Topical Analysis and Visualization of (Network) Data Using Sci2

Ted Polley

Research & Editorial Assistant

Cyberinfrastructure for Network Science Center

Library and Information Science

dapolley@indiana.edu

Please download Sci2 at http://sci2.cns.iu.edu

See documentation at http://wiki.cns.iu.edu/display/SCI2TUTORIAL

Overview

This hands-on session introduces topical analysis and visualization of network data. Specifically, we will use the Sci2 tool to extract co-word occurrence networks and to generate science map overlays.

The topic or semantic coverage of a unit of science can be derived from the text associated with it. Topical aggregations (e.g., over journal volumes, scientific disciplines, or institutions) are common.

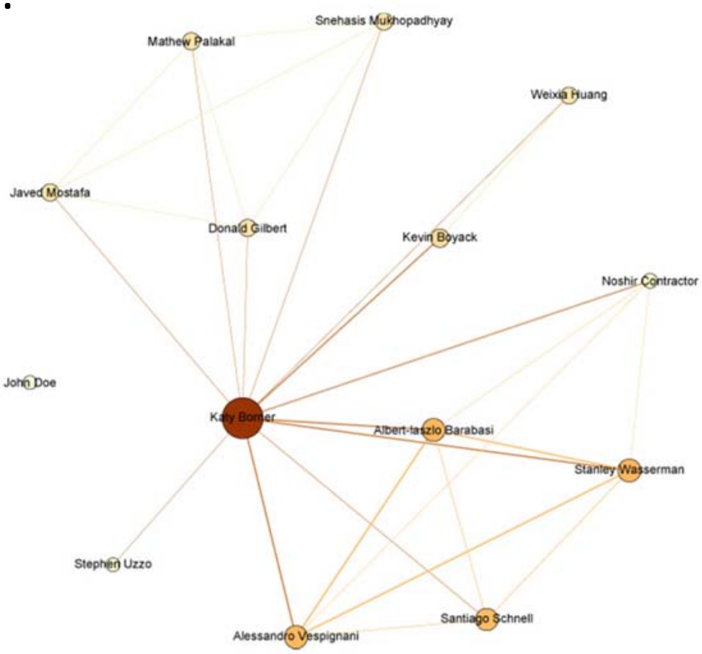
Topical analysis extracts the set of unique words or word profiles and their frequency from a text corpus. Stop words, such as 'the' and 'of' are removed, and stemming can be applied.

Stemming example:

fishing, fished, fish, and fisher to the stem “fish”

What is a Network?

- Graph – network visualized
- Nodes (vertices)
- Edges (links)

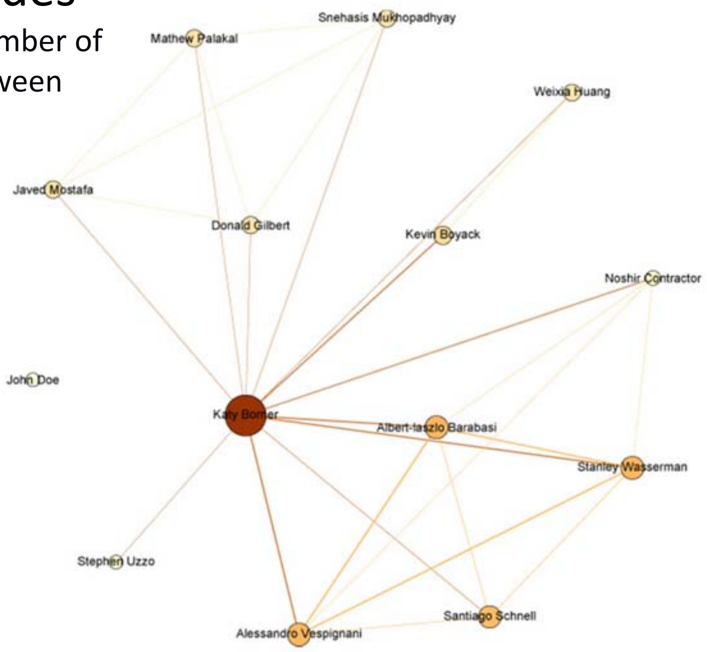


Graph Metrics - Nodes

- Betweenness Centrality – number of shortest paths a node sits between

- Degree

- Isolates



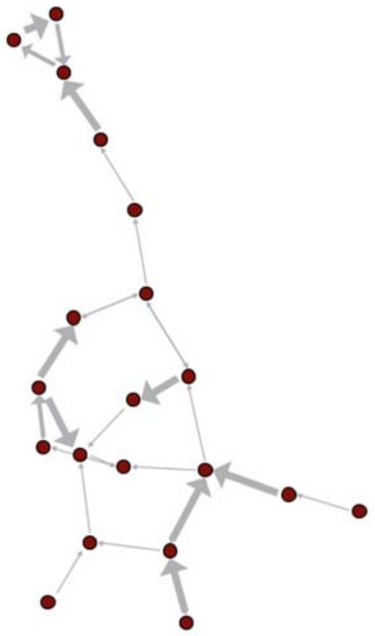
Graph Metrics - Edges

- Shortest paths – shortest distance between two nodes

- Weight – strength of tie

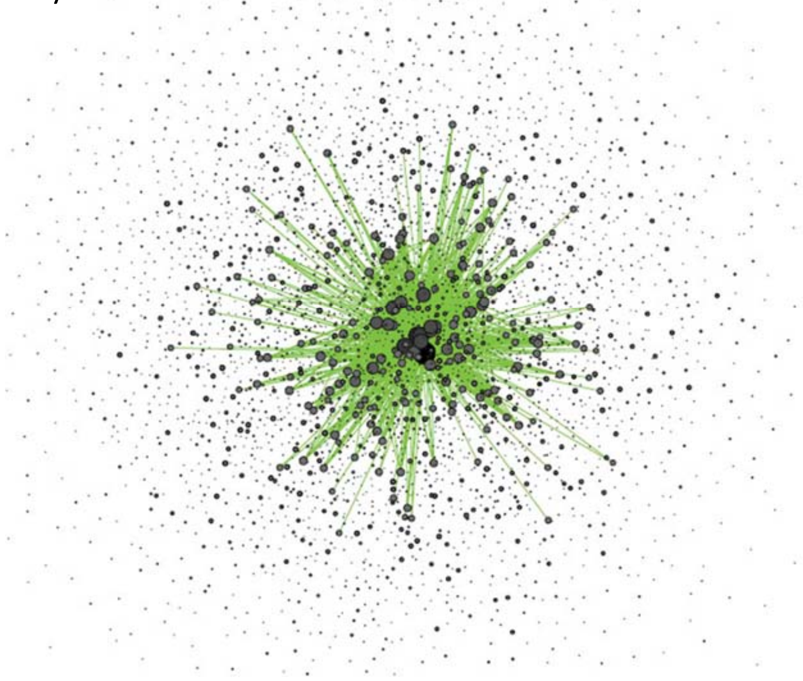
- Directionality – is the connection one-way or two-way (in-degree vs. out-degree)?

- Bridge – deleting would change structure



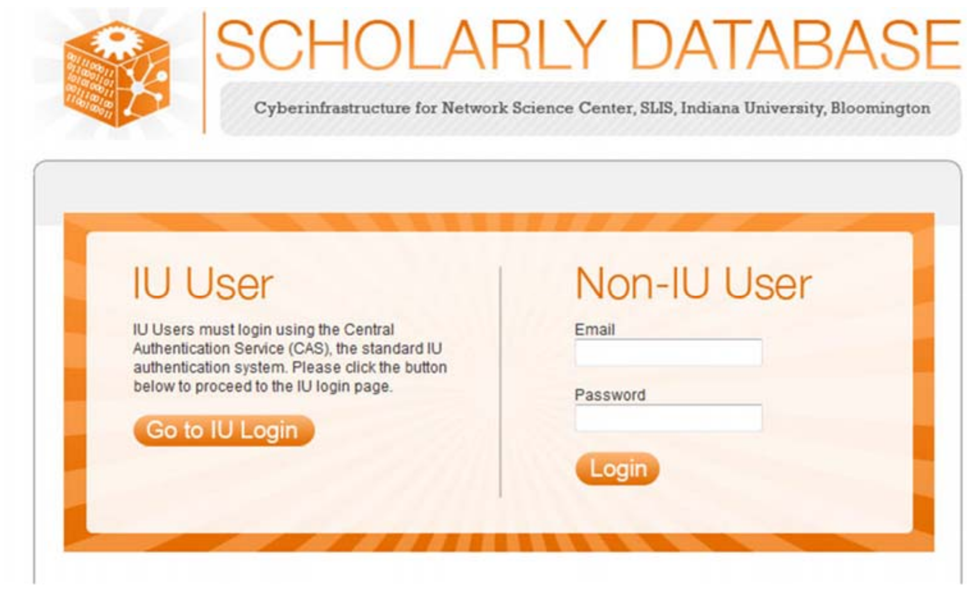
Extracting Word Co-Occurrence Network from SDB Data

Word Co-Occurrence Network from the abstracts of articles from MEDLINE with the keyword “mesothelioma” in the title…



Extracting Word Co-Occurrence Network from SDB Data

If you have registered for the Scholarly Database then go to http://sdb.cns.iu.edu and login…



SCHOLARLY DATABASE

Cyberinfrastructure for Network Science Center, SLIS, Indiana University, Bloomington

IU User

IU Users must login using the Central Authentication Service (CAS), the standard IU authentication system. Please click the button below to proceed to the IU login page.

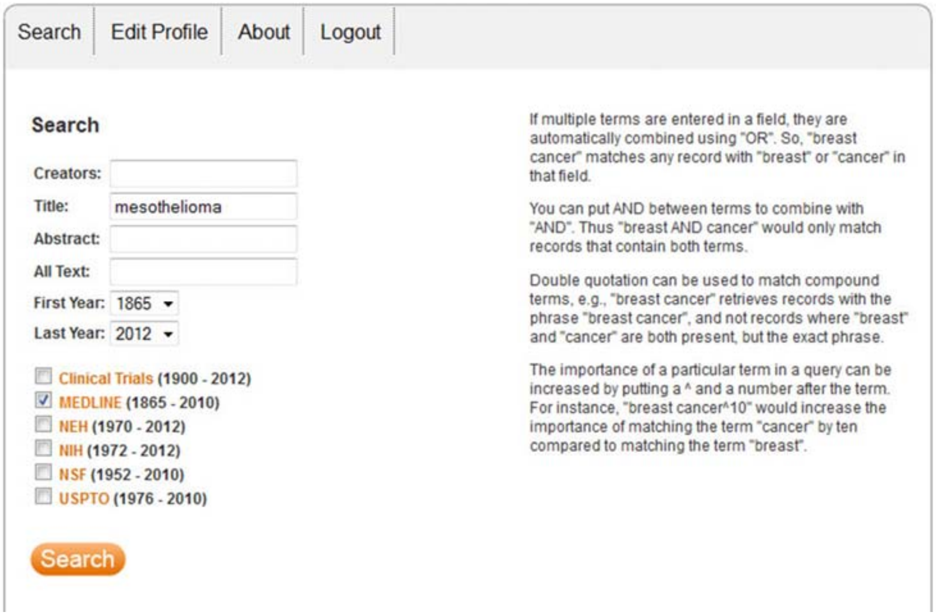
Non-IU User

Email

Password

If you do not have an account type in nwb@indiana.edu and nwb for the password

Do a keyword search in Title for “mesothelioma” and check MEDLINE…



Search

Creators:
Title: mesothelioma
Abstract:
All Text:
First Year: 1865
Last Year: 2012

- □ Clinical Trials (1900 - 2012)
- ☑ MEDLINE (1865 - 2010)
- □ NEH (1970 - 2012)
- □ NIH (1972 - 2012)
- □ NSF (1952 - 2010)
- □ USPTO (1976 - 2010)

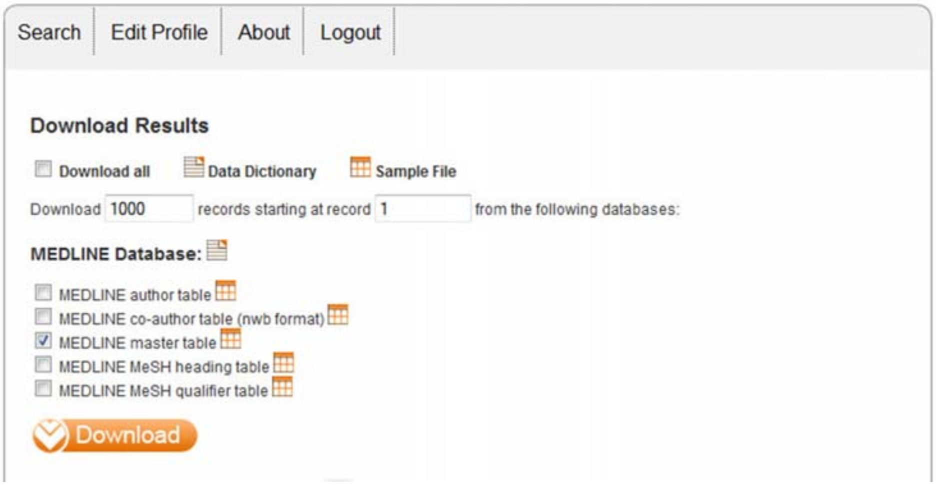
If multiple terms are entered in a field, they are automatically combined using "OR". So, "breast cancer" matches any record with "breast" or "cancer" in that field.

You can put AND between terms to combine with "AND". Thus "breast AND cancer" would only match records that contain both terms.

Double quotation can be used to match compound terms, e.g., "breast cancer" retrieves records with the phrase "breast cancer", and not records where "breast" and "cancer" are both present, but the exact phrase.

The importance of a particular term in a query can be increased by putting a ^ and a number after the term. For instance, "breast cancer^10" would increase the importance of matching the term "cancer" by ten compared to matching the term "breast".

We will only download the first 1000 results to minimize the runtime for the algorithms used in this workflow. Make sure to check MEDLINE master table since that will have all of the bibliographic data we need for this analysis.



Download Results

□ Download all   Data Dictionary   Sample File

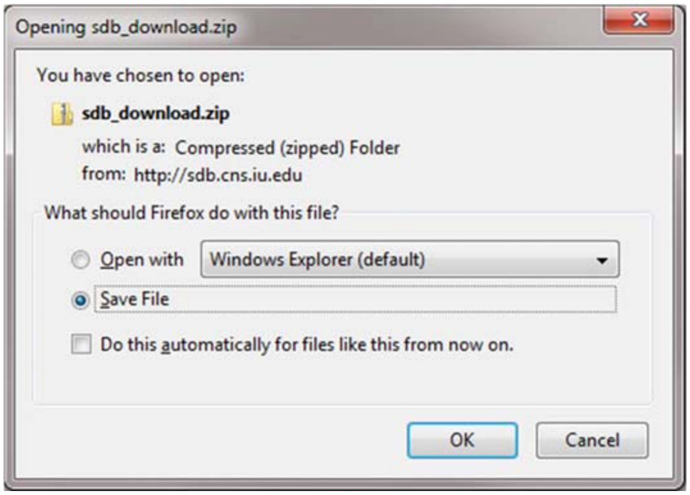
Download 1000 records starting at record 1 from the following databases:

MEDLINE Database:

- □ MEDLINE author table
- □ MEDLINE co-author table (nwb format)
- ☑ MEDLINE master table
- □ MEDLINE MeSH heading table
- □ MEDLINE MeSH qualifier table

Your download limit will initially capped at 2000 records at a time. To increase this limit, please email cns-sdb-dev-l@iulist.indiana.edu

Save the file somewhere on your computer for use later in this tutorial…



Opening sdb_download.zip

You have chosen to open:

sdb_download.zip

which is a: Compressed (zipped) Folder
from: http://sdb.cns.iu.edu

What should Firefox do with this file?

○ Open with Windows Explorer (default)

● Save File

□ Do this automatically for files like this from now on.

Extract Word Co-Occurrence

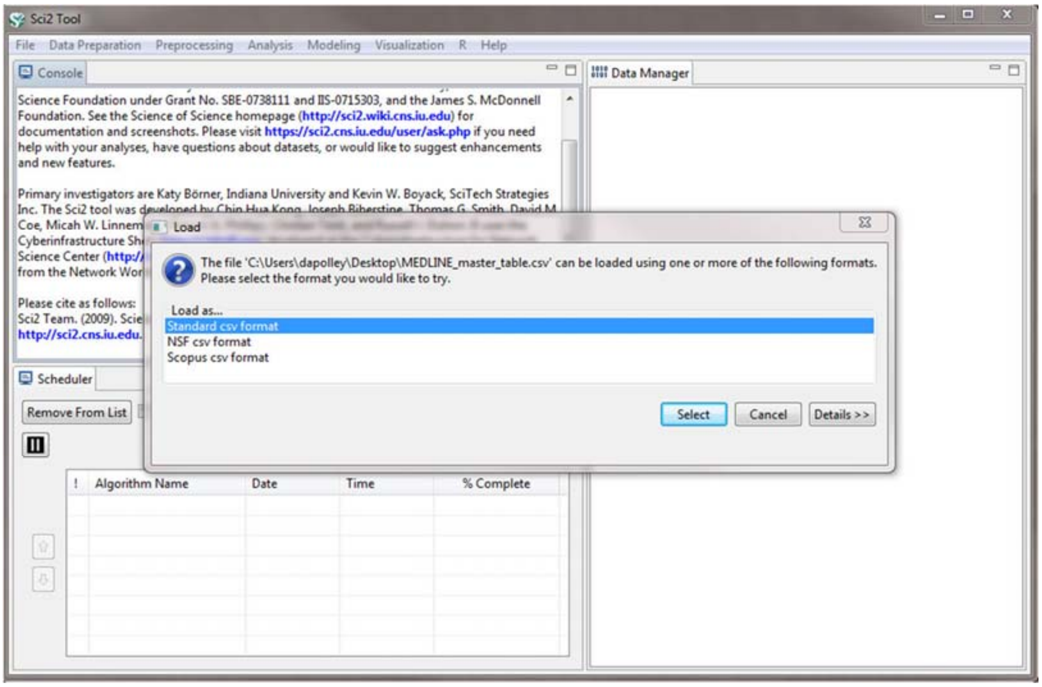
The topic similarity of basic and aggregate units of science can be calculated via an analysis of the co-occurrence of words in associated texts. Units that share more words in common are assumed to have higher topical overlap and are connected via linkages and/or placed in closer proximity.

Extract Word Co-Occurrence Network creates a weighted network where each node is a word and edges connect words to each other, where the strength of an edge represents how often two words occur in the same body of text together.

This algorithm is a shortcut for extracting a directed network using Extract Directed Network, and then performing bibliographic coupling using Extract Reference Co-Occurrence (Bibliographic Coupling) Network.

Extracting Word Co-Occurrence Network from SDB Data

Open Sci2 and load the MEDLINE_master_table.csv file as a Standard CSV file…



Sci2 Tool
File Data Preparation Preprocessing Analysis Modeling Visualization R Help
Console
Data Manager
Scheduler
Remove From List
! Algorithm Name Date Time % Complete
Load
Select Cancel Details >>

Science Foundation under Grant No. SBE-0738111 and IIS-0715303, and the James S. McDonnell Foundation. See the Science of Science homepage (http://sci2.wiki.cns.iu.edu) for documentation and screenshots. Please visit https://sci2.cns.iu.edu/user/ask.php if you need help with your analyses, have questions about datasets, or would like to suggest enhancements and new features.

Primary investigators are Katy Börner, Indiana University and Kevin W. Boyack, SciTech Strategies Inc. The Sci2 tool was developed by Chin Hua Kong, Joseph Biberstine, Thomas G. Smith, David M. Coe, Micah W. Linnem Cyberinfrastructure Sh Science Center (http:/ from the Network Wor

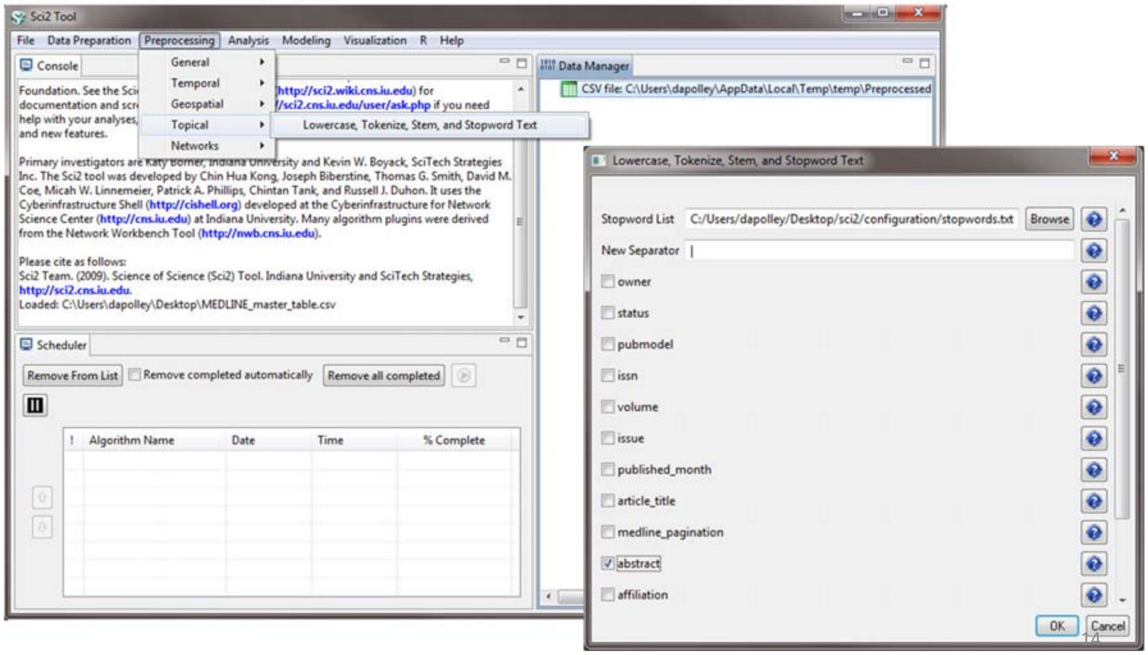
Please cite as follows:
Sci2 Team. (2009). Scie
http://sci2.cns.iu.edu.

The file 'C:\Users\dapolley\Desktop\MEDLINE_master_table.csv' can be loaded using one or more of the following formats.
Please select the format you would like to try.

Load as...
- Standard csv format
- NSF csv format
- Scopus csv format

Extracting Word Co-Occurrence Network from SDB Data

Normalize the titles by running *Preprocessing > Topical > Lowercase, Tokenize, Stem, and Stopword Text* and select "Abstract"…



Sci2 Tool
File Data Preparation Preprocessing Analysis Modeling Visualization R Help
General
Temporal
Geospatial
Topical
Networks
Lowercase, Tokenize, Stem, and Stopword Text
Console
Data Manager
Scheduler
Remove From List
Remove completed automatically
Remove all completed
! Algorithm Name Date Time % Complete
Browse
OK Cancel

Foundation. See the Sci (http://sci2.wiki.cns.iu.edu) for documentation and scr //sci2.cns.iu.edu/user/ask.php if you need help with your analyses, and new features.

Primary investigators are Katy Börner, Indiana University and Kevin W. Boyack, SciTech Strategies Inc. The Sci2 tool was developed by Chin Hua Kong, Joseph Biberstine, Thomas G. Smith, David M. Coe, Micah W. Linnemeier, Patrick A. Phillips, Chintan Tank, and Russell J. Duhon. It uses the Cyberinfrastructure Shell (http://cishell.org) developed at the Cyberinfrastructure for Network Science Center (http://cns.iu.edu) at Indiana University. Many algorithm plugins were derived from the Network Workbench Tool (http://nwb.cns.iu.edu).

Please cite as follows:
Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, http://sci2.cns.iu.edu.
Loaded: C:\Users\dapolley\Desktop\MEDLINE_master_table.csv

CSV file: C:\Users\dapolley\AppData\Local\Temp\temp\Preprocessed

Lowercase, Tokenize, Stem, and Stopword Text

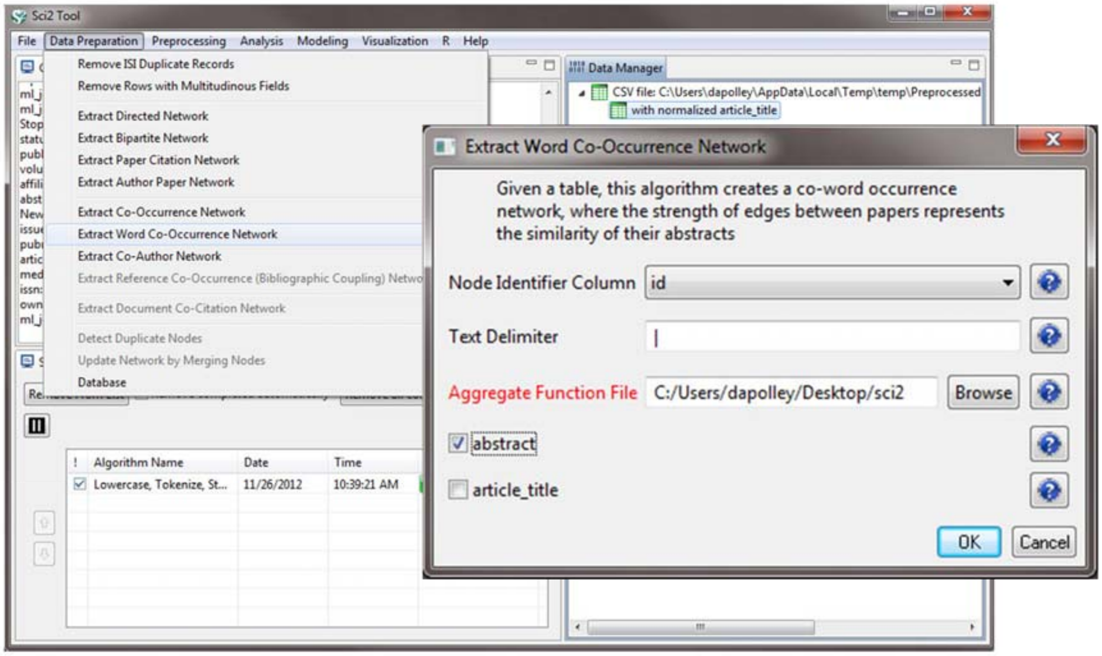
Stopword List C:/Users/dapolley/Desktop/sci2/configuration/stopwords.txt

New Separator

- □ owner
- □ status
- □ pubmodel
- □ issn
- □ volume
- □ issue
- □ published_month
- □ article_title
- □ medline_pagination
- ☑ abstract
- □ affiliation

Extracting Word Co-Occurrence Network from SDB Data

Run *Extract Word Co-Occurrence Network* and set the parameters as shown below…


Sci2 Tool
File
Data Preparation
Preprocessing
Analysis
Modeling
Visualization
R
Help
Remove ISI Duplicate Records
Remove Rows with Multitudinous Fields
Extract Directed Network
Extract Bipartite Network
Extract Paper Citation Network
Extract Author Paper Network
Extract Co-Occurrence Network
Extract Word Co-Occurrence Network
Extract Co-Author Network
Extract Reference Co-Occurrence (Bibliographic Coupling) Netwo
Extract Document Co-Citation Network
Detect Duplicate Nodes
Update Network by Merging Nodes
Database
Data Manager
Browse
OK
Cancel

CSV file: C:\Users\dapolley\AppData\Local\Temp\temp\Preprocessed
with normalized article_title

Extract Word Co-Occurrence Network

Given a table, this algorithm creates a co-word occurrence network, where the strength of edges between papers represents the similarity of their abstracts

Node Identifier Column: id

Text Delimiter: |

Aggregate Function File: C:/Users/dapolley/Desktop/sci2

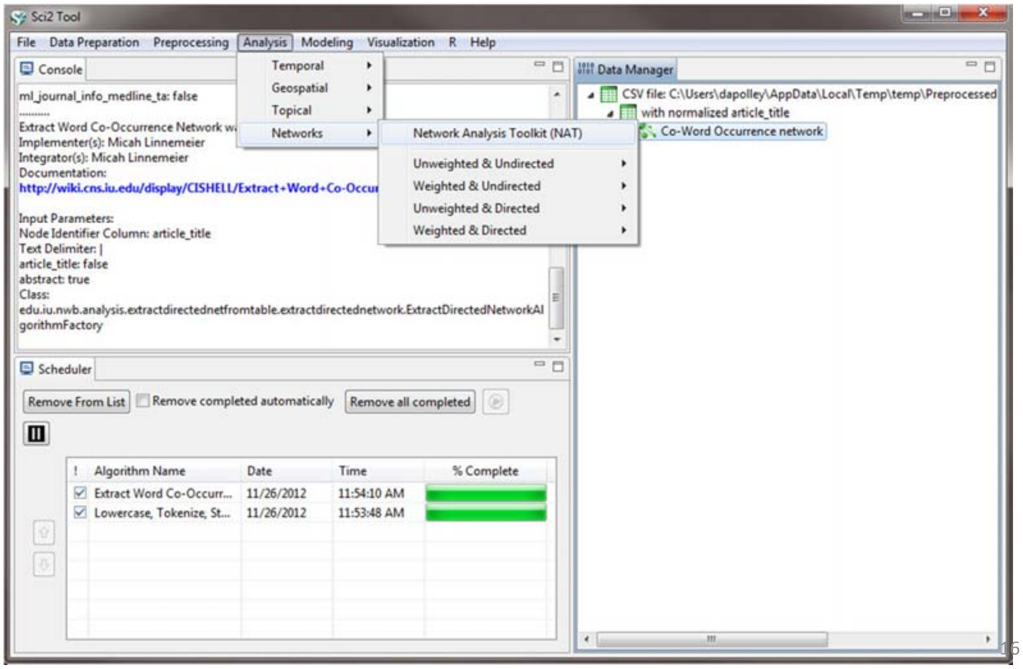
☑ abstract

□ article_title

!	Algorithm Name	Date	Time
☑	Lowercase, Tokenize, St...	11/26/2012	10:39:21 AM

Extracting Word Co-Occurrence Network from SDB Data

To see more information about your network run *Analysis > Networks > Network Analysis Toolkit*…



```
ml_journal_info_medline_ta: false
..........
Extract Word Co-Occurrence Network w
Implementer(s): Micah Linnemeier
Integrator(s): Micah Linnemeier
Documentation:
http://wiki.cns.iu.edu/display/CISHELL/Extract+Word+Co-Occur

Input Parameters:
Node Identifier Column: article_title
Text Delimiter: |
article_title: false
abstract: true
Class:
edu.iu.nwb.analysis.extractdirectednetfromtable.extractdirectednetwork.ExtractDirectedNetworkAl
gorithmFactory
```

CSV file: C:\Users\dapolley\AppData\Local\Temp\temp\Preprocessed
with normalized article_title
Co-Word Occurrence network

!	Algorithm Name	Date	Time	% Complete
☑	Extract Word Co-Occurr...	11/26/2012	11:54:10 AM	🟩
☑	Lowercase, Tokenize, St...	11/26/2012	11:53:48 AM	🟩

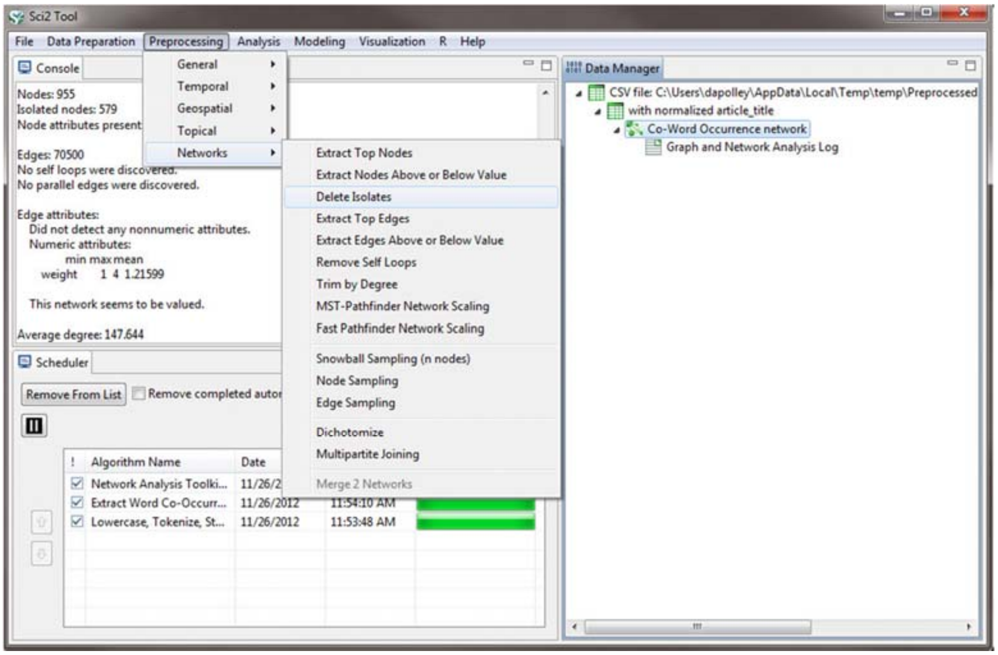
Extracting Word Co-Occurrence Network from SDB Data

The results from Network Analysis Toolkit show that there are 579 isolated nodes…

```
This graph claims to be undirected.

Nodes: 4846
Isolated nodes: 1000
Node attributes present: label, references

Edges: 499196
No self loops were discovered.
No parallel edges were discovered.

Edge attributes:
Did not detect any nonnumeric attributes.
Numeric attributes:
minmaxmean
weight     12421.7084

This network seems to be valued.

Average degree: 206.0239
This graph is not weakly connected.
There are 1001 weakly connected components. (1000 isolates)
The largest connected component consists of 3846 nodes.
Did not calculate strong connectedness because this graph was
not directed.

Density (disregarding weights): 0.0425
Additional Densities by Numeric Attribute
```

Extracting Word Co-Occurrence Network from SDB Data

Delete the isolate nodes by running *Preprocessing > Networks > Delete Isolates*



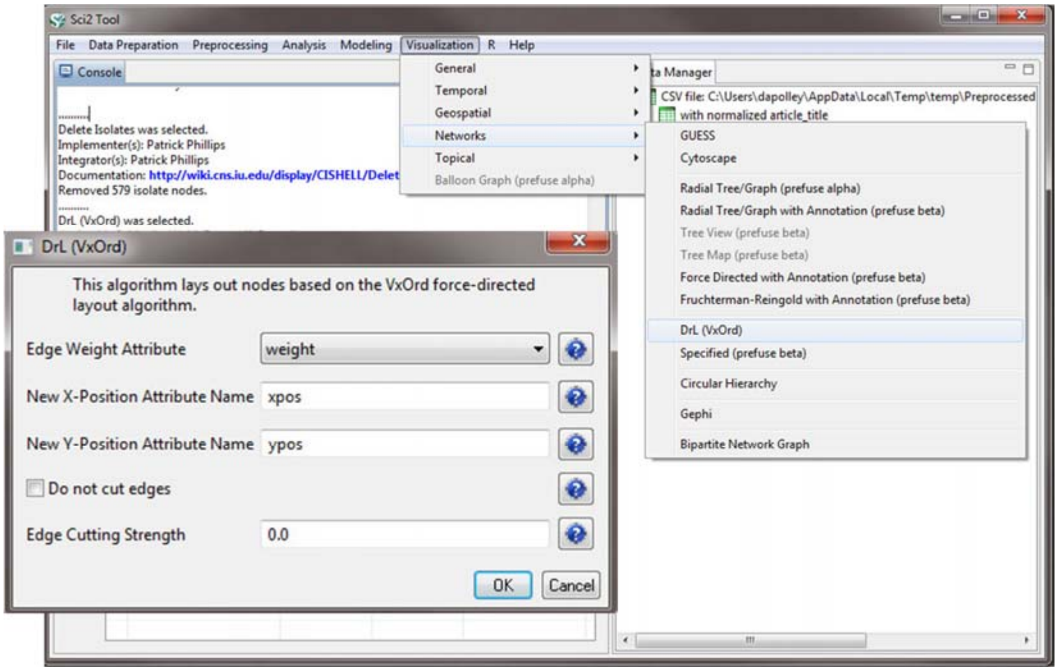
```
Nodes: 955
Isolated nodes: 579
Node attributes present

Edges: 70500
No self loops were discovered.
No parallel edges were discovered.

Edge attributes:
  Did not detect any nonnumeric attributes.
  Numeric attributes:
        min max mean
    weight    1  4  1.21599

  This network seems to be valued.

Average degree: 147.644
```

!	Algorithm Name	Date		
☑	Network Analysis Toolki...	11/26/2		
☑	Extract Word Co-Occurr...	11/26/2012	11:54:10 AM	
☑	Lowercase, Tokenize, St...	11/26/2012	11:53:48 AM	

CSV file: C:\Users\dapolley\AppData\Local\Temp\temp\Preprocessed
with normalized article_title
Co-Word Occurrence network
Graph and Network Analysis Log

Extracting Word Co-Occurrence Network from SDB Data

Apply *Visualization > Networks > DrL (VxOrd)* and words that are similar will be plotted relatively close to each other. Set the parameters to those shown below…



```
..........
Delete Isolates was selected.
Implementer(s): Patrick Phillips
Integrator(s): Patrick Phillips
Documentation: http://wiki.cns.iu.edu/display/CISHELL/Delet
Removed 579 isolate nodes.
..........
DrL (VxOrd) was selected.
```

CSV file: C:\Users\dapolley\AppData\Local\Temp\temp\Preprocessed
with normalized article_title

DrL (VxOrd)

This algorithm lays out nodes based on the VxOrd force-directed layout algorithm.

Edge Weight Attribute: weight

New X-Position Attribute Name: xpos

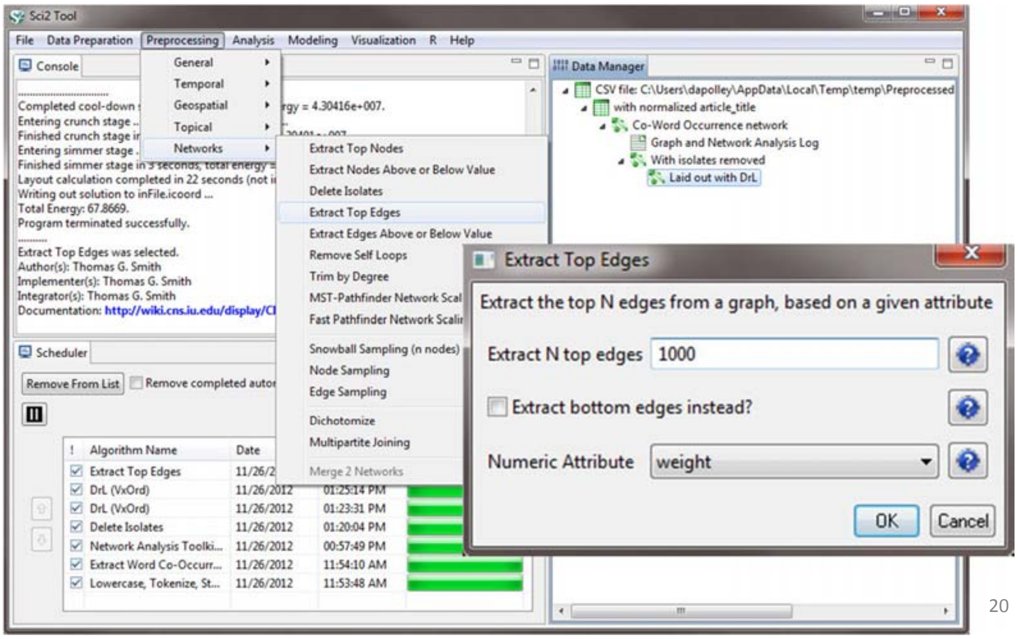
New Y-Position Attribute Name: ypos

□ Do not cut edges

Edge Cutting Strength: 0.0

Extracting Word Co-Occurrence Network from SDB Data

Laying out the network with Drl (VxOrd) may take some time, but once the algorithm is complete you will want to keep only the strongest edges, so select the "Laid out with DrL" and run *Preprocessing > Networks > Extract Top Edges* using the parameters shown below…



```
..............................
Completed cool-down s                rgy = 4.30416e+007.
Entering crunch stage ..
Finished crunch stage ir
Entering simmer stage .
Finished simmer stage in 3 seconds, total energy =
Layout calculation completed in 22 seconds (not i
Writing out solution to inFile.icoord ...
Total Energy: 67.8669.
Program terminated successfully.
..........
Extract Top Edges was selected.
Author(s): Thomas G. Smith
Implementer(s): Thomas G. Smith
Integrator(s): Thomas G. Smith
Documentation: http://wiki.cns.iu.edu/display/CI
```

!	Algorithm Name	Date		
☑	Extract Top Edges	11/26/2		
☑	DrL (VxOrd)	11/26/2012	01:25:14 PM	
☑	DrL (VxOrd)	11/26/2012	01:23:31 PM	
☑	Delete Isolates	11/26/2012	01:20:04 PM	
☑	Network Analysis Toolki...	11/26/2012	00:57:49 PM	
☑	Extract Word Co-Occurr...	11/26/2012	11:54:10 AM	
☑	Lowercase, Tokenize, St...	11/26/2012	11:53:48 AM	

- CSV file: C:\Users\dapolley\AppData\Local\Temp\temp\Preprocessed
  - with normalized article_title
    - Co-Word Occurrence network
      - Graph and Network Analysis Log
      - With isolates removed
        - Laid out with DrL

Extract Top Edges

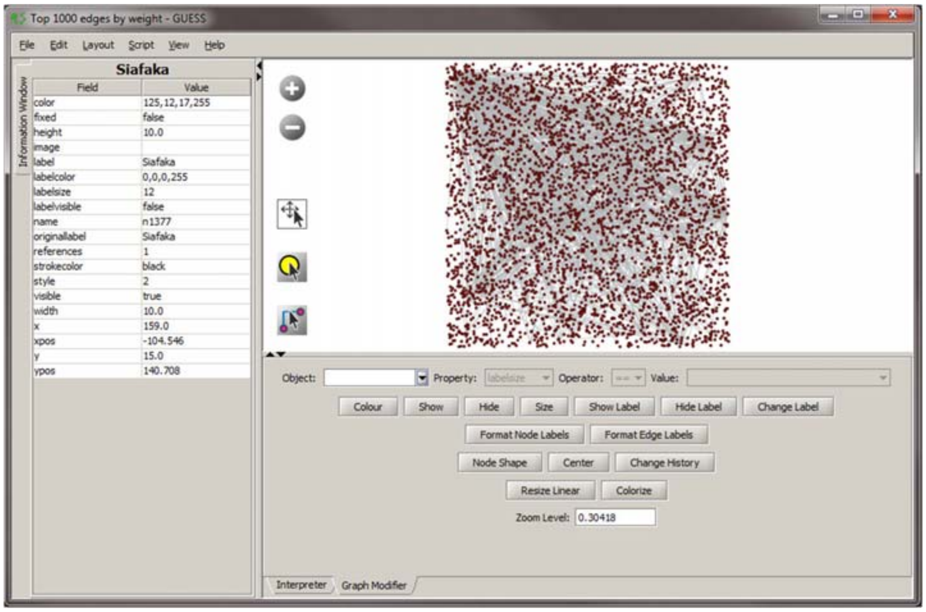
Extract the top N edges from a graph, based on a given attribute

Extract N top edges: 1000

□ Extract bottom edges instead?

Numeric Attribute: weight

Once edges have been removed, the network "top 1000 edges by weight" can be visualized by running *Visualization > Networks > GUESS*…

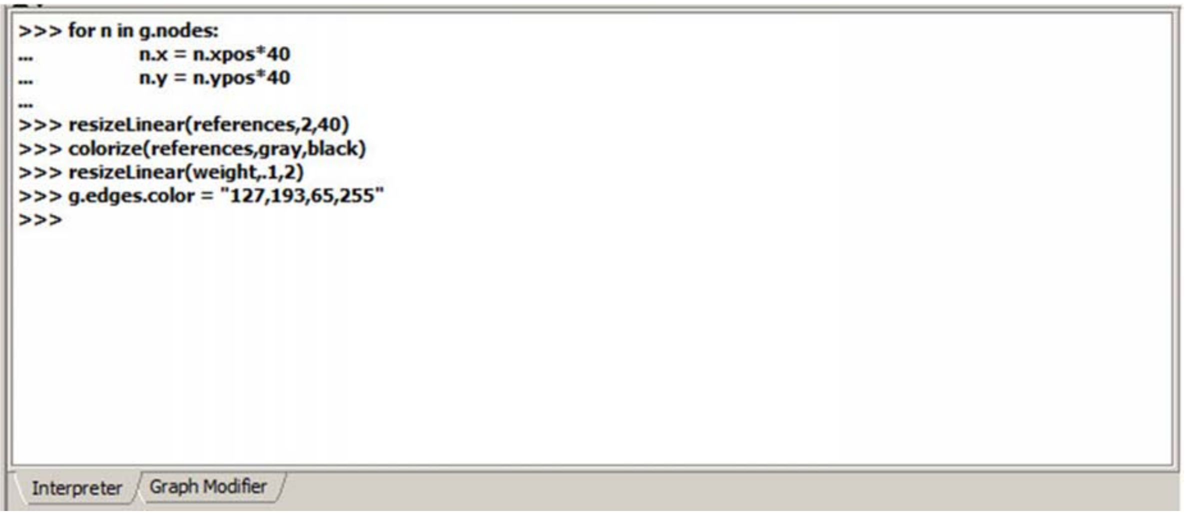


Siafaka

Field	Value
color	125,12,17,255
fixed	false
height	10.0
image	
label	Siafaka
labelcolor	0,0,0,255
labelsize	12
labelvisible	false
name	n1377
originallabel	Siafaka
references	1
strokecolor	black
style	2
visible	true
width	10.0
x	159.0
xpos	-104.546
y	15.0
ypos	140.708

Zoom Level: 0.30418

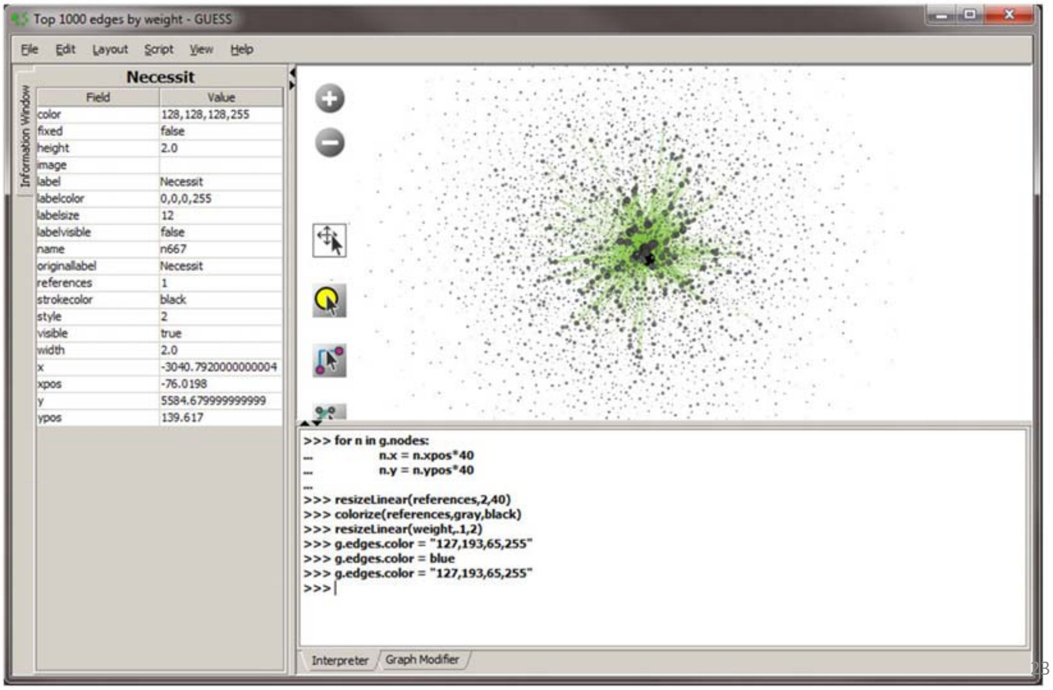
In order to make use of the DrL (VxOrd) force directed layout we applied, we need to change to the interpreter at the bottom of the screen and type in the following commands…



```
>>> for n in g.nodes:
...          n.x = n.xpos*40
...          n.y = n.ypos*40
...
>>> resizeLinear(references,2,40)
>>> colorize(references,gray,black)
>>> resizeLinear(weight,.1,2)
>>> g.edges.color = "127,193,65,255"
>>>
```

Extracting Word Co-Occurrence Network from SDB Data

Note, GUESS will not necessarily display the graph in the middle of the screen, you may have to scroll around the screen to find the graph.


Top 1000 edges by weight - GUESS
File Edit Layout Script View Help
Information Window
Interpreter
Graph Modifier

Necessit

Field	Value
color	128,128,128,255
fixed	false
height	2.0
image	
label	Necessit
labelcolor	0,0,0,255
labelsize	12
labelvisible	false
name	n667
originallabel	Necessit
references	1
strokecolor	black
style	2
visible	true
width	2.0
x	-3040.7920000000004
xpos	-76.0198
y	5584.679999999999
ypos	139.617

```
>>> for n in g.nodes:
...             n.x = n.xpos*40
...             n.y = n.ypos*40
...
>>> resizeLinear(references,2,40)
>>> colorize(references,gray,black)
>>> resizeLinear(weight,.1,2)
>>> g.edges.color = "127,193,65,255"
>>> g.edges.color = blue
>>> g.edges.color = "127,193,65,255"
>>> |
```

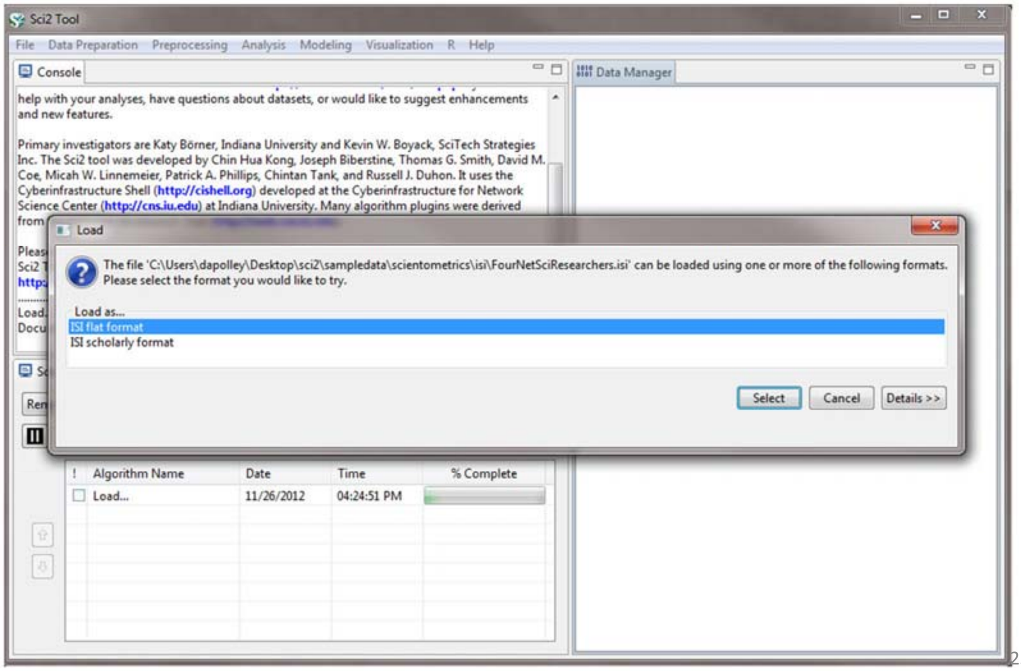
Overlaying ISI Data on the Map of Science via Journals

The Map of Science is a visual representation of 554 sub-disciplines within 13 disciplines of science and their relationships to one another, shown as points and lines connecting those points respectively. Over top this visualization is drawn the result of mapping a dataset's journals to the underlying sub-discipline(s) those journals contain. Mapped sub-disciplines are shown with size relative to the number matching journals and color from the discipline. For more information on maps of science, see http://mapofscience.com

As of the Sci2 v1.0 alpha release there is a plugin for Sci2 that allows users to visualize their own data overlaid on the Map of Science

Overlaying ISI Data on the Map of Science via Journals

Load the FourNetSciResearchers.isi file in the ISI flat format…



help with your analyses, have questions about datasets, or would like to suggest enhancements and new features.

Primary investigators are Katy Börner, Indiana University and Kevin W. Boyack, SciTech Strategies Inc. The Sci2 tool was developed by Chin Hua Kong, Joseph Biberstine, Thomas G. Smith, David M. Coe, Micah W. Linnemeier, Patrick A. Phillips, Chintan Tank, and Russell J. Duhon. It uses the Cyberinfrastructure Shell (**http://cishell.org**) developed at the Cyberinfrastructure for Network Science Center (**http://cns.iu.edu**) at Indiana University. Many algorithm plugins were derived from

Load

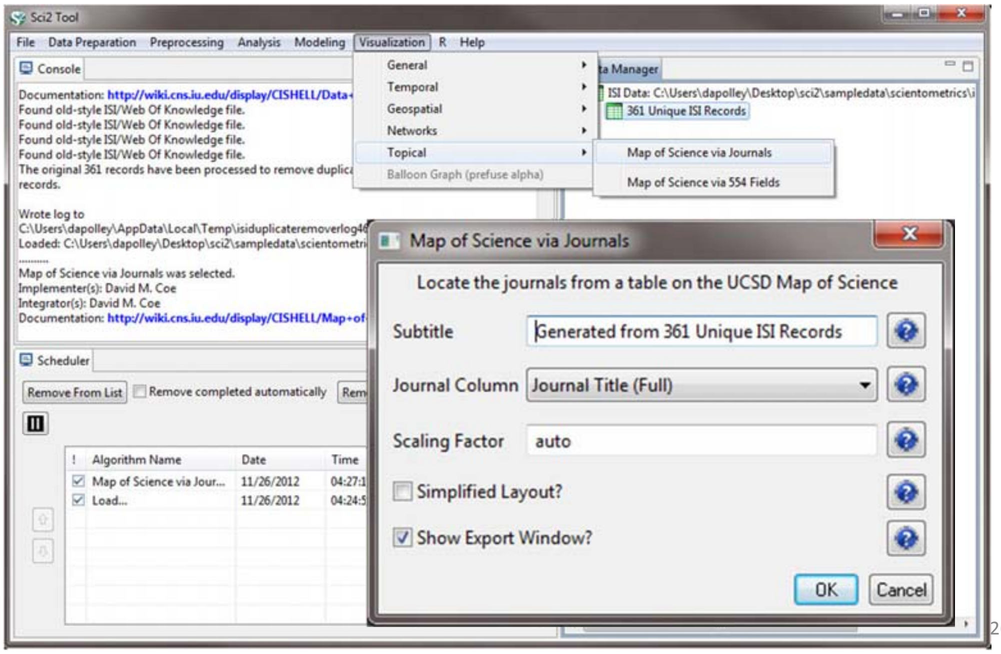
The file 'C:\Users\dapolley\Desktop\sci2\sampledata\scientometrics\isi\FourNetSciResearchers.isi' can be loaded using one or more of the following formats. Please select the format you would like to try.

Load as...
- ISI flat format
- ISI scholarly format

!	Algorithm Name	Date	Time	% Complete
□	Load...	11/26/2012	04:24:51 PM	

Overlaying ISI Data on the Map of Science via Journals

To visualize dataset overlaid on the Map of Science run *Visualization > Topical > Map of Science via Journals*



Documentation: **http://wiki.cns.iu.edu/display/CISHELL/Data+**
Found old-style ISI/Web Of Knowledge file.
Found old-style ISI/Web Of Knowledge file.
Found old-style ISI/Web Of Knowledge file.
Found old-style ISI/Web Of Knowledge file.
The original 361 records have been processed to remove duplica
records.

Wrote log to
C:\Users\dapolley\AppData\Local\Temp\isiduplicateremoverlog4
Loaded: C:\Users\dapolley\Desktop\sci2\sampledata\scientometri
....
Map of Science via Journals was selected.
Implementer(s): David M. Coe
Integrator(s): David M. Coe
Documentation: **http://wiki.cns.iu.edu/display/CISHELL/Map+of**

ISI Data: C:\Users\dapolley\Desktop\sci2\sampledata\scientometrics\i
361 Unique ISI Records

□ Remove completed automatically

!	Algorithm Name	Date	Time
☑	Map of Science via Jour...	11/26/2012	04:27:1
☑	Load...	11/26/2012	04:24:5

Map of Science via Journals

Locate the journals from a table on the UCSD Map of Science

Subtitle: Generated from 361 Unique ISI Records

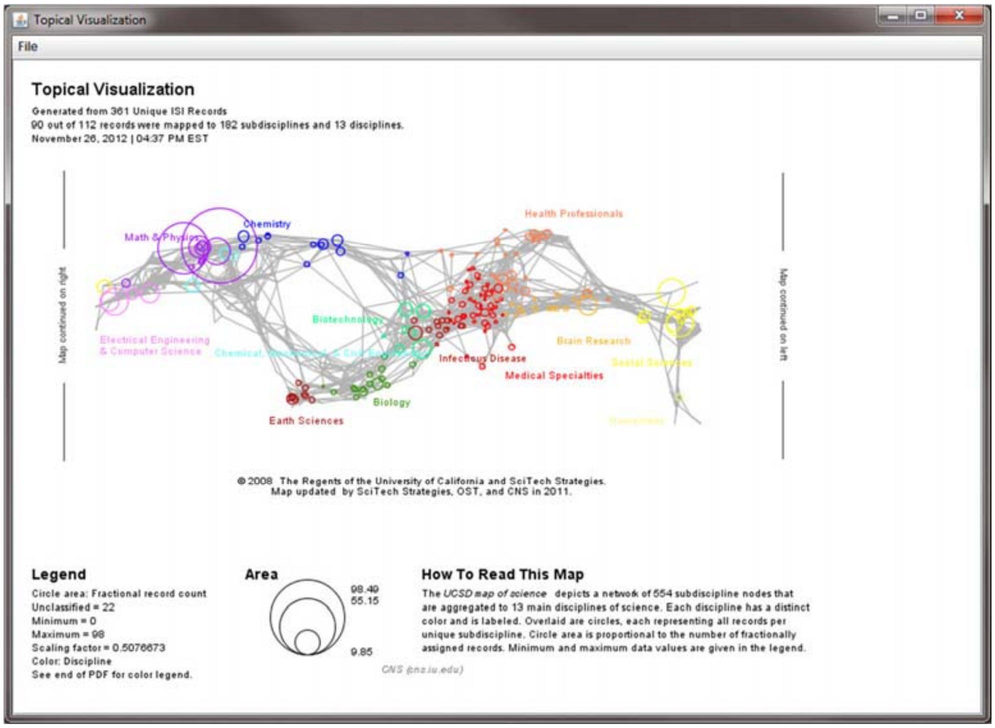
Journal Column: Journal Title (Full)

Scaling Factor: auto

□ Simplified Layout?

☑ Show Export Window?

Overlaying ISI Data on the Map of Science via Journals



Topical Visualization

Generated from 361 Unique ISI Records
90 out of 112 records were mapped to 182 subdisciplines and 13 disciplines.
November 26, 2012 | 04:37 PM EST

© 2008 The Regents of the University of California and SciTech Strategies.
Map updated by SciTech Strategies, OST, and CNS in 2011.

Legend

Circle area: Fractional record count
Unclassified = 22
Minimum = 0
Maximum = 98
Scaling factor = 0.5076673
Color: Discipline
See end of PDF for color legend.

Area

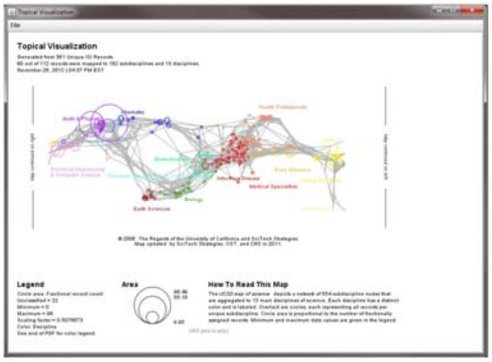
How To Read This Map

The *UCSD map of science* depicts a network of 554 subdiscipline nodes that are aggregated to 13 main disciplines of science. Each discipline has a distinct color and is labeled. Overlaid are circles, each representing all records per unique subdiscipline. Circle area is proportional to the number of fractionally assigned records. Minimum and maximum data values are given in the legend.

CNS (cns.iu.edu)

Overlaying ISI Data on the Map of Science via Journals

The journals titles are used to determine which records fit into what subdiscipline. You can view the journal titles found and those not found from the data manager of Sci2. A single journal can belong to more than one subdiscipline and thus so can the record associated with that journal. So the circle sizes are proportional to the number of fractionally assigned records.



Questions?