前言

2017年11月29日　　下午 03:11

INPUT

最主要的檔案:

resource-mentions.tsv 紀錄那些resource(rid)出現在哪些論文(mentionid or mentionid_int)內

resource-mentions-relationships.tsv 紀錄那些resource曾共同出現在那些論文內

resource-metadata.tsv 紀錄每個resource的meta information(如全名 相關網址)

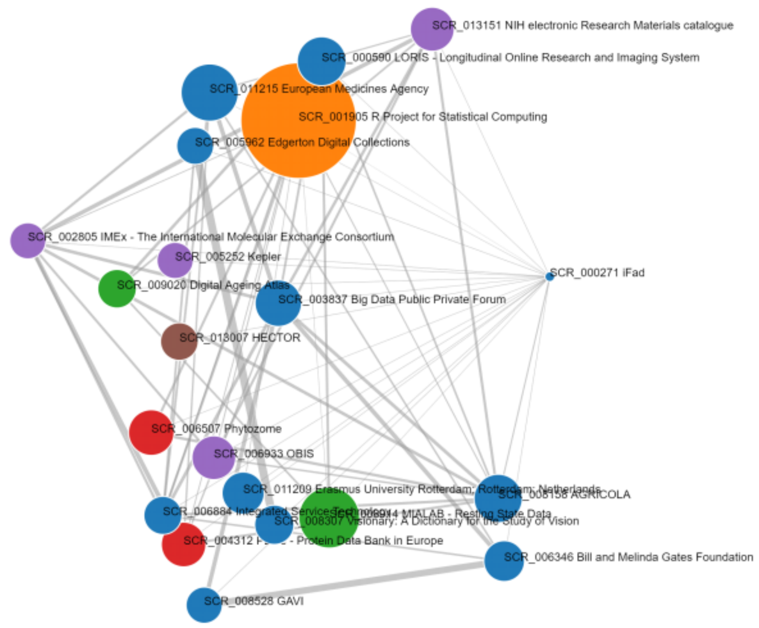
資料更正(除錯)檔:

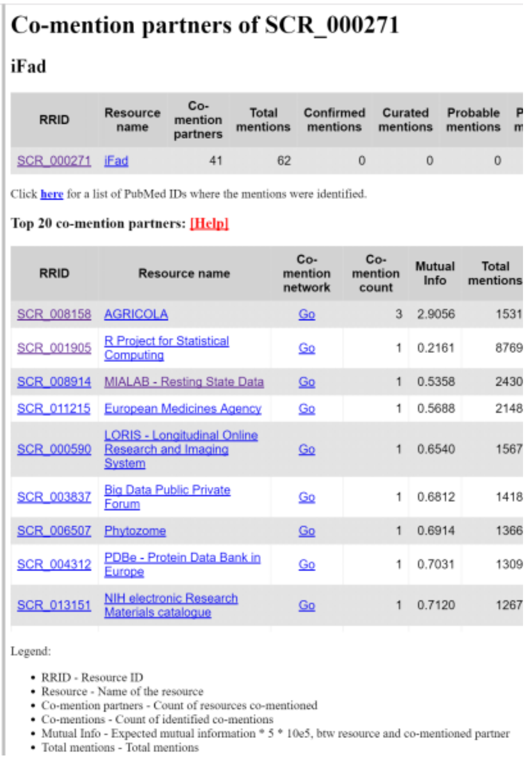
exclusion.tsv 紀錄需要移除的resource id (要從上述的mention & mention-relationship檔刪除相關資料)

resource-duplicates.tsv 紀錄一些相同resource卻不同(多餘)id的資料集合

OUTPUT

為每個resource產生 resource co-mention network (graph)還有相關的co-mention table


SCR_013151 NIH electronic Research Materials catalogue
SCR_000590 LORIS - Longitudinal Online Research and Imaging System
SCR_011215 European Medicines Agency
SCR_001905 R Project for Statistical Computing
SCR_005962 Edgerton Digital Collections
SCR_002805 IMEx - The International Molecular Exchange Consortium
SCR_005252 Kepler
SCR_000271 iFad
SCR_009020 Digital Ageing Atlas
SCR_003837 Big Data Public Private Forum
SCR_013007 HECTOR
SCR_006507 Phytozome
SCR_006933 OBIS
SCR_011209 Erasmus University Rotterdam; Rotterdam; Netherlands
SCR_008158 AGRICOLA
SCR_006884 Integrated Service Technology
SCR_008914 MIALAB - Resting State Data
SCR_008307 Visionary: A Dictionary for the Study of Vision
SCR_004312 PDBe - Protein Data Bank in Europe
SCR_006346 Bill and Melinda Gates Foundation
SCR_008528 GAVI



Co-mention partners of SCR_000271

iFad

RRID	Resource name	Co-mention partners	Total mentions	Confirmed mentions	Curated mentions	Probable mentions	P m
SCR_000271	iFad	41	62	0	0	0	

Click here for a list of PubMed IDs where the mentions were identified.

Top 20 co-mention partners: [Help]

RRID	Resource name	Co-mention network	Co-mention count	Mutual Info	Total mentions
SCR_008158	AGRICOLA	Go	3	2.9056	1531
SCR_001905	R Project for Statistical Computing	Go	1	0.2161	8769
SCR_008914	MIALAB - Resting State Data	Go	1	0.5358	2430
SCR_011215	European Medicines Agency	Go	1	0.5688	2148
SCR_000590	LORIS - Longitudinal Online Research and Imaging System	Go	1	0.6540	1567
SCR_003837	Big Data Public Private Forum	Go	1	0.6812	1418
SCR_006507	Phytozome	Go	1	0.6914	1366
SCR_004312	PDBe - Protein Data Bank in Europe	Go	1	0.7031	1309
SCR_013151	NIH electronic Research Materials catalogue	Go	1	0.7120	1267

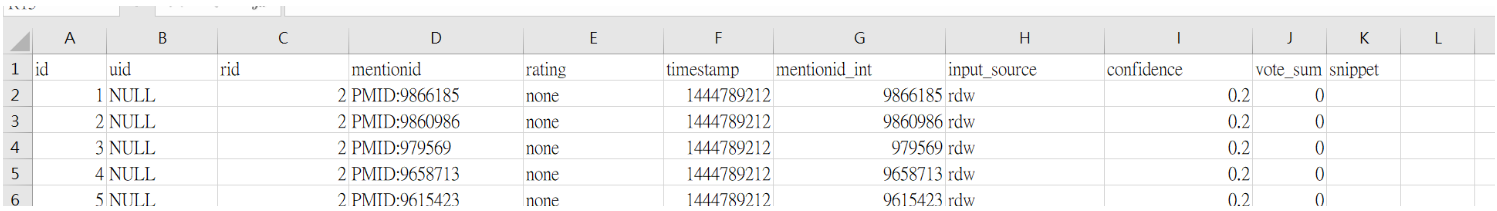
Legend:

- RRID - Resource ID
- Resource - Name of the resource
- Co-mention partners - Count of resources co-mentioned
- Co-mentions - Count of identified co-mentions
- Mutual Info - Expected mutual information * 5 * 10e5, btw resource and co-mentioned partner
- Total mentions - Total mentions

resource-mentions.tsv

2017年11月29日　　下午 03:07

紀錄每個resource出現在那些論文內



	A	B	C	D	E	F	G	H	I	J	K	L
1	id	uid	rid	mentionid	rating	timestamp	mentionid_int	input_source	confidence	vote_sum	snippet	
2	1	NULL	2	PMID:9866185	none	1444789212	9866185	rdw	0.2	0		
3	2	NULL	2	PMID:9860986	none	1444789212	9860986	rdw	0.2	0		
4	3	NULL	2	PMID:979569	none	1444789212	979569	rdw	0.2	0		
5	4	NULL	2	PMID:9658713	none	1444789212	9658713	rdw	0.2	0		
6	5	NULL	2	PMID:9615423	none	1444789212	9615423	rdw	0.2	0		

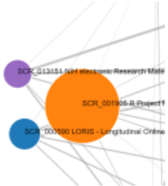
主要欄位:

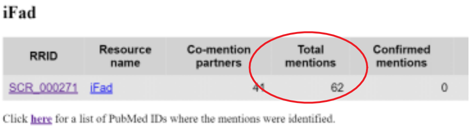
- rid: resource的id
- mentioid: 論文id
- mentionid_int: integer格式的論文id
- confidence: 介於0~1, 越大代表該resource出現在論文的可信度越高 (resource mention多半是透過information extraction (IE) 的程式擷取, 所以有些可能有錯, 而confidence = 1是人工判斷過的, 可信度最高)

主要產出資訊:

- 統計出每個resource的mention count(出現在多少篇論文內),
  Resource co-mention graph與table會使用到該值調整resource node的大小與table欄位 (如右例圖)

因為UCSD會不斷update這個檔案(修正IE程式擷取資料的錯誤),
請先了解秋中設計的db格式,
並設計Python程式來處理tsv檔以產生相關資料.


SCR_013151 NIH electronic Research Mate
SCR_001905 R Project f
SCR_000590 LORIS - Longitudinal Online



iFad

RRID	Resource name	Co-mention partners	Total mentions	Confirmed mentions
SCR_000271	iFad	41	62	0

Click here for a list of PubMed IDs where the mentions were identified.

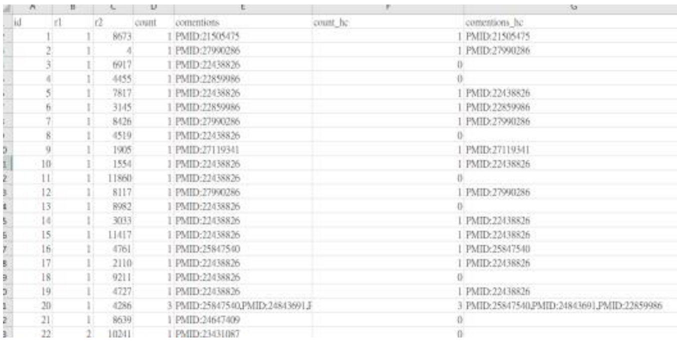
需要注意事項:

- exclusion.tsv紀錄了一些需要移除的resources, 要確定最後的db沒有這些resources
- resource-duplicates.tsv紀錄了一些多餘的resource ids, 要確定將這些多餘id的mentions (co-mention)整合到單一resource id下

resource-mentions-relationships.tsv

2017年11月30日　　上午 08:42

紀錄兩兩resource pair曾出現在那些論文內



id	r1	r2	count	comentions	count_hc	comentions_hc
1	1	8673	1	PMID:21505475	1	PMID:21505475
2	1	4	1	PMID:27990286	1	PMID:27990286
3	1	6917	1	PMID:22438826	0	
4	1	4455	1	PMID:22859986	0	
5	1	7817	1	PMID:22438826	1	PMID:22438826
6	1	3145	1	PMID:22859986	1	PMID:22859986
7	1	8426	1	PMID:27990286	1	PMID:27990286
8	1	4519	1	PMID:22438826	0	
9	1	1905	1	PMID:27119341	1	PMID:27119341
10	1	1554	1	PMID:22438826	1	PMID:22438826
11	1	11860	1	PMID:22438826	0	
12	1	8117	1	PMID:27990286	1	PMID:27990286
13	1	8982	1	PMID:22438826	0	
14	1	3033	1	PMID:22438826	1	PMID:22438826
15	1	11417	1	PMID:22438826	1	PMID:22438826
16	1	4761	1	PMID:25847540	1	PMID:25847540
17	1	2110	1	PMID:22438826	1	PMID:22438826
18	1	9211	1	PMID:22438826	0	
19	1	4727	1	PMID:22438826	1	PMID:22438826
20	1	4286	3	PMID:25847540,PMID:24843691,F	3	PMID:25847540,PMID:24843691,PMID:22859986
21	1	8639	1	PMID:24647409	0	
22	2	10241	1	PMID:23431087	0	

主要欄位:

- r1, r2: resource pair的resource id
- count: 共同出現的論文篇數
- comentions: 共同出現的論文ids

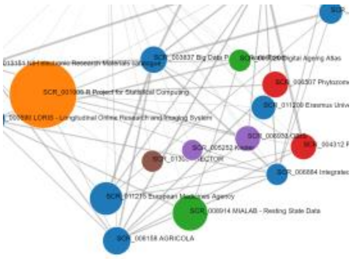
主要產出資訊:

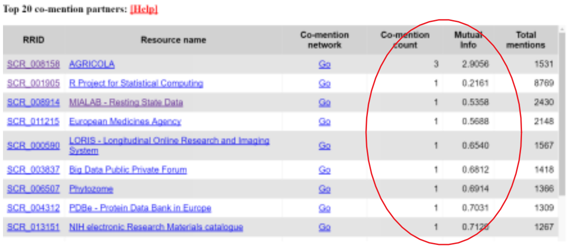
- 結合reosurce-mentions.tsv內resource出現的次數與共同出現的次數來算"**兩兩**"resources的關聯性, 關聯性數值為**expected mutual information (MI)**, 該值用來決定resource graph上link的粗細(見右圖)
- Resource table會列出與當前resource關係最強的reousrces, 且秀出其co-mention count和MI值

因為UCSD會不斷update這個檔案(修正IE程式擷取資料的錯誤),
請先了解秋中設計的db格式,
並設計Python程式來處理tsv檔以產生相關資料.

需要注意事項:

- exclusion.tsv紀錄了一些需要移除的resources, 要確定最後的db沒有這些resources的co-mentions
- resource-duplicates.tsv紀錄了一些多餘的resource ids, 要確定將這些多餘id的mentions整合到單一resource id下, 且相對應的co-mention count與comention PMID要調整


SCR_
013151 NIH electronic Research Materials catalogue
SCR_003837 Big Data P
SCR_ Digital Ageing Atlas
SCR_006507 Phytozome
SCR_001905 R Project for Statistical Computing
SCR_011200 Erasmus Univ
000680 LORIS - Longitudinal Online Research and Imaging System
SCR_006053 CIBS
SCR_004312 P
SCR_005252 Knoter
SCR_013 ECTOR
SCR_006884 Integrated
SCR_011215 European Medicines Agency
SCR_008914 MIALAB - Resting State Data
SCR_008158 AGRICOLA



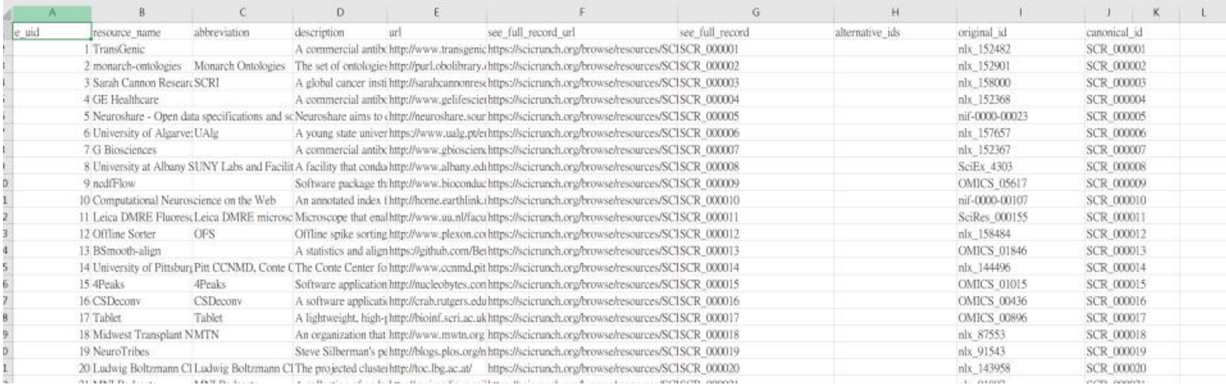
Top 20 co-mention partners: [Help]

RRID	Resource name	Co-mention network	Co-mention count	Mutual Info	Total mentions
SCR_008158	AGRICOLA	Go	3	2.9056	1531
SCR_001905	R Project for Statistical Computing	Go	1	0.2161	8769
SCR_008914	MIALAB - Resting State Data	Go	1	0.5358	2430
SCR_011215	European Medicines Agency	Go	1	0.5688	2148
SCR_000680	LORIS - Longitudinal Online Research and Imaging System	Go	1	0.6540	1567
SCR_003837	Big Data Public Private Forum	Go	1	0.6812	1418
SCR_006507	Phytozome	Go	1	0.6914	1366
SCR_004312	PDBe - Protein Data Bank in Europe	Go	1	0.7031	1309
SCR_013151	NIH electronic Research Materials catalogue	Go	1	0.7128	1267

resource-metadata.tsv

2017年11月30日　　上午 08:54

紀錄每個resource的相關資訊



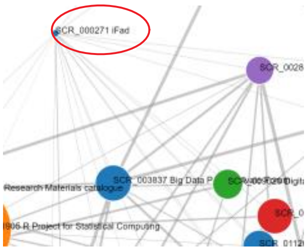
e_uid	resource_name	abbreviation	description	url	see_full_record_url	see_full_record	alternative_ids	original_id	canonical_id
1	TransGenic		A commercial antibc	http://www.transgenic	https://scicrunch.org/browse/resources/SCI	SCR_000001		nlx_152482	SCR_000001
2	monarch-ontologies	Monarch Ontologies	The set of ontologies	http://purl.obolibrary.(	https://scicrunch.org/browse/resources/SCI	SCR_000002		nlx_152901	SCR_000002
3	Sarah Cannon Researc	SCRI	A global cancer insti	http://sarahcannonresc	https://scicrunch.org/browse/resources/SCI	SCR_000003		nlx_158000	SCR_000003
4	GE Healthcare		A commercial antibc	http://www.gelifescien	https://scicrunch.org/browse/resources/SCI	SCR_000004		nlx_152368	SCR_000004
5	Neuroshare - Open data specifications and sc		Neuroshare aims to (	http://neuroshare.sour	https://scicrunch.org/browse/resources/SCI	SCR_000005		nif-0000-00023	SCR_000005
6	University of Algarve:	UAlg	A young state univer	https://www.ualg.pt/er	https://scicrunch.org/browse/resources/SCI	SCR_000006		nlx_157657	SCR_000006
7	G Biosciences		A commercial antibc	http://www.gbioscienc	https://scicrunch.org/browse/resources/SCI	SCR_000007		nlx_152367	SCR_000007
8	University at Albany	SUNY Labs and Facilit	A facility that condu	http://www.albany.edu	https://scicrunch.org/browse/resources/SCI	SCR_000008		SciEx_4303	SCR_000008
9	ncdfFlow		Software package th	http://www.bioconduc	https://scicrunch.org/browse/resources/SCI	SCR_000009		OMICS_05617	SCR_000009
10	Computational Neuroscience on the Web		An annotated index f	http://home.earthlink.(	https://scicrunch.org/browse/resources/SCI	SCR_000010		nif-0000-00107	SCR_000010
11	Leica DMRE Fluoresc	Leica DMRE microsc	Microscope that enal	http://www.uu.nl/facu	https://scicrunch.org/browse/resources/SCI	SCR_000011		SciRes_000155	SCR_000011
12	Offline Sorter	OFS	Offline spike sorting	http://www.plexon.co	https://scicrunch.org/browse/resources/SCI	SCR_000012		nlx_158484	SCR_000012
13	BSmooth-align		A statistics and align	https://github.com/Be	https://scicrunch.org/browse/resources/SCI	SCR_000013		OMICS_01846	SCR_000013
14	University of Pittsburg	Pitt CCNMD, Conte C	The Conte Center fo	http://www.ccnmd.pit	https://scicrunch.org/browse/resources/SCI	SCR_000014		nlx_144496	SCR_000014
15	4Peaks	4Peaks	Software application	http://nucleobytes.con	https://scicrunch.org/browse/resources/SCI	SCR_000015		OMICS_01015	SCR_000015
16	CSDeconv	CSDeconv	A software applicati	http://crab.rutgers.edu	https://scicrunch.org/browse/resources/SCI	SCR_000016		OMICS_00436	SCR_000016
17	Tablet	Tablet	A lightweight, high-{	http://bioinf.scri.ac.uk	https://scicrunch.org/browse/resources/SCI	SCR_000017		OMICS_00896	SCR_000017
18	Midwest Transplant N	MTN	An organization that	http://www.mwtn.org	https://scicrunch.org/browse/resources/SCI	SCR_000018		nlx_87553	SCR_000018
19	NeuroTribes		Steve Silberman's pe	http://blogs.plos.org/n	https://scicrunch.org/browse/resources/SCI	SCR_000019		nlx_91543	SCR_000019
20	Ludwig Boltzmann Cl	Ludwig Boltzmann Cl	The projected cluster	http://toc.lbg.ac.at/	https://scicrunch.org/browse/resources/SCI	SCR_000020		nlx_143958	SCR_000020

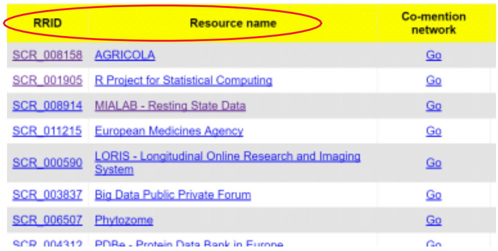
主要欄位:

- e_uid, see_full_record: resource id
- resource_name: 資源名稱
- abbreviation: 資源名稱縮寫
- url: 該資源的官方網址
- see_full_record_url: 該資源於scicrunch的說明網址

主要產出資訊:

- Resource graph與table會使用這些資訊來產生相關圖表 (見右圖)





RRID	Resource name	Co-mention network
SCR_008158	AGRICOLA	Go
SCR_001905	R Project for Statistical Computing	Go
SCR_008914	MIALAB - Resting State Data	Go
SCR_011215	European Medicines Agency	Go
SCR_000590	LORIS - Longitudinal Online Research and Imaging System	Go
SCR_003837	Big Data Public Private Forum	Go
SCR_006507	Phytozome	Go
SCR_004312	PDBe - Protein Data Bank in Europe	Go

因為UCSD會不斷update這個檔案(修正IE程式擷取資料的錯誤),
請先了解秋中設計的db格式,
並設計Python程式來處理tsv檔以產生相關資料.

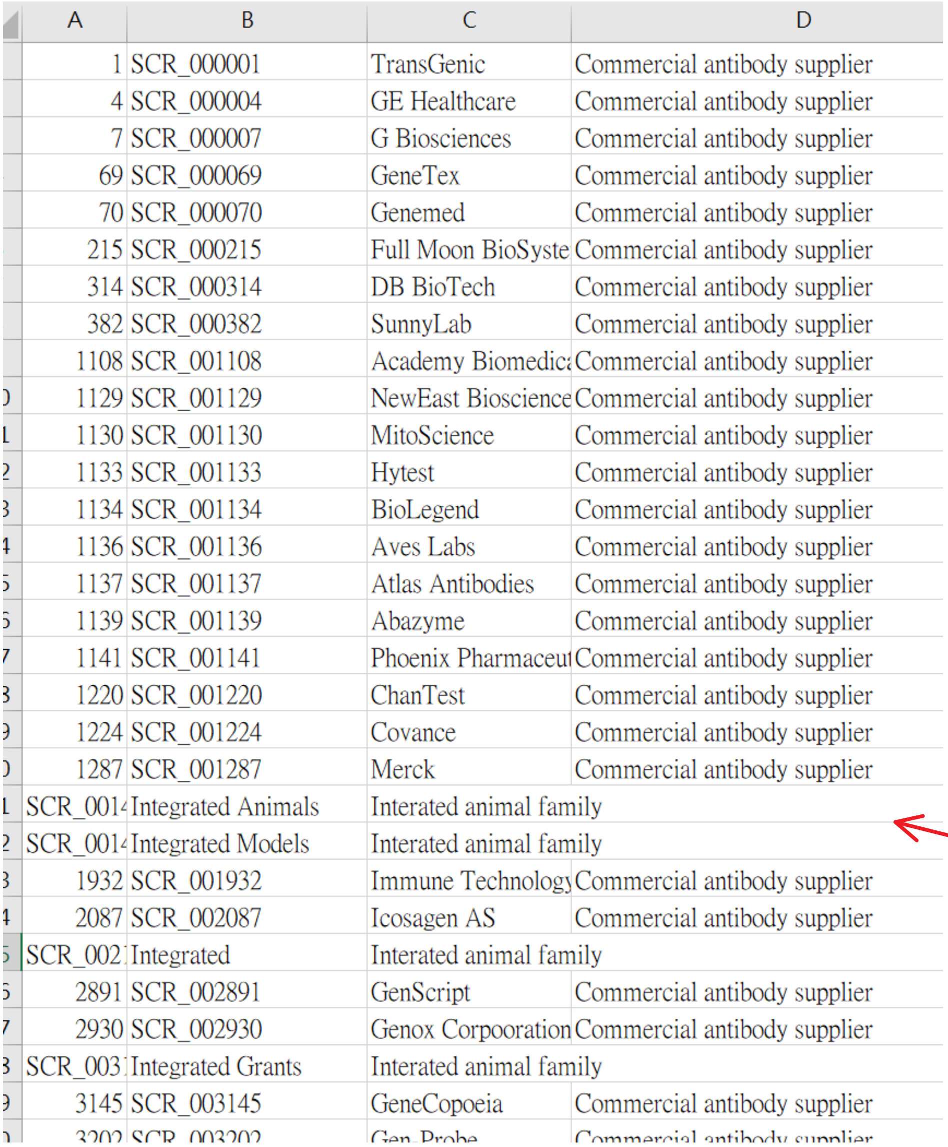
需要注意事項:

- 小心欄位會有**missing value**!! (如 1 TransGenic的abbreviation missing)

exclusion.tsv

2017年11月30日　　上午 08:55

不需產生資料的 resource ids



A	B	C	D
1	SCR_000001	TransGenic	Commercial antibody supplier
4	SCR_000004	GE Healthcare	Commercial antibody supplier
7	SCR_000007	G Biosciences	Commercial antibody supplier
69	SCR_000069	GeneTex	Commercial antibody supplier
70	SCR_000070	Genemed	Commercial antibody supplier
215	SCR_000215	Full Moon BioSyste	Commercial antibody supplier
314	SCR_000314	DB BioTech	Commercial antibody supplier
382	SCR_000382	SunnyLab	Commercial antibody supplier
1108	SCR_001108	Academy Biomedic	Commercial antibody supplier
1129	SCR_001129	NewEast Bioscience	Commercial antibody supplier
1130	SCR_001130	MitoScience	Commercial antibody supplier
1133	SCR_001133	Hytest	Commercial antibody supplier
1134	SCR_001134	BioLegend	Commercial antibody supplier
1136	SCR_001136	Aves Labs	Commercial antibody supplier
1137	SCR_001137	Atlas Antibodies	Commercial antibody supplier
1139	SCR_001139	Abazyme	Commercial antibody supplier
1141	SCR_001141	Phoenix Pharmaceut	Commercial antibody supplier
1220	SCR_001220	ChanTest	Commercial antibody supplier
1224	SCR_001224	Covance	Commercial antibody supplier
1287	SCR_001287	Merck	Commercial antibody supplier
SCR_0014	Integrated Animals	Interated animal family	
SCR_0014	Integrated Models	Interated animal family	
1932	SCR_001932	Immune Technology	Commercial antibody supplier
2087	SCR_002087	Icosagen AS	Commercial antibody supplier
SCR_002	Integrated	Interated animal family	
2891	SCR_002891	GenScript	Commercial antibody supplier
2930	SCR_002930	Genox Corpooration	Commercial antibody supplier
SCR_003	Integrated Grants	Interated animal family	
3145	SCR_003145	GeneCopoeia	Commercial antibody supplier
3202	SCR_003202	Gen-Probe	Commercial antibody supplier

數字id missing!!

不需為這些resource產生table & graph,
也不能讓這些resource出現在其他resource的graph & table內
保險起見, 在db內把他們mention co-mention的紀錄移除,

因為UCSD會不斷update這個檔案(修正IE程式擷取資料的錯誤),
請先了解秋中設計的db格式,
並設計Python程式來處理tsv檔以產生相關資料.

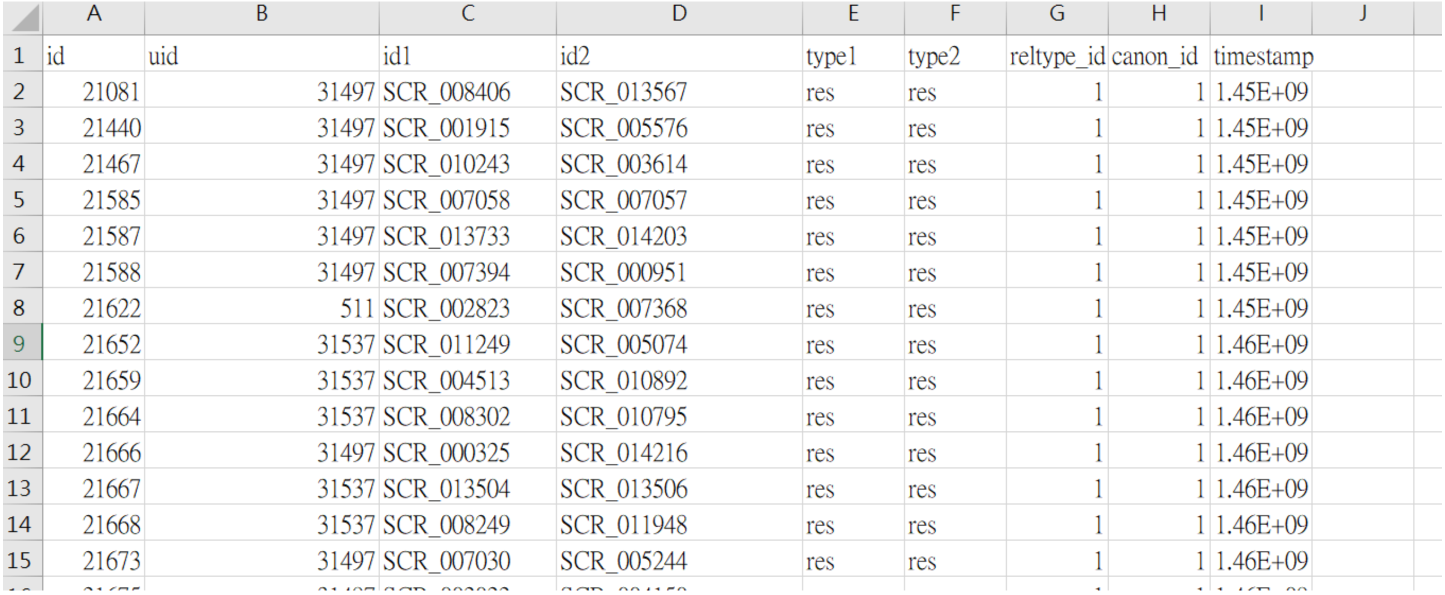
需要注意事項:

- 小心欄位會有**missing value**!!

resource-duplicates.tsv

2017年11月30日 上午 08:55

紀錄著本是同一resource, 卻因為一些緣故產生多組(餘)的resource id,
需依照該檔來將mention & co-mention整合(修正)
會影響資料一致性...需小心處理!!!



	A	B	C	D	E	F	G	H	I	J
1	id	uid	id1	id2	type1	type2	reltype_id	canon_id	timestamp	
2	21081	31497	SCR_008406	SCR_013567	res	res	1	1	1.45E+09	
3	21440	31497	SCR_001915	SCR_005576	res	res	1	1	1.45E+09	
4	21467	31497	SCR_010243	SCR_003614	res	res	1	1	1.45E+09	
5	21585	31497	SCR_007058	SCR_007057	res	res	1	1	1.45E+09	
6	21587	31497	SCR_013733	SCR_014203	res	res	1	1	1.45E+09	
7	21588	31497	SCR_007394	SCR_000951	res	res	1	1	1.45E+09	
8	21622	511	SCR_002823	SCR_007368	res	res	1	1	1.45E+09	
9	21652	31537	SCR_011249	SCR_005074	res	res	1	1	1.46E+09	
10	21659	31537	SCR_004513	SCR_010892	res	res	1	1	1.46E+09	
11	21664	31537	SCR_008302	SCR_010795	res	res	1	1	1.46E+09	
12	21666	31497	SCR_000325	SCR_014216	res	res	1	1	1.46E+09	
13	21667	31537	SCR_013504	SCR_013506	res	res	1	1	1.46E+09	
14	21668	31537	SCR_008249	SCR_011948	res	res	1	1	1.46E+09	
15	21673	31497	SCR_007030	SCR_005244	res	res	1	1	1.46E+09	

主要欄位:

- Id1, id2: 多餘的resource id

整合(修正)方式:

duplicate會造成一個resource的mentions與co-mentions的紀錄四散,
如一個resource如果有3個ids,
則應該把這三個ids的mentions, co-mentions整合

首先要確定這些duplicate ids有無串連,
如 SCR_1 跟 SCR_5 duplicate,
SCR_5又跟 SCR_10 duplicate,
要將所有相關(串聯)的ids進行整合,
即 SCR_1, SCR_5 & SCR_10 是代表同一個resource,
假設最後這三個ids都用SCR_1來統稱(代表) - **(先前挑代表id的方式有點問題, 我目前正在等 UCSD答覆如何挑對的代表id…)**

修正mention -
假如SCR_5曾出現在PMID:9000, 但SCR_1沒有, 則要修正成SCR_1有出現在PMID:9000, 且要更正SCR_1的total_mention次數 (加一)
但若SCR_1也有出現在PMID:9000, 則不修正, 且捨棄SCR_5出現在PMID:9000的資料
****若有修正, 記得要更正total_mention次數****

修正co-mention -
若SCR_5 與 SCR_2 曾一起出現在PMID:3,PMID:5, PMID:7, 而SCR_1 跟 SCR_2曾一起出現在PMID:1,PMID:5, PMID:4,
則要刪掉SCR_5與SCR_2的co-mention, 且修正SCR_1與SCR_2的co-mention為PMID:1, PMID:3, PMID:4, PMID:5, PMID:7,
**** 要同時修正SCR_1與SCR_2的co-mention count為 5(原本為3) ****

若SCR_5 與 SCR_4 曾一起出現在PMID:11,PMID:19, PMID:37, 而SCR_1 跟 SCR_4不曾一起出現過,
則要刪掉SCR_5與SCR_4的co-mention, 且修正SCR_1與SCR_4的co-mention為PMID:11, PMID:19, PMID:37
**** 要同時修正SCR_1與SCR_4的co-mention count為 3(原本沒這筆資料) ****

若發現SCR_5與SCR_1有一起出現的PMIDs, 則可以delete這些資料,
**** 要check資料的正確性, 如SCR_5與SCR_1一起出現在PMID:200, PMID:400,
則需要看"修正後的mention"是否有記錄到SCR_1出現在PMID:200, PMID:400,
如果有漏…表示資料不一致!! ****

因為UCSD會不斷update這個檔案(修正IE程式擷取資料的錯誤),
請先了解秋中設計的db格式,
並設計Python程式來處理tsv檔以產生相關資料.

需要注意事項:

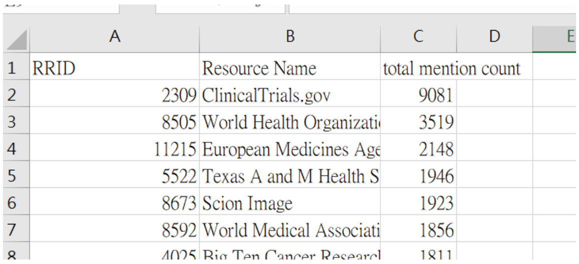
- 目前還在釐清正確的id置換法則

Community Detection

2017年11月30日　　上午 08:41

分析 global resource graph內連結強落而產生的resource clusters
每個cluster包含關係緊密的resource集合

先前秋中已產生30個clusters of resources,
每個cluster存成一個文字檔 (0.txt, 1.txt, ..., 29.txt)



	A	B	C	D	E
1	RRID	Resource Name	total mention count		
2	2309	ClinicalTrials.gov	9081		
3	8505	World Health Organizati	3519		
4	11215	European Medicines Age	2148		
5	5522	Texas A and M Health S	1946		
6	8673	Scion Image	1923		
7	8592	World Medical Associati	1856		
8	4025	Big Ten Cancer Researcl	1811		

主要欄位:

- RRID: resource id
- Resource Name: resource名稱
- total_mention_count: resouce出現的在論文篇數

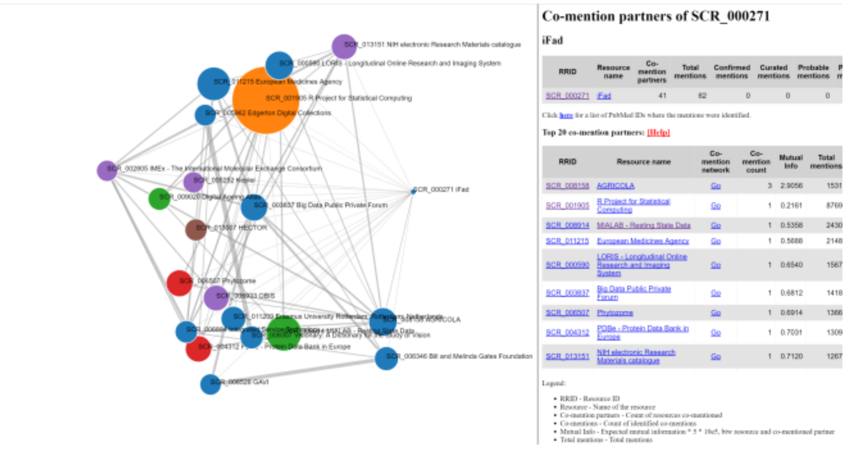
主要產出資訊:

- Resource graph內同cluster的resource會標示同個顏色 (見右圖)

Community detection的方法:
用現行的package SLM
http://www.ludowaltman.nl/slm/

需要注意事項:

- 我們先把資料(tsv & db)整理好, 確定無誤後再來研究SLM的使用方式,


SCR_013151 NIH electronic Research Materials catalogue
SCR_000590 LORIS - Longitudinal Online Research and Imaging System
SCR_011215 European Medicines Agency
SCR_001905 R Project for Statistical Computing
SCR_005962 Edgerton Digital Collections
SCR_002805 IMEx - The International Molecular Exchange Consortium
SCR_008282 Kepler
SCR_009020 Digital Ageing Atlas
SCR_003837 Big Data Public Private Forum
SCR_000271 iFad
SCR_013907 HECTOR
SCR_006587 Phytozome
SCR_008933 OBIS
SCR_011269 Erasmus University Rotterdam; Rotterdam; Netherlands
SCR_008158 AGRICOLA
SCR_006664 Integrated Service Technology
MIALAB - Resting State Data
SCR_006307 Visionary, A Dictionary for the Study of Vision
SCR_004312 PDBe - Protein Data Bank in Europe
SCR_006346 Bill and Melinda Gates Foundation
SCR_008528 GAVI

Co-mention partners of SCR_000271

iFad

RRID	Resource name	Co-mention partners	Total mentions	Confirmed mentions	Curated mentions	Probable mentions
SCR_000271	iFad	41	62	0	0	0

Click here for a list of PubMed IDs where the mentions were identified.

Top 20 co-mention partners: [Help]

RRID	Resource name	Co-mention network	Co-mention count	Mutual Info	Total mentions
SCR_008158	AGRICOLA	Go	3	2.9056	1531
SCR_001905	R Project for Statistical Computing	Go	1	0.2161	8769
SCR_008914	MIALAB - Resting State Data	Go	1	0.5358	2430
SCR_011215	European Medicines Agency	Go	1	0.5688	2148
SCR_000590	LORIS - Longitudinal Online Research and Imaging System	Go	1	0.6540	1567
SCR_003837	Big Data Public Private Forum	Go	1	0.6812	1418
SCR_006507	Phytozome	Go	1	0.6914	1366
SCR_004312	PDBe - Protein Data Bank in Europe	Go	1	0.7031	1309
SCR_013151	NIH electronic Research Materials catalogue	Go	1	0.7120	1267

Legend:

- RRID - Resource ID
- Resource - Name of the resource
- Co-mention partners - Count of resources co-mentioned
- Co-mentions - Count of identified co-mentions
- Mutual Info - Expected mutual information * 5 * 10e5, btw resource and co-mentioned partner
- Total mentions - Total mentions