ORNL

SCIENTIFIC PUBLICATION MINING

SMC Data Challenge 2017

Supriya Chinthavali, Ethan Hicks, Philip Hicks,Dakotah Maguire

Introduction: The amount of scientific information available on-line has increased at an exceptional rate, with recent estimates reporting that a new paper is getting published every 20 seconds. This report provides a brief description of our solutions to the four sub-challenges below which apply machine learning and graph modelling techniques called **probabilistic topic models** and **social network analysis** to mine the scientific publications and identify key characteristics and patterns that can be used by human researchers to develop useful knowledge and further enhance scientific discovery.

1. Identify the individual or group of individuals who appear to be the expert in a particular field or subfield: We took a social network analysis(SNA) based approach to identify experts using scientific publications data,where networks are derived from interactions between individuals/papers.For this task, we try to investigate how a research area changes over time, by building interlinked citation, co-citation, and co-authorship networks that evolve and expand constantly through the emergence of new papers and authors. More specifically, we analyze the dataset in its reduced form using techniques from social networks (key author/paper analysis), spatial analysis (relationship among involved countries) and text mining.The constructed citation networks can also help researchers identify topics that are related to a specific research topic and the subfields/communities structured around these topics.

We constructed 3 citation networks(descriptions below) using custom developed python scripts (direct_citations.py, cocitations.py, coauthors.py) using the authors and the citations data.All the data was downloaded and imported into the postgresql database.The outputs from the python scripts consisted of node and edge files with node attributes such as number of citations per articles or number of publications and edge attributes such as number of co-citations/number of coauthored publications. These files were imported into the Cytoscape tool which is an open source software platform for complex network analysis and visualization. The networks were further filtered using the node attributes(times_cited and number of publications per author) to identify the key players/authors. We also executed community detection algorithms[3](Glay) on the co-citation and co-authorship networks to identify key groups/topics and experts in those cluster groups through page rank/centrality metrics.

1. **Direct citation network**: where a directed edge (i,j) represents a citation from paper i to j. It is a directed network where the links go from one document to the other. In order to limit the size of the graph using 84M citations data, we included only articles nodes that had been cited at least 100 times.The filtered direct citation graph network had 65691 nodes and 10548 edges(Figure 1). The network was further filtered based on times_cited attribute.The size of each circle is based on the total number of citations to the article. The centrality measures degree, closeness,and betweenness centrality were applied to the largest component subnetwork, and the articles were sorted and ranked based on each of the measures(fig 4 and 5).

Fig 1: *Direct citation network after filtering 84M citation edges to nodes having at least 100 citations*

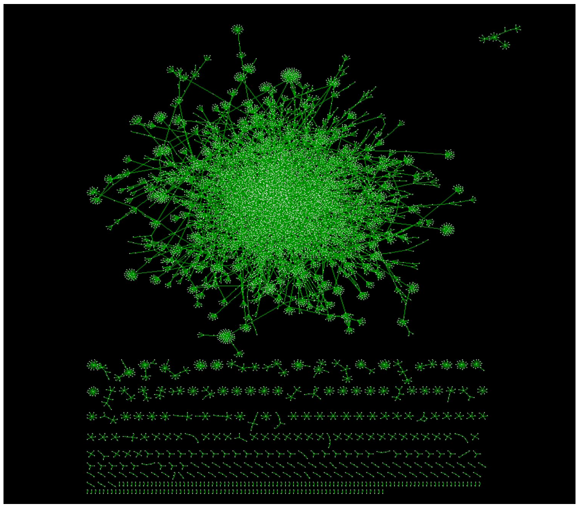


Fig 2: *Zoomed in view of largest connected component*

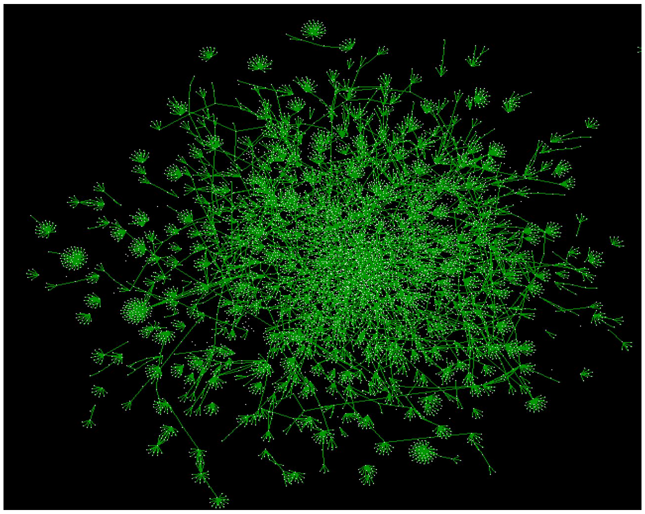
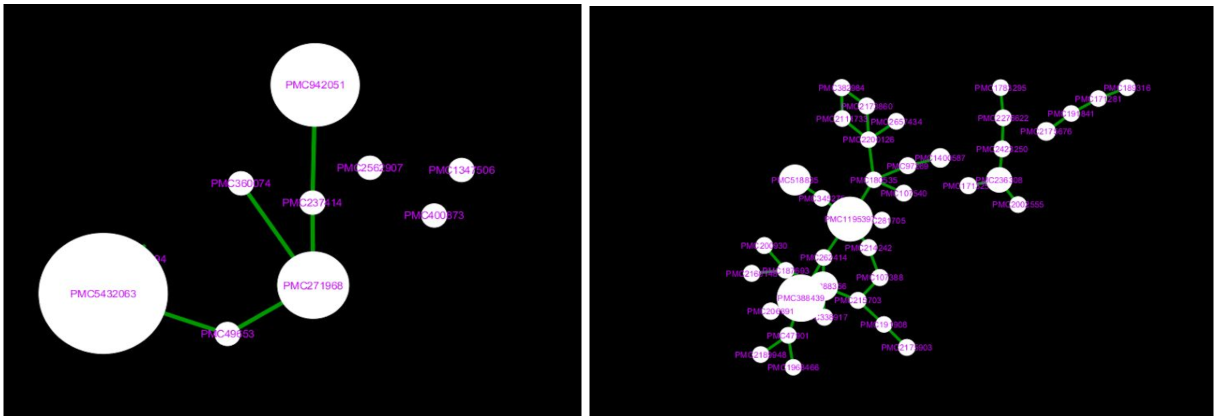
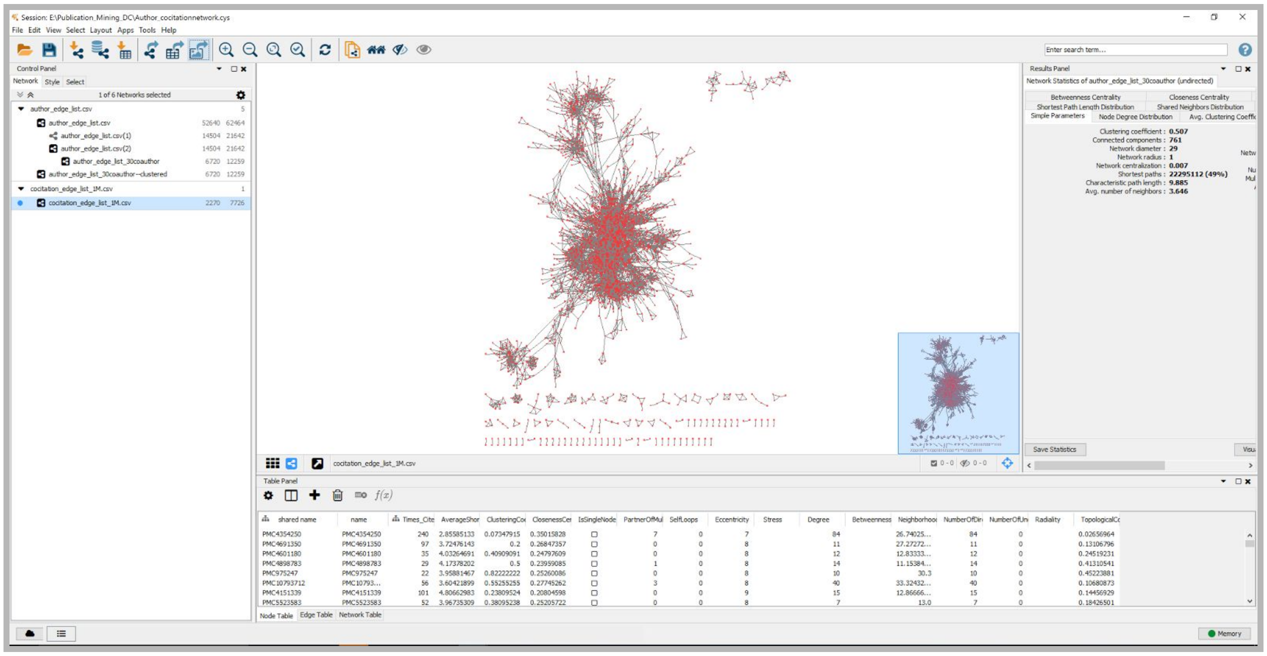


Fig 3,4: *Top nodes obtained after filtering by degree and betweenness centrality on direct citation network*



2.**Co-citation network:** Nodes are articles and edges are added when 2 articles are cited together indicating they are similar works. Links in co-authorship networks are reciprocal (symmetric).The link weights between two authors in co-authorship networks can increase as we add more articles over time.

Fig 5: *Co-citation network constructed from 1M citations dataset and filtered by nodes with at least 100 citations.*



Session: E:\Publication_Mining_DC\Author_cocitationnetwork.cys

1 of 6 Networks selected

author_edge_list.csv		5
author_edge_list.csv	52640	62464
author_edge_list.csv(1)	14504	21642
author_edge_list.csv(2)	14504	21642
author_edge_list_30coauthor	6720	12259
author_edge_list_30coauthor--clustered	6720	12259
cocitation_edge_list_1M.csv		1
cocitation_edge_list_1M.csv	2270	7726

Network Statistics of author_edge_list_30coauthor (undirected)

Betweenness Centrality | Closeness Centrality | Shortest Path Length Distribution | Shared Neighbors Distribution | Simple Parameters | Node Degree Distribution | Avg. Clustering Coeffic

```
  Clustering coefficient : 0.507
    Connected components : 761
        Network diameter : 29
          Network radius : 1
  Network centralization : 0.007
          Shortest paths : 22295112 (49%)
Characteristic path length : 9.885
Avg. number of neighbors : 3.646
```

cocitation_edge_list_1M.csv

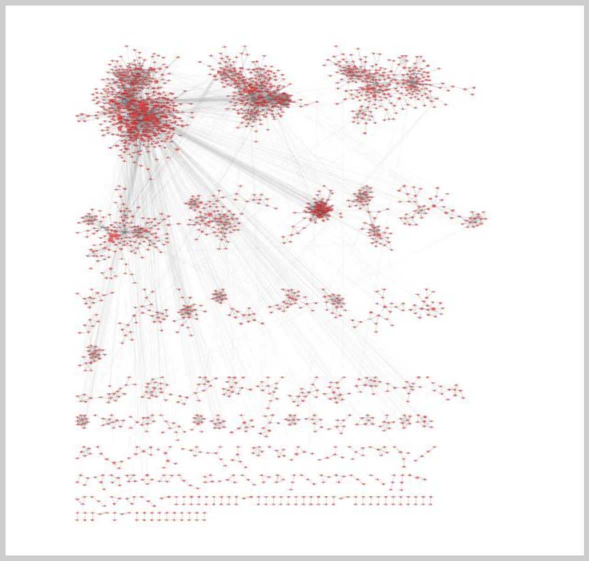
shared name	name	Times_Cite	AverageShor	ClusteringCo	ClosenessCer	IsSingleNode	PartnerOfMul	SelfLoops	Eccentricity	Stress	Degree	Betweenness	Neighborhoo	NumberOfDir	NumberOfUn	Radiality	TopologicalC
PMC4354250	PMC4354250	240	2.85585133	0.07347915	0.35015828	☐	7	0	7		84		26.74025...	84	0		0.02656964
PMC4691350	PMC4691350	97	3.72476143	0.2	0.26847357	☐	0	0	8		11		27.27272...	11	0		0.13106796
PMC4601180	PMC4601180	35	4.03264691	0.40909091	0.24797609	☐	0	0	8		12		12.83333...	12	0		0.24519231
PMC4898783	PMC4898783	29	4.17378202	0.5	0.23959085	☐	1	0	8		14		11.15384...	14	0		0.41310541
PMC975247	PMC975247	22	3.95881467	0.82222222	0.25260086	☐	0	0	8		10		30.3	10	0		0.45223881
PMC10793712	PMC10793...	56	3.60421899	0.55255255	0.27745262	☐	3	0	8		40		33.32432...	40	0		0.10680873
PMC4151339	PMC4151339	101	4.80662983	0.23809524	0.20804598	☐	0	0	9		15		12.86666...	15	0		0.14456929
PMC5523583	PMC5523583	52	3.96735309	0.38095238	0.25205722	☐	0	0	8		7		13.0	7	0		0.18426501

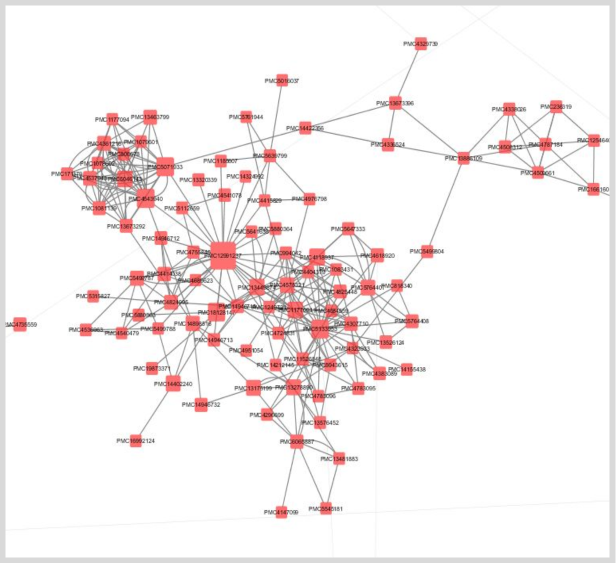
Node Table | Edge Table | Network Table

Fig 6:*Clusters identified after executing community detection algorithms on a co-citation network of 1M citations. Each cluster indicates a specific articles*
research topic articles.

Fig 7*: Zoomed in view of a cluster. Size of the node represents the number of times the articles was cited. Edge thickness represents the number of times the*

were co-cited.





3.Co-authorship network :is built based on which author writes papers with whom.Authors are nodes while edges represent co-authorship on a paper. The network in Fig 6 and 7 was constructed after filtering the network to author nodes that have atleast 30 publications.

Fig 8:*Clusters identified after executing community detection algorithms on a co-author network of 1M authors.*

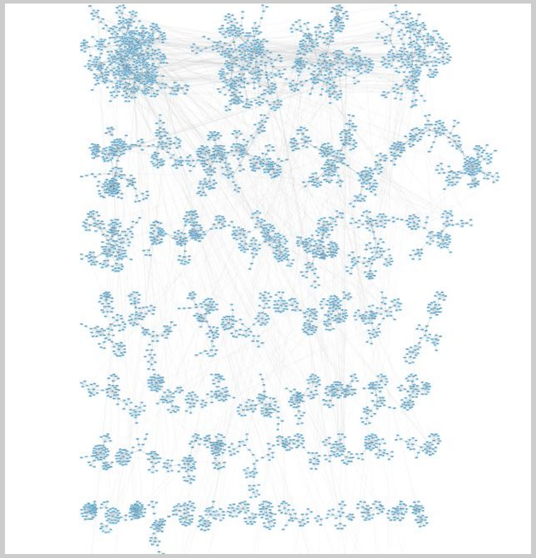
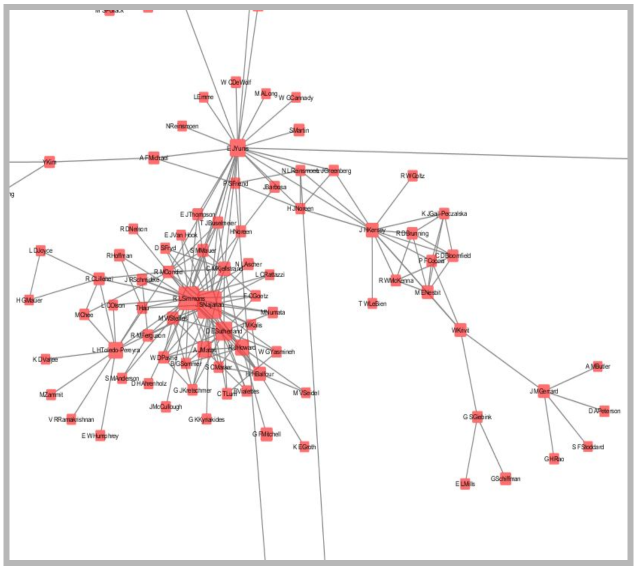


Fig 9*: Zoomed in view of a cluster. Size of the node represents the number of times the articles was cited. Edge thickness represents cocited article strength.*



2. Identify topics that have been researched across all publications. In this task, instead of exploring graph methods, we applied probabilistic topic models which are algorithms used for discovering the main themes that extend across a large and unstructured collection of documents[1]. Topic modeling is an emerging field in machine learning, which allows us to analyze streaming collections,see how the discovered themes are connected to each other, and how they change over time. The latent Dirichlet allocation (LDA) is the simplest statistical topic model.The intuition behind LDA is that documents exhibit multiple topics. A topic is defined to be a distribution over a fixed vocabulary. Each document exhibits the topics in different proportion; each word in each document is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics. The documents themselves are observed, while the topic structure—the topics, per-document topic distributions, and the per-document per-word topic assignments is hidden structure.One of the main advantages of formulating LDA as a probabilistic model is that it can easily be used as a module in more complicated models for more complicated goals. In such collections, we may want to assume that the topics change over time. One approach to this problem is the dynamic topic model—a model that considers the ordering of the documents and provides a better posterior topical structure than LDA[2]. Rather than a single distribution over words, a topic is now represented as a sequence of distributions over words. We can find an underlying theme of the collection and track how it has changed over time. Below is the workflow that we implemented to identify topics over time using the abstracts data and visualized using highlight tables and interactive dashboards(Fig 10,Fig 12).

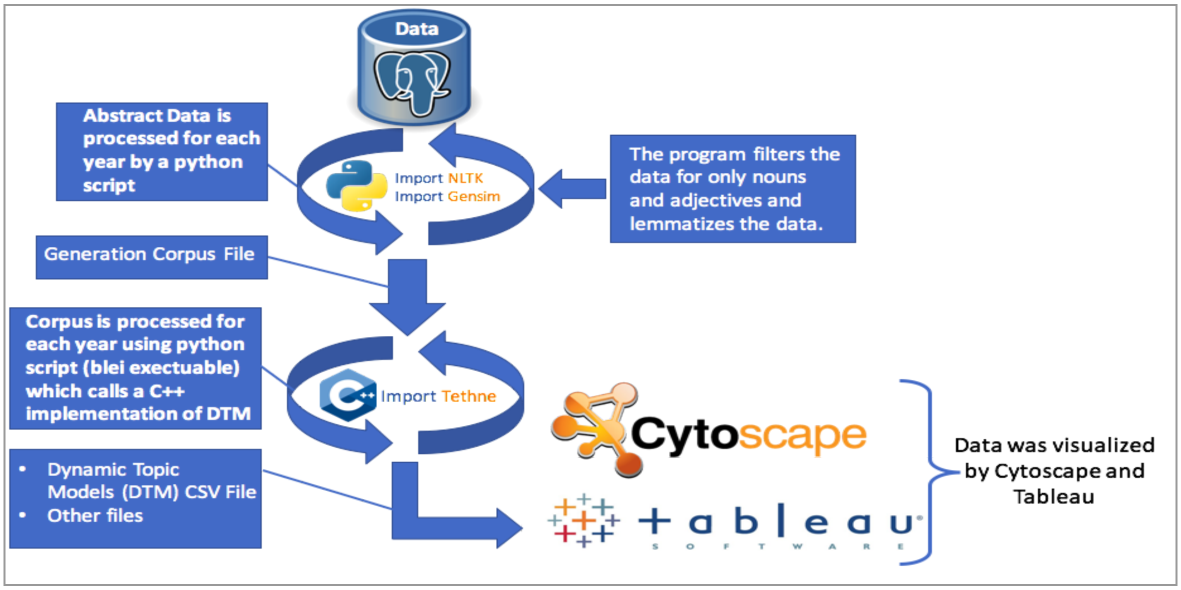
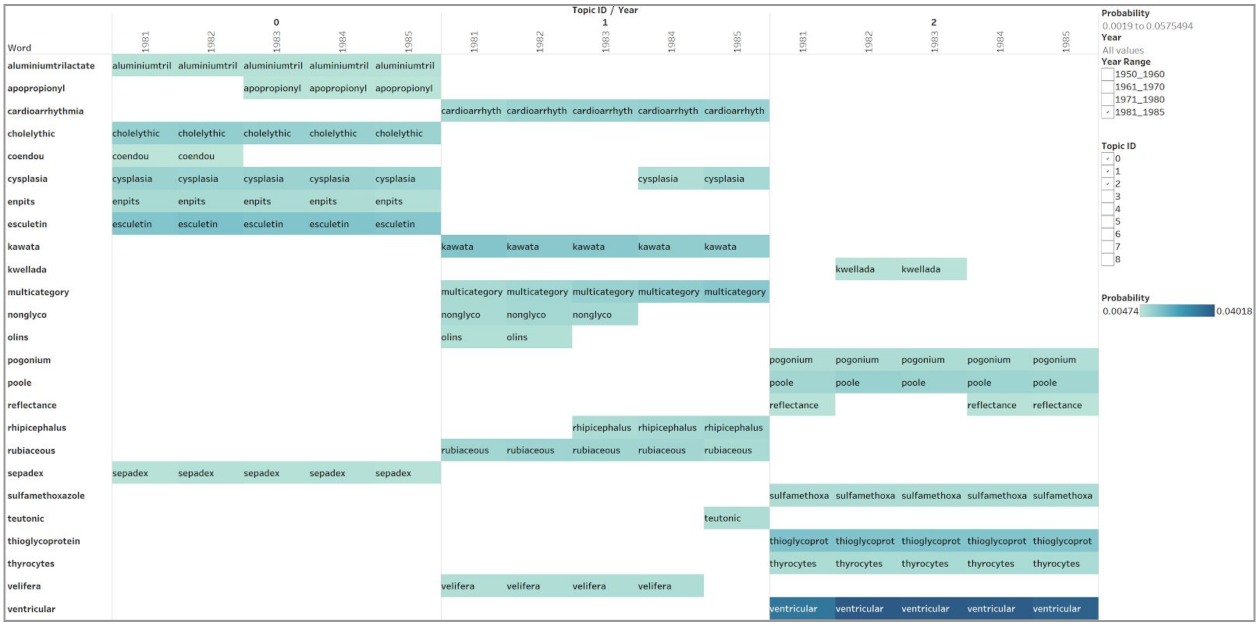

Data
Abstract Data is processed for each year by a python script
Import NLTK
Import Gensim
The program filters the data for only nouns and adjectives and lemmatizes the data.
Generation Corpus File
Corpus is processed for each year using python script (blei exectuable) which calls a C++ implementation of DTM
Import Tethne
Dynamic Topic Models (DTM) CSV File
Other files
Cytoscape
tableau SOFTWARE
Data was visualized by Cytoscape and Tableau

Fig 10: Dynamic Topic Modelling Output Dashboard: Terms used in topics 0,1and 2 in the year range 1981-1985



	Topic ID / Year														
	0					1					2				
Word	1981	1982	1983	1984	1985	1981	1982	1983	1984	1985	1981	1982	1983	1984	1985
aluminiumtrilactate	aluminiumtril	aluminiumtril	aluminiumtril	aluminiumtril	aluminiumtril										
apopropionyl			apopropionyl	apopropionyl	apopropionyl										
cardioarrhythmia						cardioarrhyth	cardioarrhyth	cardioarrhyth	cardioarrhyth	cardioarrhyth					
cholelythic	cholelythic	cholelythic	cholelythic	cholelythic	cholelythic										
coendou	coendou	coendou													
cysplasia	cysplasia	cysplasia	cysplasia	cysplasia	cysplasia				cysplasia	cysplasia					
enpits	enpits	enpits	enpits	enpits	enpits										
esculetin	esculetin	esculetin	esculetin	esculetin	esculetin										
kawata						kawata	kawata	kawata	kawata	kawata					
kwellada												kwellada	kwellada		
multicategory						multicategory	multicategory	multicategory	multicategory	multicategory					
nonglyco						nonglyco	nonglyco	nonglyco							
olins						olins	olins								
pogonium											pogonium	pogonium	pogonium	pogonium	pogonium
poole											poole	poole	poole	poole	poole
reflectance											reflectance			reflectance	reflectance
rhipicephalus								rhipicephalus	rhipicephalus	rhipicephalus					
rubiaceous						rubiaceous	rubiaceous	rubiaceous	rubiaceous	rubiaceous					
sepadex	sepadex	sepadex	sepadex	sepadex	sepadex										
sulfamethoxazole											sulfamethoxa	sulfamethoxa	sulfamethoxa	sulfamethoxa	sulfamethoxa
teutonic										teutonic					
thioglycoprotein											thioglycoprot	thioglycoprot	thioglycoprot	thioglycoprot	thioglycoprot
thyrocytes											thyrocytes	thyrocytes	thyrocytes	thyrocytes	thyrocytes
velifera						velifera	velifera	velifera	velifera						
ventricular											ventricular	ventricular	ventricular	ventricular	ventricular

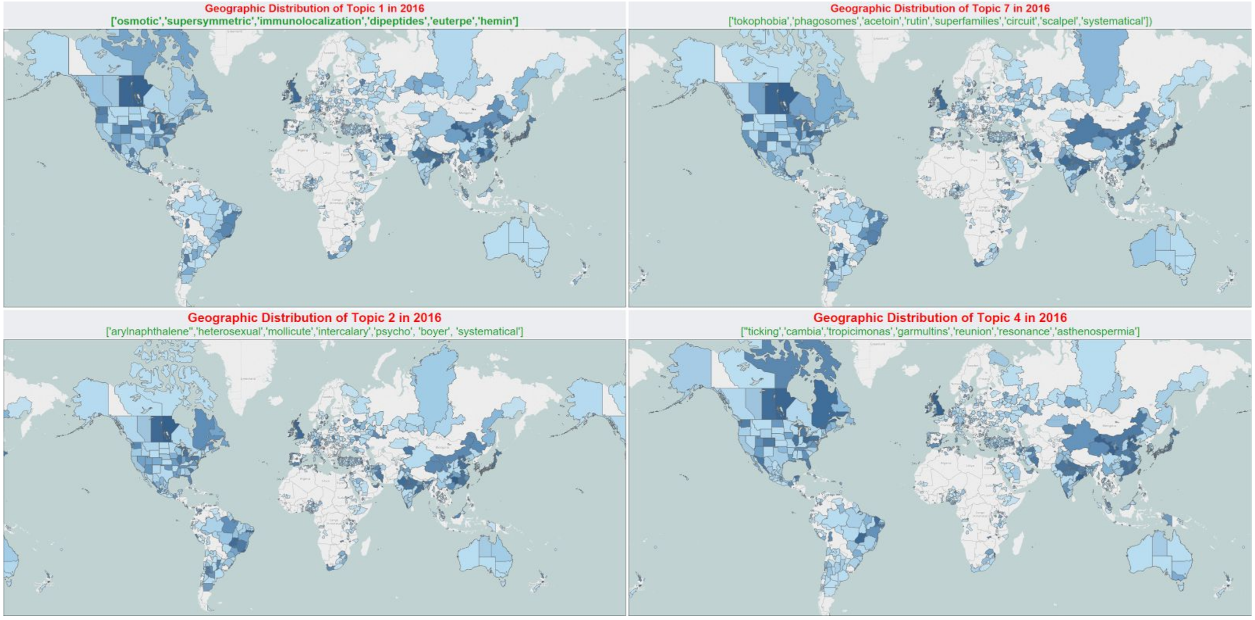
Probability
0.0019 to 0.0575494
Year
All values
Year Range
□ 1950_1960
□ 1961_1970
□ 1971_1980
☑ 1981_1985

Topic ID
☑ 0
☑ 1
☑ 2
□ 3
□ 4
□ 5
□ 6
□ 7
□ 8

Probability
0.00474 0.04018

3. Visualize the geographic distribution of the topics in the publications: For this task,we implemented custom python scripts to generate output files that combines the affiliations data with the articles data file. Geographic distribution of about 1M articles for the year 2016 was visualized on a world map using country and state data. Researchers are associated with different institutions across the globe. We also visualized the identified topic distributions of articles along with their geographical distribution using Tableau software. We chose to explore 8 topics each with 8 words associated/correlated with them for the 2016. For e.g., in figure 11, across all 4 of the 8 topics identified, there were countries that consistently produced higher probabilities of researching these topics. Within those countries, there were regions that focused on specific topics. For example, Oregon had a higher probability of researching **topic one** than the other three topics. Due to 7.5 percent of the publications for 2016 not being able to be geocoded there are possibly more regions of higher probability than shown in the maps. This will need to be corrected in future work.

Fig 11: Geographic Distribution Dashboard


Geographic Distribution of Topic 1 in 2016
['osmotic','supersymmetric','immunolocalization','dipeptides','euterpe','hemin']
Geographic Distribution of Topic 7 in 2016
['tokophobia','phagosomes','acetoin','rutin','superfamilies','circuit','scalpel','systematical'])
Geographic Distribution of Topic 2 in 2016
['arylnaphthalene'','heterosexual','mollicute','intercalary','psycho', 'boyer', 'systematical']
Geographic Distribution of Topic 4 in 2016
["ticking','cambia','tropicimonas','garmultins','reunion','resonance','asthenospermia']

4. Identify how topics have shifted over time: The DTM model also outputs the topic distributions over all the articles. These probabilities for all articles were processed as matrices using **R** and the output file was imported into Tableau for visualizing how the identified topics changed over time as **line graphs**(Fig 13).We also identified top documents that are discussing each of the identified topics(Fig 12) as box plots below in Appendix A**.**

APPENDIX A

Fig 12:Top documents discussing Topic 1 for year range 1971-1980

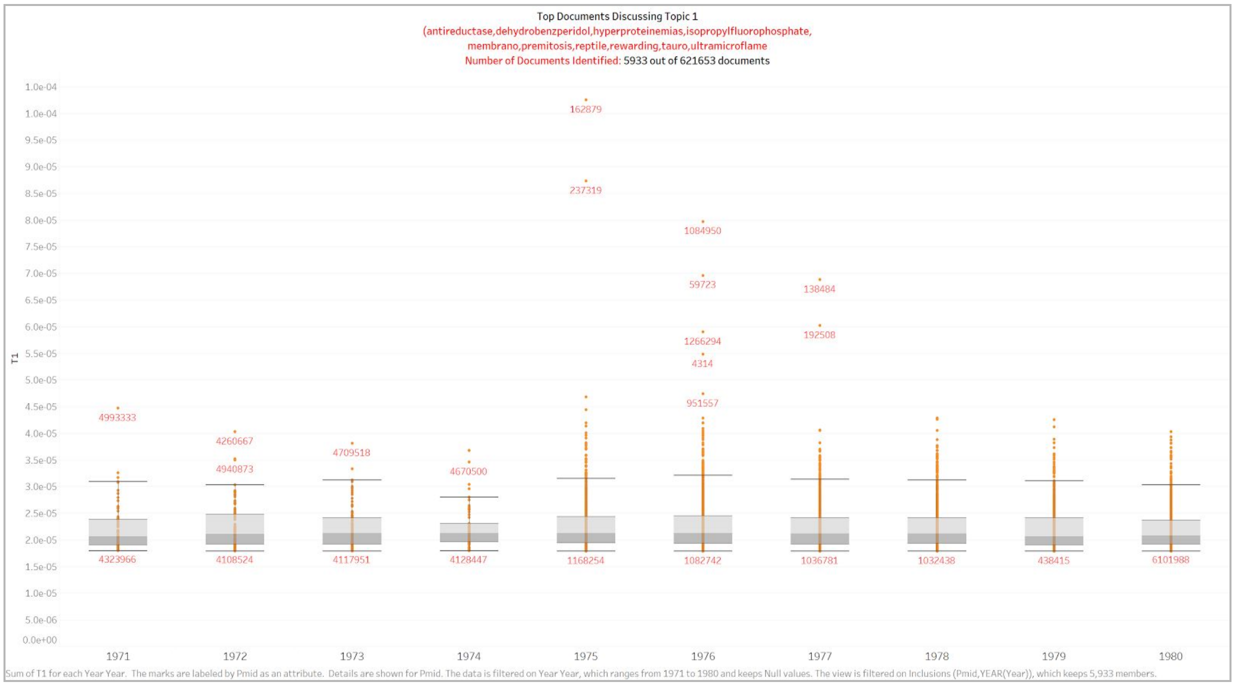


Fig 13:Topics over time for the year range 2001-2010 extracted using DTM on titles of the articles

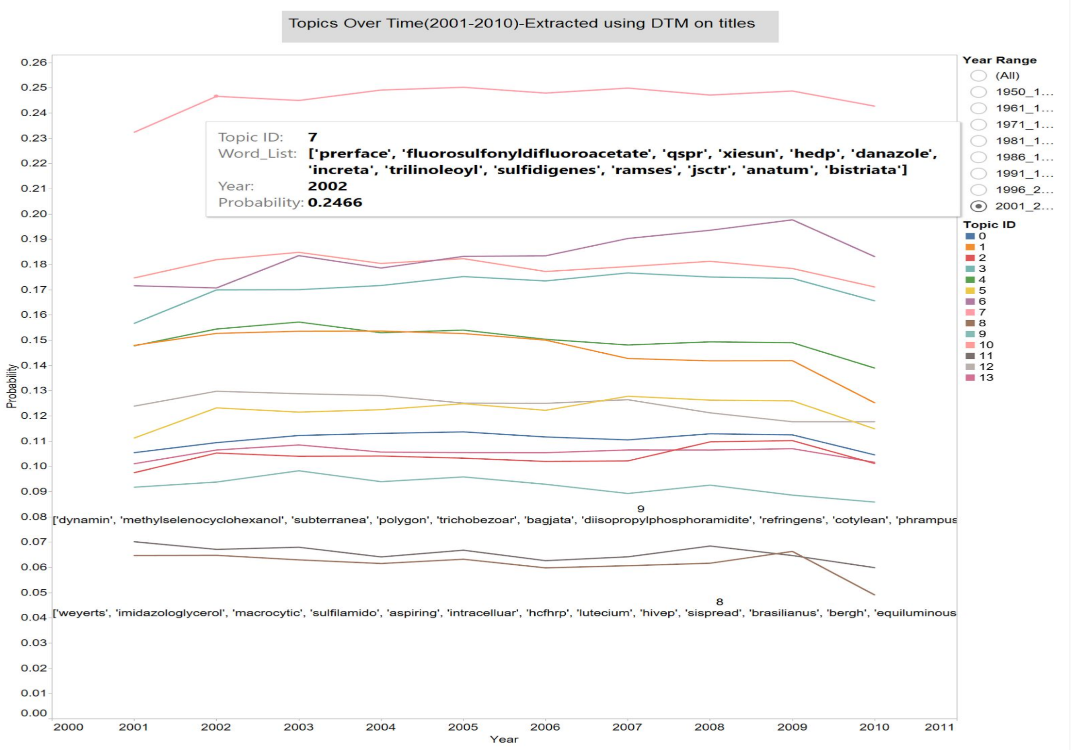
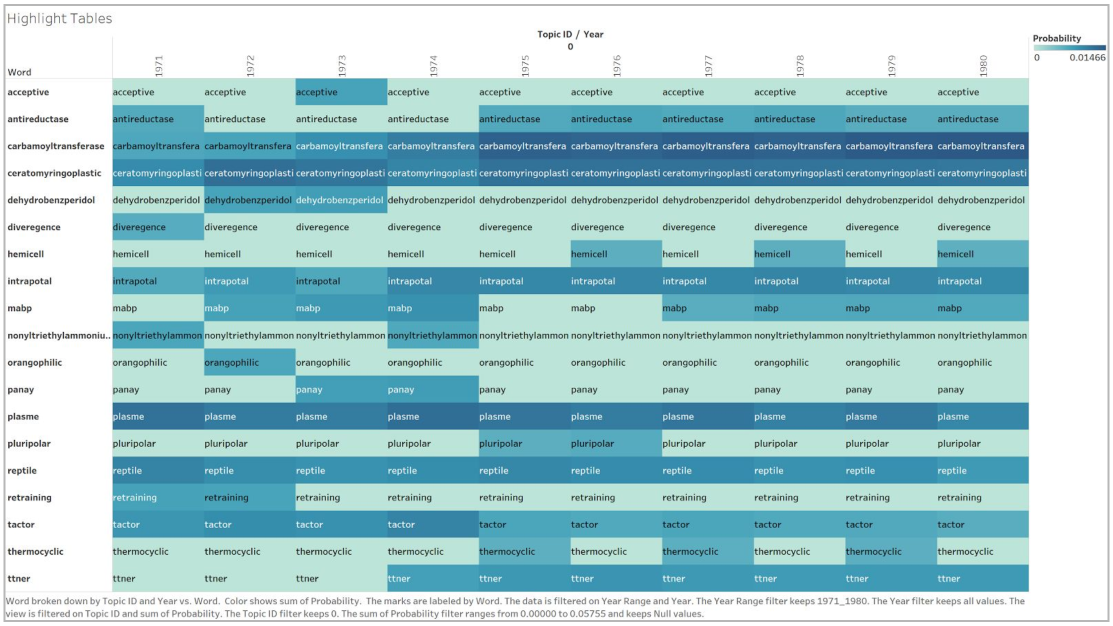


Fig 12: Dynamic Topic Modelling : Terms used in topics 0,in the year range 1971-1980



Highlight Tables

Topic ID / Year
0

Probability
0 0.01466

Word	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
acceptive	acceptive	acceptive	acceptive	acceptive	acceptive	acceptive	acceptive	acceptive	acceptive	acceptive
antireductase	antireductase	antireductase	antireductase	antireductase	antireductase	antireductase	antireductase	antireductase	antireductase	antireductase
carbamoyltransferase	carbamoyltransfera	carbamoyltransfera	carbamoyltransfera	carbamoyltransfera	carbamoyltransfera	carbamoyltransfera	carbamoyltransfera	carbamoyltransfera	carbamoyltransfera	carbamoyltransfera
ceratomyringoplastic	ceratomyringoplasti	ceratomyringoplasti	ceratomyringoplasti	ceratomyringoplasti	ceratomyringoplasti	ceratomyringoplasti	ceratomyringoplasti	ceratomyringoplasti	ceratomyringoplasti	ceratomyringoplasti
dehydrobenzperidol	dehydrobenzperidol	dehydrobenzperidol	dehydrobenzperidol	dehydrobenzperidol	dehydrobenzperidol	dehydrobenzperidol	dehydrobenzperidol	dehydrobenzperidol	dehydrobenzperidol	dehydrobenzperidol
diveregence	diveregence	diveregence	diveregence	diveregence	diveregence	diveregence	diveregence	diveregence	diveregence	diveregence
hemicell	hemicell	hemicell	hemicell	hemicell	hemicell	hemicell	hemicell	hemicell	hemicell	hemicell
intrapotal	intrapotal	intrapotal	intrapotal	intrapotal	intrapotal	intrapotal	intrapotal	intrapotal	intrapotal	intrapotal
mabp	mabp	mabp	mabp	mabp	mabp	mabp	mabp	mabp	mabp	mabp
nonyltriethylammoniu..	nonyltriethylammon	nonyltriethylammon	nonyltriethylammon	nonyltriethylammon	nonyltriethylammon	nonyltriethylammon	nonyltriethylammon	nonyltriethylammon	nonyltriethylammon	nonyltriethylammon
orangophilic	orangophilic	orangophilic	orangophilic	orangophilic	orangophilic	orangophilic	orangophilic	orangophilic	orangophilic	orangophilic
panay	panay	panay	panay	panay	panay	panay	panay	panay	panay	panay
plasme	plasme	plasme	plasme	plasme	plasme	plasme	plasme	plasme	plasme	plasme
pluripolar	pluripolar	pluripolar	pluripolar	pluripolar	pluripolar	pluripolar	pluripolar	pluripolar	pluripolar	pluripolar
reptile	reptile	reptile	reptile	reptile	reptile	reptile	reptile	reptile	reptile	reptile
retraining	retraining	retraining	retraining	retraining	retraining	retraining	retraining	retraining	retraining	retraining
tactor	tactor	tactor	tactor	tactor	tactor	tactor	tactor	tactor	tactor	tactor
thermocyclic	thermocyclic	thermocyclic	thermocyclic	thermocyclic	thermocyclic	thermocyclic	thermocyclic	thermocyclic	thermocyclic	thermocyclic
ttner	ttner	ttner	ttner	ttner	ttner	ttner	ttner	ttner	ttner	ttner

Word broken down by Topic ID and Year vs. Word. Color shows sum of Probability. The marks are labeled by Word. The data is filtered on Year Range and Year. The Year Range filter keeps 1971_1980. The Year filter keeps all values. The view is filtered on Topic ID and sum of Probability. The Topic ID filter keeps 0. The sum of Probability filter ranges from 0.00000 to 0.05755 and keeps Null values.

Table 1: Dynamic Topic Modelling : Terms used in each of the 8 topics in the year range 1971-1980

Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7
acceptive	anti reductase	bbingen	anticyclic	acceptive	dehydrobenzperidol	anticyclique	acylations
anti reductase	aphthous	bifida	bioc	accident	dimethylamino cyclohexyl	arthroscopy	bioc
carbamoyltransferase	budzik	emmens	castellino	acylations	disida	budzik	bradycardia
cerato ring plastic	case	eynde	dehydrobenzperidol	anti reductase	growth	cedrorum	ceroma
dehydrobenzperidol	cynomolgus	effects	di sida	bone metastases	healthy	chaetotaxic	effects
divergence	diazole	fluocortin	ecchronism	ceroma	hinshawii	cinnamon	grady
hemicell	elan	growth	emmonsia	dehydrobenzperidol	lethalities	degastroentero	growth
intraportal	fristoe	hache	eynde	electronystagmogram	multi banded	elan	dihydrophthalazine
mabp	gapdh	hindquarters	flfo	extnahepatic	organophilic	endolymphaticus	hyperfixation
nonyl triethylammonium	lobeless	intraportal	growth	interlaminar	oxoproline	fristoe	interactin
organophilic	mgso	isohemic	hinkle	kinkajous	proofes	grapples	longibrachia
panay	non comprehensi	issei	hyperfixation	nucleo depolymerases	spherite	herford	meshlike

	ve						
plasme	pantazocine	lifs	kreybergs	phlogosis	synostose	himmelsbach	mynah
pluripolar	phagefree	reptile	lethalities	retraining	tsioxide	hinkle	nutans
reptile	projectionists	retroperiteonal	mabp	ropers	unremitting	hydroepoxides	oligouria
retraining	reptile	unremitting	multiword	tananarive	unsplitted	inacurrate	phlogosis
tactor	uninflamed		nanoplanktonic	thby	versions	interspersal	potentiometric titration
thermo cyclic			nasr	thermocyclic		intraspecial	pseudo racemic
ttner			opportunity			phagefree	reptile
			phenyl phenyl butyl			phenylphenylbutyl	steroidal
			plasme,restructuration,smelteries,spherite,tactor,retraining			portsmouth	testiculorum

References

1. Blei, David M. "Probabilistic topic models." *Communications of the ACM* 55.4 (2012): 77-84.
2. Blei, David M., and John D. Lafferty. "Dynamic topic models." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.
3. Gang Su, Allan Kuchinsky, John H. Morris, David J. States, Fan Meng; GLay: community structure analysis of biological networks. *Bioinformatics* 2010; 26 (24): 3135-3137. doi: 10.1093/bioinformatics/btq596
4. http://socialcomputing.ing.puc.cl/uploads/DynamicTopicModellingTutorial.pdf
5. https://www.cognizant.com/whitepapers/identifying-key-opinion-leaders-using-social-network-analysis-codex1234.pdf
6. http://nealcaren.web.unc.edu/a-sociology-citation-network/
7. https://github.com/magsilva/dtm
8. http://web.mit.edu/seyda/www/Papers/ECIR_extended.pdf
9. Newman, Mark EJ. "Coauthorship networks and patterns of scientific collaboration." *Proceedings of the national academy of sciences* 101.suppl 1 (2004): 5200-5205.