลุมถึงกล่องที่ต้องการใช้ได้) ข้อมูลจำนวนที่ต้องการสกัดจะถูก เก็บลงใน Excel ไฟล์ตามที่ต้องการ กรณีที่ต้องการ Header row ในไฟล์ Excel ด้วยก็ให้เลือก AddHeaders

	Main	Expand All Collapse A
Available     App Integration     Excel     Processing	Contraction to the second time of the contraction of the contraction of the contraction of the contraction of the format contraction of the format contraction of the contraction	
<ul> <li>Table</li> <li>Append Range</li> <li>Close Workbook</li> </ul>	Excel Application Scope	,
Copy Sheet Delete Range	Workbook path. Text must be quoted	
Excel Application Scope	[‡] Do	*
Get Cell Color	Drop Acti	D ivity Here
Get Workbook Sheet		

Activities 🗸 🕈	$Main^*\times$	-	Properties	ų
	Main	Expand All Collapse All	UiPath.Excel.Activities.E	xcelWriteRange
excel X	+	<u>ه</u>	Common	
Get Selected Range			DisplayName	Write Range
Get Workhook Sheet			Destination	
the contract of the			SheetName	"Sheet1"
Get Workbook Sheets	+		StartingCell	"A1" -
Read Cell	Excel Application Scope		🗆 Input	
Read Cell Formula	"test.xlsx"		DataTable	ExtractData
Read Column			Misc	
Read Range	[ <b>*</b> ] Do		Private	
- Read Row			Options	
Save Workbook	(+)		AddHeaders	
	Write Range	~		
Select Kange	"Sheet1"	"A1"		
Color 💭 Set Range Color	ExtractDataTable			
Write Cell	LAIOCIDOISTODIE			
🔄 Write Range	(+)	_		
<b>v</b>		· · · · · · · · · · · · · · · · · · ·		

ตัวอย่างที่ยกมาเป็นตัวอย่างง่าย ๆ ที่ข้อมูลทั้หงมดอยู่ในหน้าเว็บมีการแบ่ง field ต่างๆ ชัดเจน ข้อมูลที่มีหลายหน้าก็สามารถสกัดได้ แต่ในบางเว็บ เช่น ในเว็บตัวอย่างนี้ หากดูข้อมูลที่สกัดได้ จะเห็นว่าบางรายการไม่มีตัวบทวิจารณ์มา ทั้งนี้เพราะในหน้าเว็บไม่ได้แสดงข้อมูล field นี้ทั้งหมด ผู้ ใช้ต้องกดอ่าน full review อีกที่ หากจะทำต่อจริง ๆ ก็ต้องสร้าง automation ต่อให้ดูว่ารายการที่ มีตัวบท ให้ไปเอา url ที่ดึงมา ไปสร้าง address ส่งให้เว็บแล้วสกัดบทวิจารณ์นั้นมาใส่เพิ่มใน รายการนั้น

#### โปรแกรม Octoparse

Octoparse เป็นโปรแกรมสกัดข้อมูลบนเว็บอีกโปรแกรม แต่ใช้งานได้เฉพาะบน Windows เป็นโปรแกรมที่มี UI ช่วยให้ง่ายต่อการเลือกส่วนของข้อมูลที่ต้องการสกัดนำมาใส่เป็น field ต่าง ๆ ได้ มีข้อดีคือสามารถสกัดข้อมูลที่ปรากฏเป็นรายการซ้ำ ๆ ในหน้าเว็บและที่มีรายการอยู่มากกว่าหนึ่ง หน้าก็ได้ ใน version ฟรีสามารถใช้สกัดข้อมูลได้ในปริมาณจำกัดได้ไม่เกิน 10,000 รายการ หาก ต้องการใช้งานจำนวนมากจะมีค่าใช้จ่ายรายเดือน \$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 เพื่อช่วยในการสกัดข้อมูลที่ต้องการจากเว็บเพื่อมาใช้ในงานที่ ต้องการได้ package หนึ่งที่เป็นประโยชน์เพื่อการนี้คือ rvest ตัวอย่างข้างล่างดัดแปลงจาก <u>http://stat4701.github.io/edav/2015/04/02/rvest\_tutorial/</u>

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 อยู่ แล้ว สกัด text ในนั้นออกมา แปลงเป็นตัวเลข เครื่องหมาย %>% เป็นการ pipe หรือส่งผ่าน ข้อมูลต่อ 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
```

```
<strong title="8.5 based on 15,578 user ratings"><span
```

itemprop="ratingValue">8.5</span></strong>

```
<div class="ratingValue">
</strong title="8.5 based on 15,578 user ratings">
</strong="ratingValue">
</span itemprop="ratingValue">8.5</span>
</strong>
</span class="grey">/</span>
```

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

```
> 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"
```



จะเห็นว่าการใช้ rvest ช่วยให้เราสามารถดึงข้อมูลส่วนที่ต้องการในหน้าเว็บมาใช้ได้เลย เพียงแต่เรา ต้องรู้ว่าข้อมูลนั้นอยู่ที่ไหน ภายใต้แท็กอะไร เพื่อจะได้เขียนคำสั่งสกัดให้ถูกตำแหน่งได้

#### ใช้ Python package สำหรับ web crawler

Python มีโมดูล webscrapy ที่ช่วยในการสกัดหน้าเว็บที่ต้องการได้ และมักใช้ร่วมกับF,ดูล BeautifulSoup ที่ช่วยให้เข้าถึงแท็กต่าง ๆ ใน html เพื่อเข้าถึง text ที่ต้องการได้ ตัวอย่างการเขียน โปรแกรม 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 โดย ใช้ BeautifulSoup เก็บโครงสร้ง html ไว้ในตัวแปร soup จากนั้น เราต้องดูข้อมูลใน page นั้นว่า ส่วนที่ต้องการสกัดอยู่ภายใต้แท็กชื่ออะไร บทวิจาร์แต่ละอันอยู่ในแท็ก <div> ที่มี attribute "class" ที่มีค่าเป็น 'review-container' วนลูปด้วยการใช้ find\_all ของ BeautifulSoup จากนั้น ในแต่ละ review เราจะสกัดสองอย่าง คือ rating 1-10 และตัวบทที่เป็นความเห็นของคนวิจารณ์ ตัวความเห็นอยู่ในแท็ก <div> ที่มี attribute "class" ที่มีค่าเป็น 'text show-more\_control' ดึง ออกมา (ตัวอย่างคือ <div class="text show-more\_control">This film will be a classic one day! ......) ส่วนคะแนน review อยู่ภายใต้ <span> .. <span> (ตัวอย่างคือ <span>8</ span><span class="point-scale">/10</span>) จึงใช้ find ซ้อนกันสองครั้ง การสกัดข้อมูลจาก แต่ละเว็บจึงต้องเขียนแบบเจาะจงให้เข้ากับโครงสร้างข้อมูลแท็กภายในไฟล์ 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])

А	В	с	D	E	F	G	н	I.	J	к
10	Growing up	Mary Poppins	was one of n	ny favorite fil	ms. When Sa	ving Mr. Bank	s came out a	few years ba	ck I re-watch	ed the ori
10	This film wil	l be a classic	one day! It w	as great! Emi	ly Blunt as Ma	ary Poppins w	as the only p	erson I would	I have picked f	for the rol
8	Greetings ag	ain from the	darkness. The	e 1964 classic	Disney film N	MARY POPPIN	S is much bel	oved and has	been shared	across ge
10	Mrs. Shulliva	in and I treate	ed ourselves t	o a special ea	rly Christmas	present by at	tending the o	pening night	of Mary Popp	ins Retur
3	With the mu	Itiple referen	ces to the ori	ginal Mary Po	ppins through	nout this movi	e, and even ir	n the title, m	y hope was th	is movie
10	This is profo	undly one of t	he years very	best pictures	l It is filled w	ith pure bliss	and the magi	c Disney has	been aiming	to posses:
2	Excellent per	rformances of	thoroughly u	inimaginative	music, a pale	reflection of	the original,	which was al	I about the m	usic. It's a
 3	I can't begin	to say how e	cited I was f	or this movie.	I've been wa	iting over a ye	ear for it. We	brought my	daughter to se	e it with:
2	Long and bor	ring! Fell asle	ep several tin	nes and I just	wanted it to	end. None of	the songs ma	de any sense	and they are	totally fo
2	The acting w	as good. The	characters w	ere likeable b	ut the music	let it down. Ti	his film wasn	't necessary.		
1	I went to see	e this, against	my basic ins	tinct, because	I thought the	e original film	was so marv	ellously good	that they cou	ıldn't go s
1	Why every ti	me they need	to sing every	damn thing	I have never b	een more ann	noyed with a	movie as I ha	ve watching t	his crap.
3	First things									
2	It took all of	me not to get	t up and walk	outand I've	sat through s	some bad mo	vies! But the	story &	music were ju	ist awful.
2	Since Disney	is incompete	nt of coming	up with new	ideas, and mu	ust resort to u	sing older sto	ries they did	years ago, th	ey certain
1	The great									
1	I believe that	t 'Mary Poppi	ns Returns' is	the soulless	Disney nostal	gia cash-grab	number 112?	I don't know	, I lost count	long ago.l
1	This film wa	s a huge disa	ppointment. I	t has a cast o	f classic actor	s including Er	nily Blunt, Me	eryl Street, Co	olin Firth and	Angela Ln
1	Honestly, I w	as cringing b	y the end of t	his movie and	would have	eft if I weren	't with a large	e group. Man	Poppins didr	n't really d
1	It was alway	s going to be	hard to go up	o against a ch	erished classi	c but this film	is just dire. [	Don't waste y	our time. I ha	d to walk
 2	***Warning	Mild Spoilers	***Where w	as the joy? Th	e charm? The	e humor? Mar	y Poppins Ret	turns is such	a slog of a me	ovie and p
2 1 1 1 1 1 2	Since Disney The great I believe that This film wa Honestly, I w It was alway ***Warning	is incompete t 'Mary Poppin s a huge disar vas cringing br s going to be Mild Spoilers	nt of coming ns Returns' is ppointment. I y the end of t hard to go up ***Where wa	up with new the soulless t has a cast o his movie and against a ch as the joy? Th	ideas, and mu Disney nostal f classic actor d would have l erished classi ne charm? The	ust resort to u gia cash-grab 's including Er left if I weren c but this film e humor? Mar	sing older sto number 112? nily Blunt, Me 't with a large is just dire. I y Poppins Re	ries they did I don't know eryl Street, Co e group. Man Don't waste y turns is such	years ago, th y, I lost count olin Firth and y Poppins didr your time. I ha a slog of a mo	ey certa long ag Angela n't really d to wa ovie and

### โปรแกรมสำหรับทำ Dimension Reduction

การวิเคราะห์โดยใช้ multivariate analysis เป็นวิธีการทางสถิติที่ได้รับความนิยมมากขึ้นในงาน วิจัยที่ข้อมูลมีตัวแปรต้น (independent variable) หลายตัว ในกลุ่มนี้ เทคนิคการลด dimension เป็นวิธีการที่ถูกนำมาใช้ในการศึกษาภาษา งานที่เป็นที่รู้จักดีคืองานของ Biber (1993) ที่ใช้ Factor Analysis ลดตัวแปรทางภาษาจำนวนมากที่พบใน text ลงเหลือ 5-6 dimension ที่สามารถใช้แยก ความต่างทาง genre ของตัวบทได้ หรืองานของ Binongo (2003) ที่ใช้วิธีการลด dimension เพื่อ หาลักษณะการเขียนของแต่ละคนและสรุปว่างานชิ้นที่สงสัยควรเป็นงานที่เขียนโดยคนไหน โปรแกรมที่ใช้สำหรับทำงานทั้งสองแบบมีผู้พัฒนาให้ใช้ ดังนี้

โปรแกรม Multidimensional Analysis Tagger เป็นโปรแกรมที่เขียนขึ้นโดย Andrea Nini เพื่อวิเคราะห์ตัวบทโดยอาศัยลักษณ์ต่าง ๆ ที่พบในตัวบท ตามแนวการวิเคราะห์ที่ 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 ตาม แบบ Biber โดยลด dimension ออกมาเป็น 6 dimension ตัวอย่างข้างล่างแสดงผลการวิเคราะห์จะ เห็น folder ย่อยที่ถูกสร้างขึ้นสองแฟ้มภายใต้ folder ที่เราเก็บข้อมูล

•	MAT_test-ma
	1841Harrison_MAT.txt
	ap20010629_MAT.txt
	BB_Pat3_MAT.txt
	bern_cv_MAT.txt
	Cop_basic_MAT.txt
	design_R_MAT.txt
	Statistics
•	ST_test-ma
	ST_1841Harrison.txt
	ST_ap20010629.txt
	ST_BB_Pat3.txt
	ST_bern_cv.txt
	ST_Cop_basic.txt
	ST_design_R.txt

folder ST\_ เก็บข้อมูลที่มีการแท็ก POS folder MAT\_ เก็บข้อมูลที่แท็กลักษณ์ภาษาเพิ่มเติม แสดงใน [] ภายใต้ folder MAT\_ จะมี folder statistics ที่เก็บผลการวิเคราะห์ไว้ ในตัวอย่าง corpus นี้มีไฟล์ 6 ไฟล์ แต่ละไฟล์จะถูกวิเคราะห์ว่าเมื่อมองใน dimension 1-6 แล้วมีค่าคะแนนเท่าไร และควรจัดเป็นตัวบทประเภทใด เช่น learned exposiiton, scientific exposition และถ้ามองภาพ รวมทั้ง 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 ที่ทอลองเป็นเส้นขวาสุดเทียบกับตัวบทมาตรฐานที่เป็น ตัวแทนตัวบทต่างๆ 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 มีการ แปลง text เป็นเว็กเตอร์และลดมิติด้วยเทคนิค PCA เพื่อใช้วิเคราะห์เปรียบเทียบระหว่าง text ได้ โปรแกรมนี้พัฒนาโดย Harri Siirtola ในโครงการ DAMMOC project สามารถดาวน์โหลดได้จาก <u>http://www.uta.fi/sis/tauchi/virg/projects/dammoc/tve.html</u> เป็นโปรแกรม Java จึงใช้ได้กับ ทั้ง Windows และ Mac
Type/token			0.38 - 0.54 - 0.6
and the second data as	Additional Address of the Additional States		All and the same
Hapax legomena/type			0.64 - 0.79 - 0.5
Average word length			3.52 - 4.60 - 6.0
	manufacture and a second		the second second
		Clustering:	Pronouns
and dive and yellow nouses and the ros	egaroens and the jessamine and		Edit words
peraniums and cactuses and Gibraltar as of the mountain yes when I put the rose	s a girl where I was a Flower : in my hair like the Andalusian		Word Count
girls used or shall I wear a red yes and h Moorish wall and I thought well as well I	iow he kissed me under the him as another and then I asked		
him with my eyes to ask again yes and t	then he asked me would I yes to		
menus my mountain Games and first to	aut muy some around him use and		
say yes my mountain flower and first I p drew him down to me so he could feel m	out my arms around him yes and ny breasts all perfume yes and his		
say yes my mountain flower and first I p drew him down to me so he could feel n heart was going like mad and yes I said	out my arms around him yes and ny breasts all perfume yes and his yes I will Yes.		
say yes my mountain flower and first I p drew him down to me so he could feel n heart was going like mad and yes I said Trieste-Zurich-Paris 1914–1921	aut my arms around him yes and ny breasts all perfume yes and his yes I will Yes.		
say yes my mountain flower and first I p deew him down to me so he could feel n heart was going like mad and yes I said Inieste-Zurich-Paris 1914–1921	out my arms around him yes and ny breasts all perfume yes and his yes I will Yes.		
say yes my mountain flower and first I p drew him down to me so he could feel n heart was going like mad and yes I said Inieste-Zurich-Paris 1914–1921	aut my arms around him yes and ny breasts all perfume yes and his yes I will Yes.	0	
say yes my mountain flower and first I p dew him down to me so he could feel n heart was going like mad and yes I said Inieste-Zurich-Paris 1914–1921	Nord break:		
say yes my mountain flower and first I p dnew him down to me so he could feel m heart was going like mad and yes I said Inieste-Zurich-Paris 1914–1921 Window: 994 words	Word break: -+/=#%,:*		
say yes my mountain flower and first I p deew him down to me so he could feel n heart was going like mad and yes I said Inieste-Zurich-Paris 1914–1921 Window: 994 words	Word break: -+/=#%,;;* Word count: 26540	17	
Say yes my mountain flower and first I p deew him down to me so he could feel n heart was going like mad and yes I said Trieste-Zurich-Paris 1914-1921 Window: 994 words	Word break: -+/=#%,:* Word count: 26540	0 17	
say yes my mountain flower and first I p dnew him down to me so he could feel n heart was going like mad and yes I said Inieste-Zurich-Paris 1914-1921 Window: 994 words Overlap: 0	Nord break: -+/=#%,;* Word break: -+/=#%,;* Word count: 26540 Frag. count: 26	0 57 58	
say yes my mountain flower and first I p deew him down to me so he could feel n heart was going like mad and yes I said Inieste-Zurich-Paris 1914–1921 Window: 994 words Overlap: 0	Word break: -+/=#%,:* Word count: 26540 Frag. count: 21	57	
say yes my mountain flower and first I p deew him down to me so he could feel n heart was going like mad and yes I said Inieste-Zurich-Paris 1914-1921 Window: 994 words Overlap: 0	Nut my arms around him yes and ny breasts all perfume yes and his yes I will Yes. Word break: -+/=#%,:* Word count: 26540 Frag. count: 20	57	
say yes my mountain flower and first I p deew him down to me so he could feel n heart was going like mad and yes I said Inieste-Zurich-Paris 1914–1921 Window: 994 words Overlap: 0	Word break: -+/=#%,;;* Word count: 26540 Frag. count: 22	i7 i8	
window: 994 words Overlap: 0	Word break: -+/=#%,:* Word count: 26540 Frag. count: 21	57 58	w regions 2 :

ตัวอย่างการใช้งานโปรแกรมนี้คืองานของ Binongo (2003) ที่ต้องการศึกษาว่า หนังสือที่เป็น ประเด็นตั้งคำถามกันว่า หนังสือ The Royal Book of Oz เป็นหนังสือเล่มสุดท้ายที่เขียนโดย Lyman Frank Baum หรือเป็นเล่มแรกที่ Ruth Plumly Thompson เขียนกันแน่ วิธีการตรวจสอบคือ การนำหนังสือที่แน่ใจแล้วว่าเขียนโดย Baum และที่เขียนโดย Thompson แยกเป็นข้อมูลสองชุด และเลือกคำที่ไม่แปรไปตามเนื้อเรื่องแต่เป็นคำที่สามารถใช้บ่งบอกลักษณะสไตล์การเขียนของแต่ละ คนได้ เช่น คำกลุ่มคำไวยากรณ์ ใช้คำเหล่านี้ที่ใช้มากสุด 50 คำมาเป็นตัวแทนสร้างเว็กเตอร์ 50 มิติ ของ 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.

# โปรแกรม Topic Modeling

เป็นโปรแกรมสำหรับวิเคราะห์เอกสารต่าง ๆ ที่มี เพื่อมองหาว่าในเอกสารนั้นมี topic หรือหัวข้ออะไร บ้าง และหัวข้อนั้น ๆ แสดงออกผ่านทางรูปคำอะไรบ้าง มีโปรแกรมที่สามารถดาวน์โหลดมาใช้ได้ฟรี ทั้งบน Windows และ Mac เช่น โปรแกรม GUI สำหรับ MALLET Topic Modeling (<u>https://github.com/senderle/topic-modeling-tool</u>) 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 ก็ จะสามารถสร้างเมทริกซ์ขนาด d x n ได้ ซึ่งเมทริกซ์ d x n นี้ สามารถมองได้ว่าเป็นผลมาจากการ รวมกันของเทริกซ์ขนาด d x k และ k x n ได้ ซึ่งถ้าเรามองว่า k นี้คือจำนวน topic ทั้งหมดที่มี ในข้อมูล เราก็สามารถหาคำตอบนี้ได้ เพราะจะได้ผลที่บอกว่า คำที่เกี่ยวข้องกับ topic ต่างๆ (1..k) มีคำอะไรบ้าง เกี่ยวข้องมากน้อยเพียงใด ส่วนผลอีกตารางจะบอกว่าเอกสารแต่ละขิ้นมีความ เกี่ยวข้องกับ topic ต่างๆ กี่เรื่องและแกี่ยวข้องมากน้อยแค่ไหน อัลกอลิทึมพื้นฐานที่มักใช้กันในการ ทำ topic modeling คือ latent Dirichlet allocation (LDA; Blei, Ng, and Jordan 2003)



เพื่อให้เห็นผลที่ชัดเจนการการทำ topic modeling ในตัวอย่างนี้จึงทดลองใช้ไฟล์ข้อมูลจาก โดเมนที่แตกต่างกันขัดเจน 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 เป็นสิ่งที่ผู้วิจัยต้องใช้ความรู้ของตนระบุชื่อ เอง ถ้าต้องการสำรวจผลที่ได้ว่า จำนวน topic ที่กำหนดและจำนวนรอบการ run นั้นให้ผลดีพอ หรือยัง ก็ควรเปิด folder output\_html เพราะมี link ที่โยงข้อมูลให้คลิกดูได้สะดวกมากกว่า ส่วน ใน output\_csv จะเห็นไฟล์ csv ที่สรุปความสัมพันธ์ระหว่าง word, topic, doc เมื่อเปิดดูไฟล์ top\_in\_doc.csv จะเห็นว่าแต่ละ doc นั้นมีเนื้อหาเกี่ยวกับ topic อะไรมากน้อยกว่ากันโดย คำนวณมาเป็นน้ำหนัก จากข้อมูลที่นำมาทดสอล จะเห้นว่าไฟล์เนื้อหาทางด้านเศรษฐกิจถูกจัดว่ามี น้ำหนักเป็น topic 1 และ 6 มากสุด ไฟล์สุนทรพจน์มีเนื้อหาหนักไปทาง topic 5 ไฟล์นวนิยายมี เนื้อหาเป็น topic 7 ส่วนไฟล์ top\_in\_word.csv แสดงถึงรายการคำสำคัญในแต่ละ topic ที่วิเคราะห์มาโดยเรียง ตามลำดับความสำคัญของคำต่าง ๆ

TopicMod	elingTool
'Corpus/test-topic-model	🚭 Input Dir
/Users/macbook/Desktop	🚭 Output Dir
Number of topics: 10	Optional Settings
🚱 Lear	n Topics
Cons	ole
ose Input Dir Directory: /User en command cancelled by use ose Stopword File File: /Users	s/macbook/Cloud/Dropbe er. /macbook/Cloud/Dropbo:
porting and Training s could take minutes or days	depending on settings and
mporting From /Users/macbo	ook/Cloud/Dropbox/Corp
llet command: nallet import-dir \	
extra-stopwor remove-stopw token-regex [\	ds /Users/macbook/Clouc ords \ p{L}\p{N}_]+ \
input /Users/n output /Users/ keep-sequence	nacbook/Cloud/Dropbox/ /macbook/Desktop/outpu e
Clear C	onsole

4	A	В	С	D	E	F	G	н	1	J	к	L Fo
1	docld	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 Jel/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

1	А	В	с	D	E	F	G	н	1	J	к	L	м
1	Topic Id	Top Word	i										
2	0	elizabeth	darcy beni	net bingley	jane miss	wickham (	collins lydi	a mother c	atherine h	ope sister	friend fath	ner gardine	r lizzy lor
3	1	percent m	arket rate	1 1999 gro	wth prices	2 2000 rate	es quarter :	1998 4 fina	ncial feder	al debt eco	onomic rea	al inflation	investme
4	2	time day v	world hom	e house fo	und attent	tion visit p	assed busi	ness situat	ion detern	nined brou	ght told su	ubject expe	ected dou
5	3	elinor ma	rianne edv	vard moth	er dashwoo	od jenning	s willough	by miss sis	ter lucy joł	nn brandor	colonel f	errars mido	ileton ho
6	4	anne capt	ain elliot v	ventworth	charles ru:	ssell walte	r mary mu	sgrove loui	isa miss ba	th father e	lizabeth la	dy friend u	uppercros
7	5	people go	vernment	public con	gress laws	american	country po	licy law cit	tizens unit	ed busines	s executiv	e power na	ational fre
8	6	sales distr	rict prices of	demand ac	tivity repo	rted conta	cts continu	ed constru	uction repo	rt strong p	ercent ma	nufacturin	g growth
9	7	lady youn	g dear sir s	ister proje	ect feelings	happy crie	ed morning	g family rep	plied left l	ove pleasu	re mind gu	itenberg le	tter ever
10	8	copyright	rights auth	nor law ow	ner protec	tion prope	rty act lite	rary licence	e moral tra	de united	article inte	ellectual ar	tistic king
11	9	informatio	on confide	nce public	confidenti	ial law def	endant bre	ach emplo	yee emplo	oyer duty p	laintiff ob	ligation na	ture cont
12													

# โปรแกรม Corpus Annotator

โปรแกรม annotator คือโปรแกรมที่ช่วยในการ markup ข้อมูลเพิ่มเติมในตัวบท โปรแกรม บางตัวออกแบบมาเพื่อใช้เหมือนเป็นการทำ note เพิ่มเติมในเอกสาร เช่น โปรแกรม Annotator (<u>http://annotatorjs.org</u>) ที่เขียนด้วยภาษา Javascript ให้ผู้ใช้สามารถเติมข้อความเพิ่มเติมไปใน เอกสารผ่านทางเว็บได้ ไม่ว่าจะเป็นหน้าเว็บที่เป็น html หรือ pdf หรือ EPUB จึงเหมาะสำหรับการ markup ทั่วไปแต่ไม่เหมาะกับการกำกับข้อมูลทางภาษาศาสตร์เพื่อสร้างข้อมูลที่จะนำมาวิเคราะห์ต่อ

โปรแกรม brat rapid annotation tool (http://brat.nlplab.org) เป็นโปรแกรมที่ทำงานบนเว็บสำหรับกำกับข้อมูลทางภาษา เช่น ชื่อเฉพาะ หมวดคำ ความสัมพันธ์ ระหว่างคำ จึงเป็นโปรแกรมที่เหาะกับการกำกับข้อมูลภาษาศาสตร์เพื่อนำมาวิเคราะห์ต่อมากกว่า

โปรแกรมนี้ต้องติดตั้งบน server ทำงานด้วยโปรแกรม Python 2.5 ขึ้นไป (ใช้บน Python 3 ไม่ได้) ขั้นตอนนี้จะต้องให้ผู้ดูแล server เป็นคนติดตั้งให้ก่อนจึงจะใช้งานได้ เมื่อติดตั้งเรียบร้อย ผู้ที่ ต้องการใช้ จะต้องมีไฟล์ข้อมูล และไฟล์ configuration ที่กำหนดแท็กต่างๆ และความสัมพันธ์ที่ จะมีระหว่างแท็กได้ 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 อาจมีลำดับชั้นได้ ด้วยการจัดเรียงแล้วเคาะแท็บหน้ากลุ่มเป็นลูกกลุ่ม



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

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

[events] ใช้กำหนดความสัมพันธ์ของ Entity หรือ Relation ที่มีใน Event นั้น จะมีองค์ประกอบหนึ่ง ตัวหรือมากกว่าสองก็ได้ มีลักษณะคล้ายยการกำหนด Frame ของเหตุการณ์ว่าประกอบด้วยอะไร บ้าง องค์ประกอบเขียนในรูป Role:Type เช่น Participant1:Person Org:Company เช่น ตัวอย่างนี้



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

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

สำหรับ configuration อื่นเป็นเรื่องการกำหนดการแสดงผลหน้าจอและสีต่างๆ ที่แสดง ราย ละเอียดสามารถดูได้จากเว็บโครงการเอง Brat annotation นี้ได้ถูกใช้สำหรับกำกับข้อมูลทาง ภาษาศาสตร์หลายเรื่อง เช่น การกำกับชื่อเฉพาะ การกำกับโครงสร้างต้นไม้แบบพึ่งพา การกำกับ Event ดังตัวอย่างที่แสดงข้างล่างนี้





RESULTS: Angiostatin	-Regulation Cell inhibited endothelial	Cell_proliferatio	n by 50-60%.
Drug/comp <sup>Cause</sup> Regul Thalidomide had no direc	t effect on endothelial c	ells.	
GGP Drug/con Angiostatin and thalidom	ide both inhibited	PathF Th Growth	by about 55%.

# 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 <u>http://127.0.0.1:8001</u> จะเห็นหน้าจอของ Brat ขึ้นมา จะมีตัวอย่างให้ทดลองใช้ ให้ login เป็น admin รหัสผ่าน adminbrat (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 เรา ต้องการข้อมูลที่ผ่านการตัดคำและใส่ POS ตามแบบ UD เราต้องเตรียมไฟล์ ann กับ .txt ให้ โปรแกรม 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|น่า วิตก/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|อย่าง/NOUN|ที่/SCONJ|พุทธศาสนิกชน/VERB|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|ก៏/ 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|เทตุ/NOUN|อธิบาย/VERB|ไม่ได้/AUX|จึง/SCONJ|นับถือ/VERB|<s/>/PUNCT|<Fail>ที่นับถือ เพราะขลั</Fail>/NOUN|<s/>

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 ข้อมูลที่

กำกับจะถูกเติมเพิ่มในไฟล์ .ann เช่น

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

ผลลัพธ์ที่ได้สามารถแปลงไปเป็น format อื่นๆ ตามที่ต้องการ เช่น แปลงเป็น CONLLU โดยการใช้ โปรแกรมแปลง เช่น brat2conllu.py (Chonlathorn Kwankajornkiet) -i บอกชื่อ folder ที่เก็บ ข้อมูล brat annotation -o บอกชื่อไฟล์ที่เป็น output -p บอกให้สร้างไฟล์เก็บข้อมูลที่เป็น nonprojective 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	ข้อควรปฏิบั	ดิ _	NOUN	NOL	JN	_	0	ROOT _	SpaceAfter=No
2	เพื่อ _	ADP ADP	_ 3	Х	_	Spac	eAfte	r=No	
3	ความ _	NOUN	NOUN	Prefix	k=Yes	1	Х	_ SpaceAfter	r=No
4	ปลอดภัย	_ VERB	VERB_	3	Х	_	Spac	eAfter=No	
5	หาก _	SCONJ	SCONJ	_	6	Х	_	SpaceAfter=No	
6	เกิด _	VERBVERB	_ 1	Х	_	Spac	eAfte	r=No	
7	วิกฤต _	NOUN	NOUN	_	6	Х	_	SpaceAfter=No	
8	น้ำท่วม	_ NOU	N NOU	IN	_	7	Х	_ SpaceAfter	r=No

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

กรณักำกับข้อมูลแล้วไม่ได้เป็น dependency tree ออกมา ไฟล์ที่มีชื่อ -non-projective.conllu จะ เก็บข้อมูลที่มีปัญหาเหล่านี้ 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 ที่รับข้อมูลกำกับภาษาใดก้ได้ นำมา 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 และให้นำโมเดลนี้ไปใช้กับข้อมูลใหม่ต่อ ไปได้ วิธีการ parse ก็ใช้คำสั่งลักษณะเดียวกัน ในตัวอย่างข้างล่าง s1.conllu เป็นไฟล์ input ที่ คำแต่ละคำพร้อม 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 (<u>https://portal.catma.de/catma/</u>)

เป็นโปรแกรมสำหรับใช้กำกับข้อมูลใน corpus พร้อมช่วยวิเคราะห์ข้อมูลที่กำกับ ตัวโปรแกรม อยู่ที่ server ของ University of Hamburg ประเทศเยอรมัน CATMA เอง เป็น web-based โปรแกรมจึงสะดวกสำหรับผู้ใช้โดยทั่วไป เพียงแค่สมัครใช้งานก็สามารถเข้าไปทำงานได้เลย จะ ทำงานคนเดียวหรือแซร์งานทำด้วยกันหลาย ๆ คนก็ได้

CATMA สนับสนุนการกำกับข้อมูลแบบ external stand-off markup คือตัวข้อความกับ ข้อมูลที่กำกับถูกแยกจากกัน แต่เชื่อมโยงโดยาร cross-link ไปที่ตัว original text ได้ แท็กต่าง ๆ ที่ ใช้สามารถกำหนดเองได้ตามวัตถุประสงค์ของงานที่ต้องการ เช่น แท็ก named entities หรือแท็ก เพื่อใช้ในการวิเคราะห์ metaphor การกำกับข้อมูลทำง่าย ๆ ด้วยการเลือกข้อความส่วนที่ต้องการ แล้วบอกว่าต้องการแท็กเป็นอะไร ข้อมูลที่กำกับแล้วสามารถนำออกมาเป็น XML ไฟล์ตามมาตรฐาน ของ TEI ได้

# ขั้นตอนการใช้งานโดยคร่าว ๆ มีดังนี้

Manage Resource ขั้นแรกให้สร้าง corpus โดยการ upload ไฟล์ต่าง ๆ ที่ต้องการทำงานขึ้นไป อาจใช้วิธีการ upload ไฟล์ที่มีในเครื่องหรือระบุ url ที่มีข้อมูลนั้นก็ได้ หลังจาก upload ไฟล์เข้าไป แล้ว ขั้นต่อไปคือ กำหนดแท็กที่ต้องการใช้ในการกำกับข้อมูล ให้เลือก Create Tag Library แล้วจึง Open Tag Library ที่สร้างขึ้นมา จากนั้นกำหนดชื่อแท็กเซ็ตที่ต้องการ ภายในแท็กเซ็ตจึงตั้งชื่อ แท็กแต่ละอันที่จะใช้ในการแท็ก พร้อมระบุสีที่ต้องการให้แสดงในหน้าจอ แท็กต่าง ๆ ที่สร้างขึ้นมานี้ คือสิ่งที่จะไปใช้กำกับข้อมูลส่วนที่ต้องการ เช่น ในงาน NER หรือ named entities recognition เราต้องการกำกับข้อมูลชื่อสามประเภท คือ ชื่อบุคคล ชื่อองค์กร และชื่อสถานที่ ก็เลยสร้างแท็กไว้ สามตัวในเซ็ตของ NER

CATMA5.0 Manage Resources Manage Ta	rgs Annotate Analyze Visualize About Terms Of Use Imprint Privacy State
Repositories Overview         CATMA DB Repository ×           Document Manager	
Corpora	Documents
All documents	<ul> <li>ap20010619 AP-Japan-Markets Tokyo Stocks Fall</li> </ul>
test	<ul> <li>ap20010701 Tokyo Stocks Decline</li> </ul>
	ap20010709 Big Movers in the Stock Market
Create Corpus More actions ~	Open Annotations Add Document More actions Y

sets	Tag Color	
named entity		C Reload Tagsets
organization		Load Tagset into currently active
♦person		Document
location		A Transf
		A collect
		Create Tagset
		Remove Tagset
		Edit Tagset
		◆ Tag
		Create Tag
		Create Tag Remove Tag

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

CATMA5.0 Manag	e Resources Manage Tags	Annotate Analyze Visualize		About Terms Of Use Impri	int Privacy Statement Manual 🛛 awir
DH-AI x			Active Tagsets	Active Annotations	
1. The History of Humanities Computing Susan Hockey			Open Tagset		Ð
Introduction Tracing the history of any interdisciplinary questions. What should be the scope of th on the development of the activity? What disciplines? Does a straightforward chron Might there be digressions from this, which	academic area of activity raises a n re area? Is there overlap with relater has been the impact on other, perha ological account do justice to the de h could lead us into hitherto unexplo	umber of basic d areas, which has impacted ups more traditional, velopment of the activity? red avenues? Each of these	Tagsets	Tag Color	◆ Tag Create Tag Remove Tag
questions could form the basis of an essi the approach taken is to present a chrono computing. Within this, the emphasis is or has been made or where work done with substantially within other disciplines.	ey in itself but within the space and c logical account which traces the dev n highlighting landmarks where signi in humanities computing has been a	context available here, velopment of humanities ficant intellectual progress dopted, developed or drawn on	Writable Annotation	Collection:	Remove Annotation Edit Property values
It is not the place of this essay to define w within this Companion indeed sends plan with the applications of computing to rese III I /733	hat is meant by humanities computi ty of signals about this. Suffice it to s arch and teaching within subjects th H Analyze Document	ng. The range of topics any that we are concerned at are loosely defined as			Annotation info Collection Path
Open Tagset			+ × 1		
Select a Tag Library:					
Tag Libraries		Create Tag Lit	orary		
Example Tag Library					
Metaphor					
NER		1			
Select a Tagset:					
Tagsets	Tag Color	Load Tagset			
<ul> <li>Anamed entity</li> </ul>		currently			
organization		Document			
♦person					
Iocation		Tagset			

การกำกับข้อมูลทำได้ง่าย ๆ โดยการเลือกส่วนของ text ที่ต้องการใน Document นั้น แล้วไปกด เลือกแท็กที่ต้องการในตารางของ Tagset ให้กดที่ตัวไอคอนสีของแท็กนั้น จะเห็นว่าข้อความที่ ถูกกำกับจะมีการระบายสีแท็กนั้นไว้ข้างใต้ให้เห็น ให้ทำการแท็กข้อมูลที่ต้องการไปจนครบใน Document นั้น แล้วจึงจะเข้าสู่ขั้นตอนต่อไป คือ การ Analyze Document

CATMA5.0 Manage Resources Manage Tags Annotate Analyze Visualize	Niteatos Imprint Privacy Of Statemy Use	Manual (	awiroteRgo 	nail.com out
DH-All x 1949, an Italian Jesuit priest, Father Roberto Busa, began what even to this day is a monumental task:	Active Tagsets	Active Annotations		
to make an index verborum of all the words in the works of St Thomas Aquinas and related authors, totaling some 11 million words of medieval Latin. Father Busa imagined that a machine might be able to	Open Tagset			Ð
help him, and, having heard of computers, went to visit Thomas J. Watson at IBM in the United States in	Tagsets	Tag Color		
search of support (Busa 1980). Some assistance was forthcoming and Busa began his work. The entire texts	<ul> <li>Anamed entity</li> </ul>			Tag
were gradually transferred to punched cards and a concordance program written for the project. The	organization			
intention was to produce printed volumes, of which the first was published in 1974 (Busa 1974).	person			Create Tag
	location			Remove
A purely mechanical concordance program, where words are alphabetized according to their graphic forms (sequences of letters), could have produced a result in much less time, but Busa would not be satisfied	Writable Annotation Col	lection: DH-All,	awirote@gmail.c	om_2018-10-16T06:40:1
with this. He wanted to produce a "lemmatized" concordance where words are listed under their dictionary	Annotation	* Col	Remove Ar	notation
headings, not under their simple forms. His team attempted to write some computer software to deal with this and, eventually, the lemmatization of all 11 million words was completed in a semiautomatic way with	person		Edit Prope	rty values
human beings dealing with word forms that the program could not handle. Busa set very high standards for his work. His volumes are elegantly typeset and he would not compromise on any levels of scholarship in order to get the work done faster. He has continued to have a profound influence on humanities computing,			Annotation	Info
● H 4 2 /733 → H Analyze Document 1 _ 100 □ 🖽 🗸			Collection	DH-AI_awirote@gmail.
			Path	/person

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

ในขั้นแรกสามารถเลือกจาก wordlist เพื่อจะได้เห็นว่าใน document นั้นมีคำอะไรมากน้อยแค่ ไหน เลือกคำที่ต้องการ ก็จะเป็นคำนั้นพร้อมบริบทซ้ายขวา (kwic) ในหน้าจอด้านขวา เรา สามารถใช้ขั้นตอนนี้กำกับข้อมูลคำหรือวลีที่เห็นให้เป็นแท็กที่ต้องการด้วยการเลือก Annotate selected result โดยเลือกแถวที่ต้องการกำกับข้อมูลแท็ก แล้วลากแท็กที่ต้องการมายังหน้าต่าง kwic นั้น ตัวคำค้นก็จะถูกแท็กด้วยสิ่งที่ลากมา ทำให้ไม่ต้องเสียเวลากำกับคำหรือวลีนั้นทีละคำก็ได้ หรือจะใช้ประโยชน์เป็นเหมือน concordance โปรแกรมดึงคำพร้อมบริบทออกมาใช้ก็ได้

Query freq>0 Query Builder	Vordlist 😧		Execute Query	Documents and annotations co DH-All	nstraining this	+ New Query
Result by Phrase	Result by Tag					
Phrase	Frequency	Visible in Kwic	Document/Anr	Left Context	Keyword	Right Context
analyzer	1	0	DH-AII	, Force, Moralite,	Intelligence	." Of course,
Intelligence	14	0	DH-AI	, Force, mechant,	Intelligence	. PONCT[pt]
▶ ordinal	1	0	DH-AI	In The Foundations of Artificial	Intelligence	, ed. D.
<ul> <li>Their</li> </ul>	12		DH-AI	Intelligence without Representation. A	Intelligence	Journal 47: 139 59
<ul> <li>Existing</li> </ul>	3	0	DH-AII	Thirteenth National Conference on Art	Intelligence	: 598 603. Cunningham
<ul> <li>Carroll</li> </ul>	4		DH-AII	Thirteenth National Conference on Art	Intelligence	2: 1089 74.
<ul> <li>ceases</li> </ul>	1	0	DH-AII	Spolsky), and Artificial	Intelligence	theory (CHUM 1993,
Total count: 22.349	Total frequency: 314.362		DH-AII	is the rise of Artificial	Intelligence	(AI) the effort
	197 Select all Deselect all	Select all for Kwic Deselect all	DH-AI	in the Al (Artificial 30 Anno 5 token(s) context	Intelligence	soults Select all

การค้นให้ใช้ Query Builder ได้ แล้วเลือกค้นคำหรือค้นแท็ก หรือจะค้นคำเขียนคล้ายคลึง คำปรากฏ ร่วมกัน (collocation) ก็สามารถระบุผ่านการใช้ Query Builder ได้

Query				Documents and annotations co	instraining this	search + New Query
tag="person%"			Execute Query	<ul> <li>DH-All</li> </ul>		
Query Builder Wordlin Result by Phr. 40 Result	O at by Tag					
Phrase	Frequency	Visible in Keic	Document/Anr	Left Context	Keyword	Right Context
<ul> <li>Wisbey</li> </ul>	1		DH-All	Middle High German texts (	Wisbey	1963). In the
Father Busa	2		DH-All	words of medieval Latin.	Father Busa	imagined that a machine m
<ul> <li>St Thomas Aquinas</li> </ul>	1		DH-AII	English, and Italian.	Father Busa	himself was the first recipie
<ul> <li>Busa</li> </ul>	6		DH-AI	words in the works of	St Thomas Ar	and related authors, totalin
Thomas J. Watson	1		DH-All	Some assistance was forthcoming an	Busa	began his work. The
Dolores Burton	1		DH-AII	was published in 1974 (	Busa	1974). A purely
Burton	1		DH-AII	in search of support (	Busa	1980). Some assistance
Father Roberto Busa	1		DH-All	much less time, but	Busa	would not be satisfied with
Total count: 9	Total frequency: 15	-	DH-AII	cum hypertextibus") (	Busa	1992) and was accompanie
	Select all Deselect all for K	t all Deselect all for Kwic		O 30 a token(s) context	otate selected re	sults Select all

ส่วนสุดท้ายคือ Visualize ซึ่ง ณ ปัจจุบันยังไม่มีอะไรมาก มีเพียงการแสดงผลของการกระจายความถี่ ของคำ และข้อมูลการปรากฏของคำติดกัน (bigram) ในรูปของ double tree โดยเราต้องเลือกคำที่ ต้องการก่อน แล้วกดไปคอนที่อยู่ข้างล่าง ถ้ามีความต่างของความถี่มากจะเห็นความใหญ่เล็กของคำ ชัดเจนขึ้น

Ø by Tag Frequency	Visible in Kwic		
Py Tag Frequency	Visible in Kwic		
Frequency	Visible in Kwic		
Frequency	Visible in Kwic		
10			
10	0		
22			
38			
12			
7			
9			
8			
Total frequency: 283,864			
Select all Deselect all Selec	t all for Kwic Deselect a		
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	22 38 12 7 9 3 Total frequency: 283,864 Select all Deselect all Select		



โปรแกรม CATMA เหมาะสำหรับการกำกับข้อมูล เพื่อนำข้อมูลที่กำกับนั้นมาหา pattern หรอือที่ เรียกในเว็บอธิบายโปรแกรมว่าคือ model คือการพยายามหาข้อสรุปบางอย่างจากการวิเคราะห์ ข้อมูลที่กำกับไว้ เป็นงานเน้นด้าน qualitative literary หรือ text studies ที่อาศัยการวิเคราะห์ text ด้วยผู้วิจัยมากกว่าจะเป็นงานเน้นข้อมูลมาก ๆ แบบ text mining และเนื่องจากการกำกับ ข้อมูลอาศัย tagset ที่สร้างขึ้นได้ตามที่เรากำหนด เราจึงสามารถกำหนด tageset หลายชุดกำกับ ข้อมูลเดียวกันในระดับต่างๆ ได้

## การแปลงข้อมูล XML annotation

การกำกับข้อมูลด้วย XML เป็นมาตรฐานที่ได้รับการยอมรับในการใช้กำกับคลังข้อมูลภาษา จึงมี โมดูลหรือเครื่องมือที่ใช้ในการตรวจสอบการกำกับว่า 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 สามารถใช้วิธีนี้ถ้ารู้แน่ว่าไฟล์ เดิมมีส่วนของข้อมูลที่ไม่ 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 สามารถเขียนข้อมูลที่ได้ เป็นไฟล์ 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 ที่ไม่มาก 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 เพราะหากมีการสูญหาย ของข้อมูลไม่ว่าจะในขั้นตอนใด ก็อาจทำให้งานวิจัยนั้นล้มเหลวไม่สำเร็จตามเวลาที่กำหนด

# การสำรองข้อมูลและการรักษาความปลอดภัย

### การสำรองข้อมูลและการรักษาความปลอดภัย

การสำรองข้อมูล เป็นกิจวัตรที่ผู้วิจัยควรต้องกระทำอย่างสม่ำเสมอทุกวัน หรืออย่างน้อยทุก สัปดาห์ ข้อมูลควรจะต้องมีการสำรองไว้หลายชุด และเก็บไว้ในหลายที่ทั้งที่บ้าน ที่ทำงานหรือ มหาวิทยาลัย หรือที่เป็น cloud storage เช่น dropbox, google drive หรือ onedrive หรือบางครั้ง อาจสำรองข้อมูลเป็นหลาย ๆ version คือไม่ได้ลบไฟล์เก่าทิ้ง แต่เก็บไฟล์ข้อมูลใหม่ วิธีนี้จะทำให้มีข้อ มูลก่าเมื่อสัปดาห์ก่อน เมื่อเดือนก่อน อยู่ตลอดเวลาการวิจัย ทำให้หากต้อการย้อนกลับไป ณ เวลา เดิมก็สามารถทำได้ หากจะทำเช่นนี้สิ่งที่จะต้องคำนึงคือการวางระบบให้ชัดเจนว่าข่อมูลชุดไหนเป็น ของช่วงเวลาไหน

ในกรณีที่ข้อมูล (ทั้งที่เป็นข้อมูลภาษา ข้อมูลงานวิจัย ฯลฯ) มีจำนวนมาก อาจสะดวกกว่าถ้าใช้ วิธีการ sync ข้อมูลแทนที่จะ copy หรือ backup ไฟล์ทั้งหมด โดยสามารถใช้โปรแกรม เช่น Synkron (<u>http://synkron.sourceforge.net</u>/) ซึ่งเป็นโปรแกรมฟรีสำหรับทั้ง PC และ Mac Allway Sync (<u>http://allwaysync.com</u>/) สำหรับ PC ซึ่งมีทั้งที่ฟรีและต้องซื้อ หรือโปรแกรมที่ขายอย่าง GoodSync ได้ <u>http://www.goodsync.com/</u> แต่ทั้งนี้ต้องรอบคอบและระวังไม่ให้ข้อมูลเก่าทับ ข้อมูลใหม่ หรือลบข้อมูลใหม่ทิ้ง ควรทดลองใช้โปรแกรมกับข้อมูลทดสอบอื่นๆ ก่อนจนเข้าใจวิธีการใช้ โปรแกรมดีแล้วจึงจะใช้กับข้อมูลจริงได้

การ backup ระบบคอมพิวเตอร์ก็จำเป็นกรณีที่เครื่องที่ใช้มีปัญหา จะได้ restore กลับได้โดย ง่าย 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 (<u>http://www.macrium.com/ReflectFree.asp</u>), สำหรับ Mac สามารถใช้โปรแกรม SuperDuper (<u>http://www.shirt-pocket.com/SuperDuper/SuperDuperDescription.html</u>) อย่างไรก็ตาม ก่อนใช้งานโปรแกรมเกี่ยวกับการจัดการ partition ให้ทดลองให้แน่ใจว่าขั้นตอนเป็น อย่างไร ก่อนจะใช้งานกับข้อมูลที่สำคัญ ในการ backup ควรใช้ external harddisk ในการ backup drive เมื่อมีปัญหาก็สามารถ restore จาก external harddisk นั้นได้ โปรแกรมประเภทนี้จำเป็นจะ ต้องให้เราสร้าง CD สำหรับ boot เครื่องขึ้นมาเองเวลาที่ต้องการ restore ให้สร้าง boot CD แล้วเก็บ ไว้ในที่ที่ปลอดภัย หรือถ้าสามารถสร้าง backup บน usb harddisk แล้วกำหนดให้เครื่อง boot จาก usb drive ได้เลยก็จะสะดวก

# ไวรัสคอมพิวเตอร์

#### ไวรัสคอมพิวเตอร์

ผู้วิจัยควรมีโปรแกรม Antivirus ติดตั้งและต้องตั้งค่าให้ update ตัวเองอยู่ทุกวัน เครื่องที่ใช้จึง ควรต่ออินเทอร์เน็ตอยู่เสมอเพื่อให้สามารถนำข้อมูลใหม่มา update ได้ โปรแกรมป้องกันไวรัสที่ได้ แถมมากับเครื่องมักมีระยะเวลาให้ใช้ได้ช่วงหนึ่ง เช่น Norton, McAffee เมื่อพ้นกำหนดแล้วถ้าไม่จ่าย เงิน จะไม่ได้ update ตัวเอง จึงเป็นอันตรายที่ไวรัสใหม่ๆ จะเข้ามาได้ง่ายๆ หากเราไม่รู้และไม่ได้จ่าย เงินรายปีต่อ ทางที่ดีคือให้ใช้โปรแกรม Anti virus ที่มี free version ให้ใช้ อย่างน้อยหนึ่งตัว เช่น AVG (<u>http://free.avg.com/</u>), Avast (<u>http://www.avast.com/free-antivirus-download</u>), Avira (<u>http://www.free-av.com</u>/) และที่สำคัญ จะต้องคอย update โปรแกรมต่างๆ บนเครื่อง ไม่ ว่าจะเป็น windows, Internet Explorer, Firefox, etc. เพราะโปรแกรมเก่าจะมีช่องว่างให้ไวรัสหลุด เข้ามาได้ง่าย

ไวรัสคอมพิวเตอร์เป็นโปรแกรมที่เขียนขึ้นมาเพื่อแฝงตัวในคอมพิวเตอร์และแพร่กระจายไป คอมพิวเตอร์เครื่องอื่นๆ ไวรัสสามารถแพร่ผ่านสื่อต่างๆ diskette, flash drive, CD, internet ผลก ระทบที่เกิดจากการติดไวรัสมีความแตกต่างกันไปหลายแต่ว่าไวรัสนั้นถูกเขียนมาเพื่อให้ทำอะไร

โปรแกรมไวรัสมีหลายประเภท มีที่เป็นแบบไวรัสคือต้องอาศัยไฟลือื่นเป็น host ในการแพร่ ที่ เป็นแบบ Trojan คือแฝงมาในโปรแกรมที่คิดว่านำมาใช้งานหนึ่งแต่แอบซ่อนโปรแกรมร้ายไว้ ที่เป็น แบบ worm คือกระจายตัวได้เองพยายามแพร่ตัวเองผ่านเตรือช่ายจนระบบล่มไป

ปัจจุบันไวรัสมาในรูปแบบต่างๆ ส่งผ่านอีเมล์หรือเว็บ เพื่อหลอกผู้ใช้ให้เปิดดู เมื่อเปิดก็จะถูก จู่โจมและติดทันที บางครั้งมาในรูปการหลอกว่าเครื่องเราได้ติดไวรัส หากต้องการ remove หรือ ป้องกันให้ install โปรกแกรมป้องกันที่ให้โดยด่วน วิธีป้องกันคือไม่เปิดไฟล์ที่ไม่ใช่โปรแกรมที่เรารู้จัก และได้ติดตั้งเองในเครื่อง ไม่เปิดไฟล์คนแปลกหน้า หรือแม้แต่ไฟล์ที่ส่งมาจากคนรู้จัก หากเนื้อความ ในจดหมายไม่มีข้อความอะไรเป็นสาระมากพอให้สงสัยว่า อีเมล์นั้นถูกส่งมาด้วยโปรแกรมไวรัสโดย อาศัยรายซื่ออีเมล์ที่คนนั้นเก็บไว้ในเครื่อง

Get Mall Write Addre	55 8	DOR Reply Reply All Forward Tag Delete Arris Print	G. O	2.0	tire Message	
All Folders ***  All Folders ***  Lung573  Lung741  MR02546  ANR02546  ARosearch U  Saved message Soudent53  Thai Thai Thai Thai Thai Thai Thai Tha		8 Subject ENEED YOU IN MY LIFE Regarding your familyIts important Unadroted Transaction Con You Account	Sender     prince     Trask Miller	Date 4/2/2010 3:12 AM 4/20/2010 6:56 AM	<ul> <li>Status</li> <li>Read</li> <li>Read</li> <li>Read</li> </ul>	Size 4 348 - 243
		Greetings to you     Reference no: UK/UA2CT-1309/08.     Reference no: UK/UA2CT-1309/08.     Subject: Unachterised Transaction On Your Account     Form: InSEC ater) - security/sects@online.indo.co.ate.     Date: 5/3/2010 11:22 PM     Te: everote@Online.tb	<ul> <li>Grew Phil</li> <li>CLAINS DEFARTMENT UK LO.</li> <li>LIK LOTTERY ROARD</li> </ul>	5/15/2010 919 PM 5/15/2010 3014 PM 5/22/2010 11:00 PM	Read Read Read	243 29043 29043
		We're sorry, but there appears to be a problem loading the mes	nape <u>Click here to tex again</u>			

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

เราอยากจะแจ้งให้ทราบว่าขณะนี้เรากำลังดำเนินการบำรุงรักษาที่กำหนดและอัพเกรดการให้บริการ เว็บของเราและเป็นผลจากการนี้ HTK4S ไวรัสได้รับการตรวจพบในโฟลเดอร์ที่บัญชีของคุณและบัญชี ของคุณจะต้องมีการอัพเกรดใหม่ของเรา F-Secure รุ่น HTK4S anti-virus/anti-Spam 2013 เพื่อ ป้องกันไม่ให้เกิดความเสียหายต่อไฟล์สำคัญของคุณ เติมคอลัมน์ด้านล่างและส่งกลับมาหรือบัญชี อีเมลของคุณจะถูกระงับชั่วคราวจากการให้บริการของเรา

# ล้มเหลวที่จะทำเช่นนี้ภายใน 24 ชั่วโมงทันทีจะทำให้บัญชีอีเมลของคุณปิดการใช้งานจาก chula.ac.th ฐานข้อมูลของเรา, ลิขสิทธิ์ 2013 chula.ac.th (ค) สิทธิเครือข่ายทั้งหมด

อย่างไรก็ตาม การติดไวรัสอาจมาจากการใช้เครื่องสาธารณะที่มีผู้อื่นนำไฟล์ที่ติดไวรัสมาใช้ เมื่อ เรานำเอา flash drive ไปใช้ก็สามารถติดไวรัสมาได้ flash drive จึงไม่ควรใช้เพื่อเก็บข้อมูลสำคัญ อย่างถาวร 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 ควรมีการ เปลี่ยน password อยู่เสมอ อย่าใช้ password ชุดเดียวกันกับหลายๆ account อย่าตั้ง password ที่ง่ายเกินไป สามารถเดาได้ อย่าตั้งไว้สั้นเกินไป มีผู้แนะนำว่าให้ใช้ password ยาวแต่จำได้ง่าย อาจใช้ คำในภาษามาผสมกันแต่ไม่เป็นวลีที่ใช้ในภาษาเพื่อช่วยให้จำง่าย เช่น "microwave someone Turkish" เป็นต้น

มีตัวอย่างนักคอมพิวเตอร์ ที่ถูก hack Google account, Twitter, Apple ID โดยอาศัยช่องโหว่ ของระบบการตรวจสอบ โดยได้โทรศัพท์ไปที่ Amazon ขอเพิ่มเลขบัตรเครดิตใน account โดยมี ข้อมูลชื่อ ที่อยู่ และ email เท่านั้น หลังจากนั้น โทรไปหา Amazon ว่าลืม password และใช้ข้อมูล ชื่อ email และหมายเลขบัตรเครดิตที่เพิ่งเพิ่มเข้าไป เมื่อสามารถเข้า account Amazon ก็สามารถ เห็นเลขสี่ตัวสุดท้ายบัตรเครดิตแท้จริง 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 โดย ส่งไปที่อีก email ที่เรากำหนดไว้ได้ (ดู

http://www.wired.com/gadgetlab/2012/08/apple-amazon-mat-honan-hacking/)