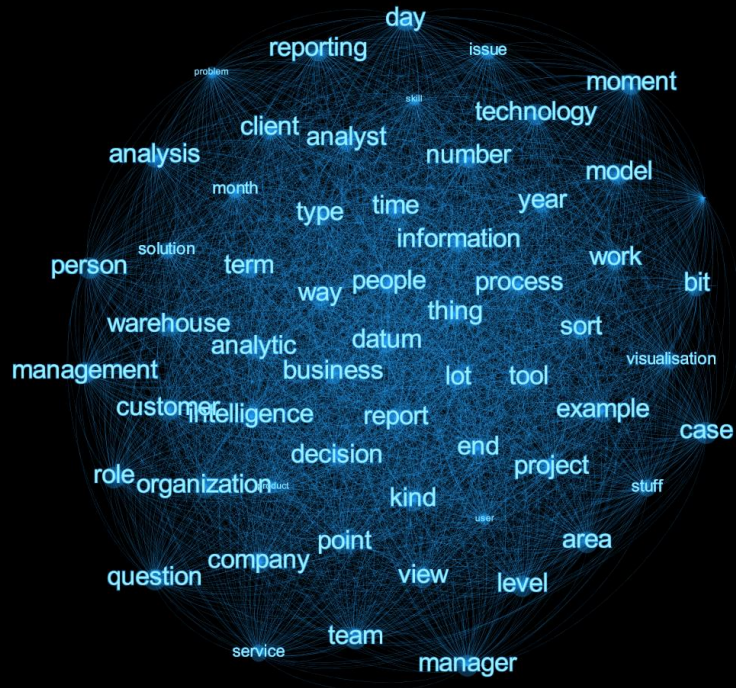


PRINCIPLES OF TEXT VISUALISATION:

TEXT REPRESENTATION, EXPLORATION AND INSIGHT GENERATION



Jacob L. Cybulski

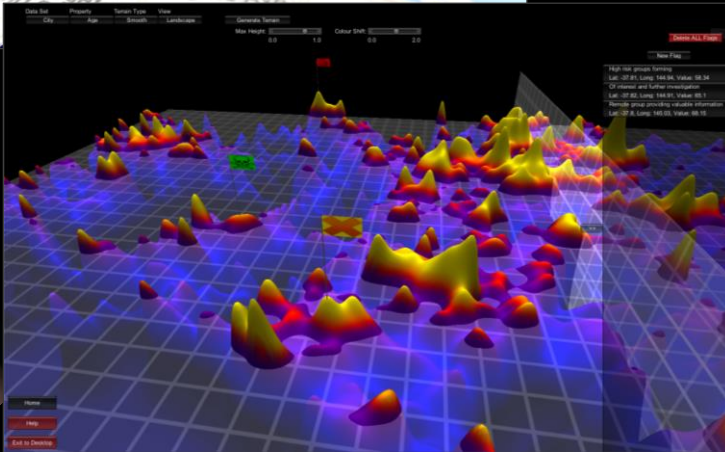
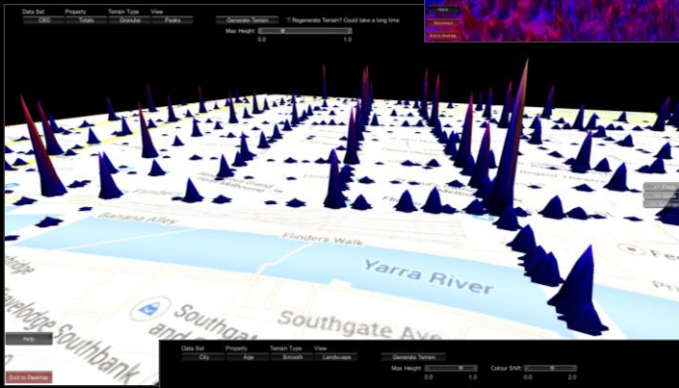
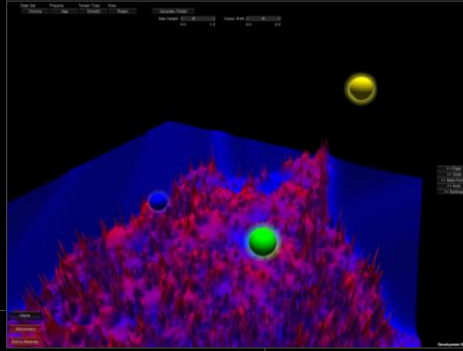
*Visual Analytics Collaboratory
Dept of Info Sys and Bus Analytics*

*Deakin Business School
Faculty of Business and Law
Deakin University*

*To capture the essence of
information in the moment of time*

VISUAL ANALYTICS COLLABORATORY / JACOB'S PROJECTS

Research

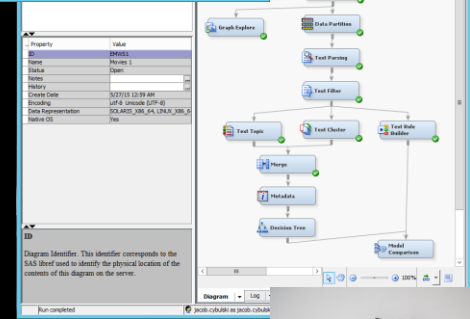
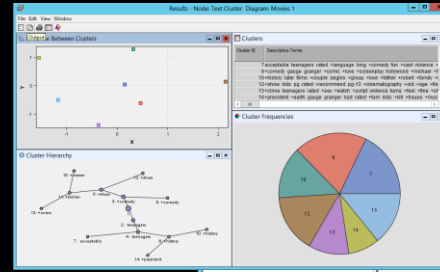


Collaborative & Interactive 3D Visual Analytics

SAS, R
RapidMiner
D3.js Three.js
Unity 3D

Immersive
Visual Analytics

Exploration
Interactivity &
Collaboration



Devices

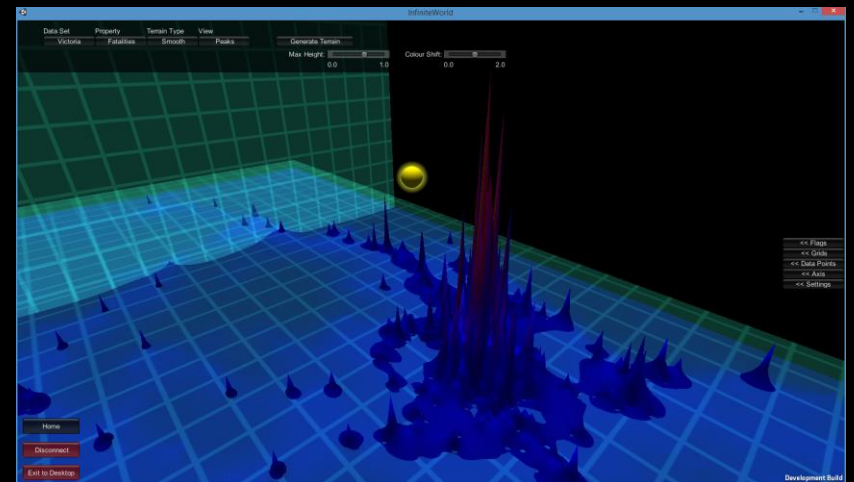
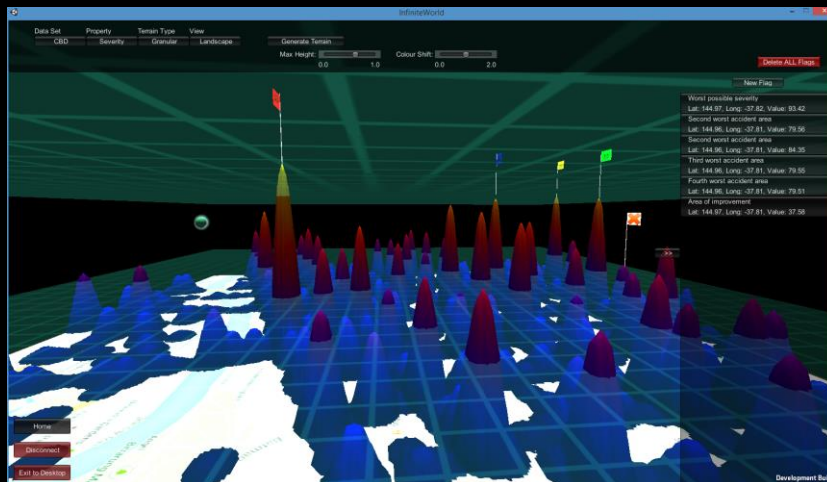
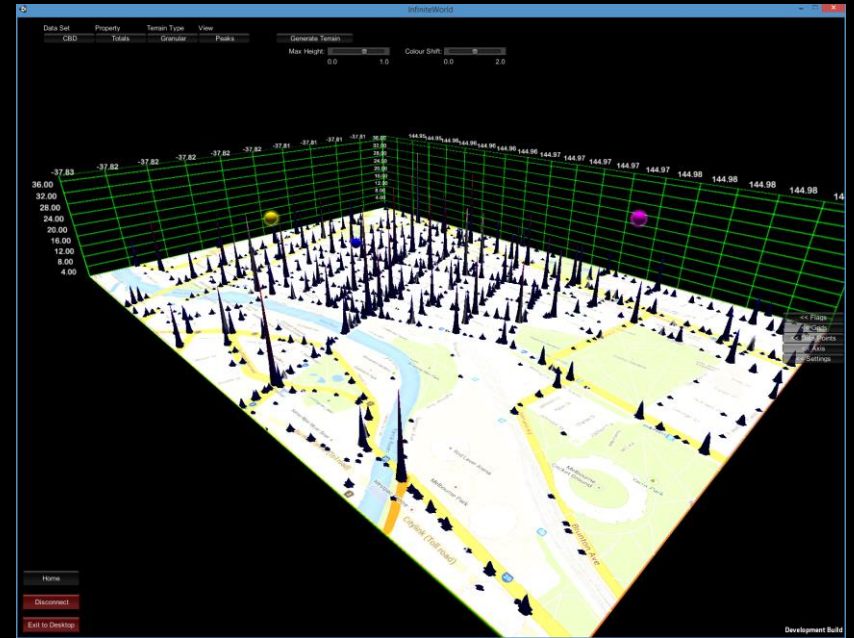
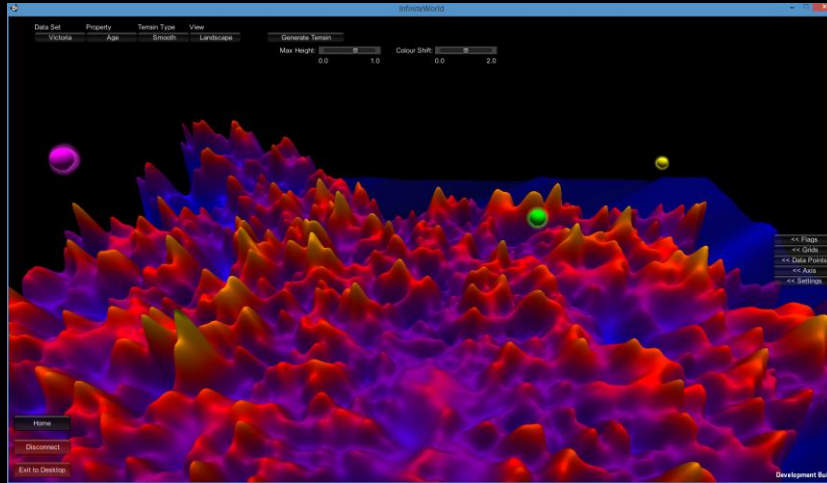


Education



RESEARCH: 3D VISUALISATION INTERACTIVITY AND COLLABORATION

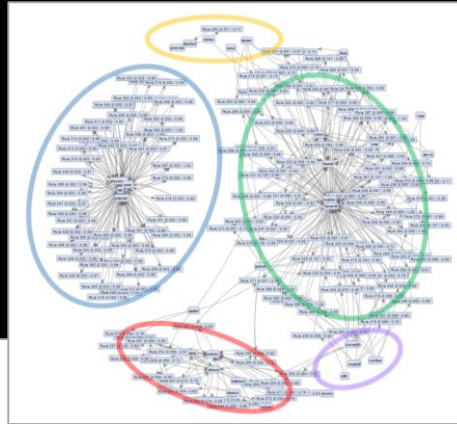
Custom-built environment: Visual Analyst 3D (Unity 3D / C#)



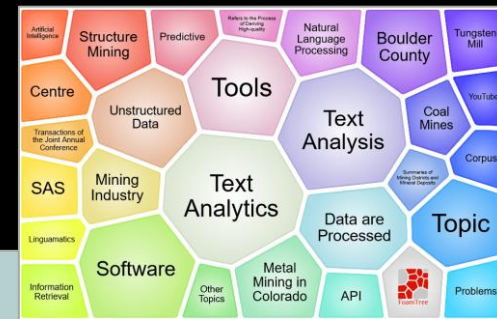
TEXT MINING and TOOLS

WHY TEXT?

Association Rules with RapidMiner

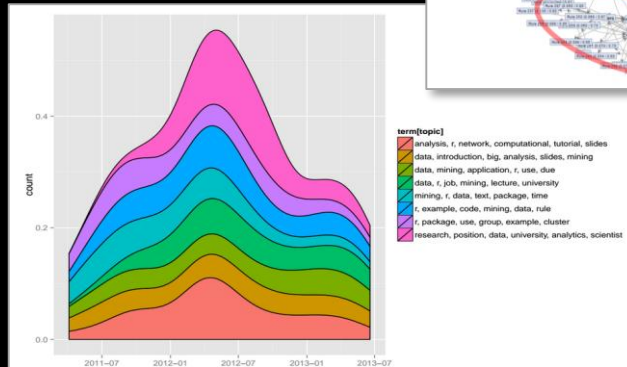


Foam Trees with Carrot 2



However much can be done with some of the off-the-shelf visualisation tools

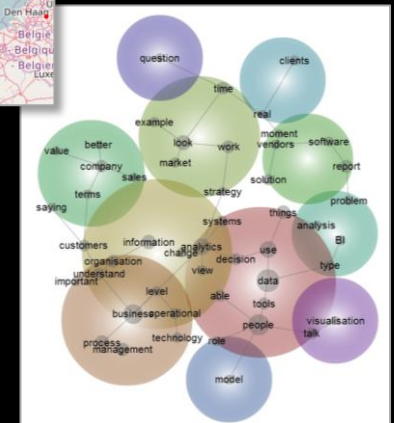
Topic Modelling with R



Twitter Analysis with Python



Term Clusters with Leximancer



Open Source

- KH Coder
- Carrot2
- RapidMiner, KNIME, Orange
- R and Python
- And more...

Commercial

- Leximancer
- IBM SPSS Modeler
- SAS Enterprise Miner
- Microsoft Cortana Intelligence
- And more...

QUESTION

How can data mining and data visualization assist analysis of interview transcripts?

KH Coder (Koichi HIGUCHI, Ritsumeikan University)
Gephi (Gephi Consortium)

FREE TEXT

At the beginning there was text.

And the text was without *form*, and *void*; and darkness was upon the face of the data analyst...

Interview with Daniel

Facilitator: So what's your role here? What type of tasks and jobs do you do?

Daniel: So, data analyst is my role and a lot of times it's dealing with clients about knowing and finding out exactly what they want to report on and helping them and engaging them with tools. So, the tools we're using currently are ProcureTrak or Omniscope, and that's a data visualisation tool. And so it's a lot to do with client relationship but also, obviously, with their data. So, it's pretty much end to end - so, get the data from them, go through the data, put some business rules around the data, see what type of reports they want, giving them back the reports seeing if that's what they want to see or if they want to see something else. So adjust the reports based on what they want to see.

... and 5891 more paragraphs of this kind in 27 interviews ...

KH CODER

- KH Coder was developed by Koichi Higuchi, at Ritsumeikan University, Japan.
- It is a free software for quantitative analysis of text.
- While it was originally developed for Japanese language, it is also available for the analysis of text in English, French, German, Italian, Portuguese, Spanish, Russian, Chinese (simplified) and Korean.
- It has a rich suite of tool for lexical analysis, as well as, document and term analysis and visualisation.
- It includes the following features:
 - Lexical analysis of text
 - Statistical reports
 - Correspondence analysis
 - Multi-dimensional scaling
 - Term clustering and analysis
 - Bayesian modelling and classification

The image displays three windows from the KH Coder software interface:

- KWC Concordance:** Shows a search entry for the word 'client' with various filters and options. The result window displays a list of text segments where 'client' appears, such as "you do? (1) Daniel: So... data analyst is my role and a lot of times it's dealing with clients about".
- Collocation Stats:** A table showing the frequency of words co-occurring with 'client'. The table has columns for Word, POS, Total, and various RT (Right Triangles) and R4 (Right Squares) counts. For example, 'want' has a total of 17 and RT1 of 16.
- Multi-Dimensional Scaling of Words:** Two 3D scatter plots showing word positions in a multi-dimensional space. The top plot shows words like 'understand', 'process', 'client', 'mean', 'start', 'year', 'company', 'talk', 'project', 'number', 'question', 'information', 'datum', 'be', 'term', 'scope', 'real', 'report', 'type', 'analytic', 'investigation', 'decision', 'intelligence', 'scope', 'real'. The bottom plot shows words like 'report', 'client', 'decision', 'customer', 'process', 'understand', 'intelligence', 'level', 'visualisation', 'business', 'information', 'organisation', 'type', 'company', 'like', 'team', 'model', 'help', 'term', 'project', 'example', 'start', 'work', 'year', 'happen', 'number', 'time', 'date', 'day', 'work', 'question', 'bit', 'stuff', 'know', 'look', 'way', 'people', 'help', 'model', 'year', 'work', 'start', 'example', 'project', 'bit'. The stress values are 0.047 and 0.150 respectively.

UNIT OF ANALYSIS: DOCUMENTS AND TERMS

Sentences

The first step in text analytics is to decide on your unit of analysis:

- Sentence
- Paragraph
- Interview
- Record

id	length_c	length_w	datum	people	thing	business	report	lot	time	way
7	287	154	6	0	0	1	3	2	1	0
8	25.5	11	0	0	0	0	0	0	0	0
9	12	5	0	0	0	0	0	0	0	0
10	95	40	0	0	1	0	0	1	0	0
11	37	14	0	0	0	0	0	1	0	0
12	5.5	3	0	0	0	0	0	0	0	0
13	119.5	61	0	0	0	0	0	0	0	0
14	275.5	157	5	1	0	0	0	0	1	1

Paragraphs

We often call our chosen unit of analysis *document*

Each document is then parsed and split into words and the words are changed into a standard form, known as “stem”, “root” or “lemma”

id	length_c	length_w	datum	people	thing	business	report	lot	time	way
1	165	81	0	0	0	2	0	0	0	0
2	164.5	78	0	0	0	0	0	0	0	0
3	31.5	18	0	0	0	0	0	0	0	0
4	287	154	6	0	0	1	3	2	1	0
5	37.5	16	0	0	0	0	0	0	0	0
6	95	40	0	0	1	0	0	1	0	0
7	37	14	0	0	0	0	0	1	0	0
8	5.5	3	0	0	0	0	0	0	0	0
9	119.5	61	0	0	0	0	0	0	0	0
10	275.5	157	5	1	0	0	0	0	1	1

We will call it a *term*

A common representation of documents is in the form of *vectors of term frequencies*

Such vectors could be huge!

Interviews

name	length_c	length_w	datum	people	thing	business	report	lot	time	way
Jeffrey	22384.5	11195	64	64	31	65	48	18	17	9
Ross	35140.5	18169	148	61	119	28	7	28	26	48
Andrew	20419	10599	72	32	51	23	4	27	28	20
Clark	24425.5	12513	64	67	51	69	22	41	26	35
Ian	13739	6873	58	12	18	20	42	13	19	5
Rachel	21534.5	10744	49	36	32	55	26	22	9	10
Ruofan	24076	12297	167	35	18	59	27	15	34	19
Sahil	19800	10009	119	17	20	22	32	7	31	17

Term parts of speech and frequency table

Noun	ProperNoun	TAG	Adj	Adv	Verb	
datum	2021 BI	216 Facilitator:	1353 different	513 just	1116 think	1135
people	1033 Excel	112 Facilitator1:	272 able	326 actually	606 say	1090
thing	1031 Australia	75 Facilitator2:	191 good	270 really	566 know	1057
business	1024 Cognos	55 Shaun:	140 big	233 probably	356 use	936
report	738 Data	46 Hill:	139 sure	197 quite	229 look	892
lot	591 ASP	43 Ian:	126 new	188 maybe	166 want	806
time	562 Uni	42 Daniel:	117 important	166 right	163 need	762
way	535 SQL	40 Nathan:	111 particular	155 basically	149 make	689
tool	474 Microsoft	35 Jeffrey:	107 okay	143 usually	107 come	560
decision	468 SLT	34 Glenn:	105 great	127 obviously	94 work	514
process	462 White	34 Facilitator3:	93 interesting	126 certainly	89 try	411
analytic	452 Business	31 Jordan:	87 little	125 necessaril	79 understand	405
sort	440 ETL	31 Rachel:	87 better	117 absolutely	76 talk	371
informatio	438 Office	31 Madison:	83 real	117 exactly	75 mean	316
client	415 SharePoir	31 Arnaldo:	81 operatione	116 pretty	67 start	287
type	329 IBM	24 Myla:	81 certain	111 generally	66 happen	258
organizati	326 Melbourne	24 Ross:	76 right	95 definitely	65 like	205
project	324 Warehouse	23 Alfred:	73 specific	84 yes	63 help	203
company	323 PowerPoir	22 Andrew:	72 interested	83 far	56 ask	189
bit	293 SAS	22 Emily:	64 simple	81 quickly	55 create	189
year	280 SPSS	22 Scott:	60 technical	79 away	49 build	177
work	277 Oracle	21 Clark:	58 best	74 particularl	48 tell	156
number	276 Victoria	21 Ruofan:	57 financial	73 especially	46 provide	154
level	270 Group	20 Matt:	56 long	71 mainly	45 involve	144
term	263 Daniel	19 Robert:	56 useful	64 fairly	44 change	143
example	257 SAP	19 Chandler:	51 strategic	62 currently	39 run	137
kind	246 iPad	18 Sahil:	49 easy	61 directly	38 guess	135

TEXT UNITS CHARACTERISTICS

Documents, paragraphs or sentences can also be analysed for the co-occurrence of terms.

The statistics about text units can be used to define their "similarity" or "correspondence" (e.g. TF-IDF, term-context or entropy), which in turn can be interpreted as their distances.

Text units can then be visualized as points in multidimensional space.

Term-to-term "context" distance

As a side effect of the initial analysis we have access to lexical attributes of every term which appeared in each document. We also have vector representation of each document.

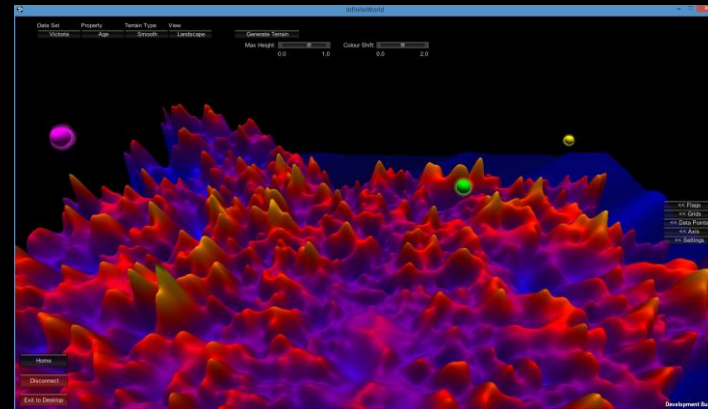
More importantly:

- Based on what terms appear in each document we can judge the documents similarity.
- Based on what documents the terms appear in we can also judge the terms similarity.

Word Context	cw: datum	cw: people	cw: thing	cw: business	cw: report
datum(2021)	1.657916325	0.23461854	0.2707137	0.273174733	0.150943396
people(1033)	0.47044335	1.272167488	0.307881773	0.333743842	0.227832512
thing(1031)	0.582070707	0.296717172	1.301767677	0.284090909	0.151515152
business(1022)	0.627192982	0.346491228	0.302631579	1.494152047	0.156432749
report(738)	0.487854251	0.299595142	0.208502024	0.230769231	1.493927126
lot(591)	0.636363636	0.393939394	0.426262626	0.341414141	0.195959596
time(562)	0.644210526	0.277894737	0.338947368	0.271578947	0.216842105
way(535)	0.723684211	0.320175439	0.412280702	0.353070175	0.155701754
tool(474)	0.549723757	0.281767956	0.218232044	0.29281768	0.138121547
decision(468)	0.484419263	0.359773371	0.198300283	0.348441926	0.235127479
process(461)	0.59375	0.321022727	0.261363636	0.397727273	0.15625
analytic(453)	0.491712707	0.314917127	0.201657459	0.580110497	0.143646409
sort(440)	0.673740053	0.305039788	0.557029178	0.297082228	0.137931034
information(438)	0.65	0.352941176	0.305882353	0.388235294	0.205882353
client(415)	0.648318043	0.155963303	0.302752294	0.226299694	0.226299694
type(329)	0.600682594	0.307167235	0.38225256	0.300341297	0.211604096
organization(326)	0.608058608	0.340659341	0.278388278	0.472527473	0.131868132
project(324)	0.612068966	0.262931034	0.284482759	0.379310345	0.137931034
company(323)	0.434456929	0.314606742	0.239700375	0.438202247	0.108614232

TEXT VISUALISATION

- At this point any collection of documents can be represented as a space of term vectors.
- Each of those terms defines a variable value, which can indicate frequency of that particular term in a document.
- Term frequencies are often weighted against the term frequency across all documents (TF-IDF) or they are transformed into entropy measures.
- There can be a very large number of term-variables, possibly up to 20,000.
- The purpose of text visualization is to transform this vector space and reduce its dimensionality so that it could be displayed in 2 (planar) or 3 (spatial) physical dimensions, plus a number of perceptual dimensions, using colour, density, texture, transparency, etc.



A number of mathematical and computational methods can assist in this process, e.g.

- Principal component analysis (PCA)
- Singular value decomposition (SVD)
- Multi-Dimensional scaling (MDS)
- Correspondence analysis (CA)
- Non-linear dimensionality reduction (NLDR)
- Manifold learning algorithms
- Projection-based dimensionality reduction
- Clustering techniques

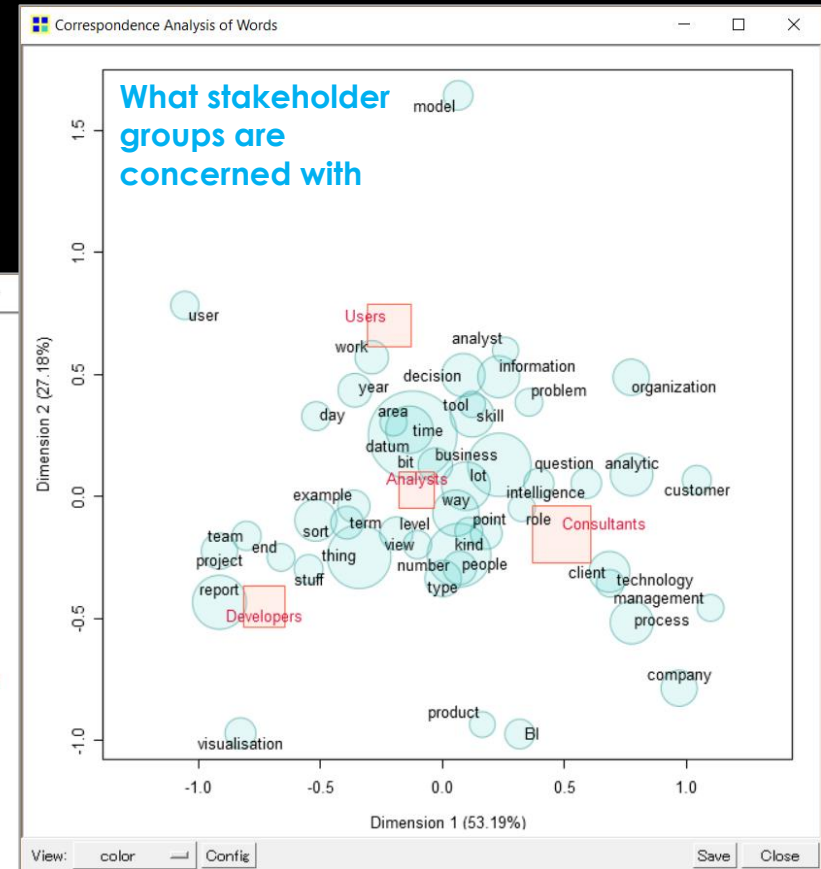
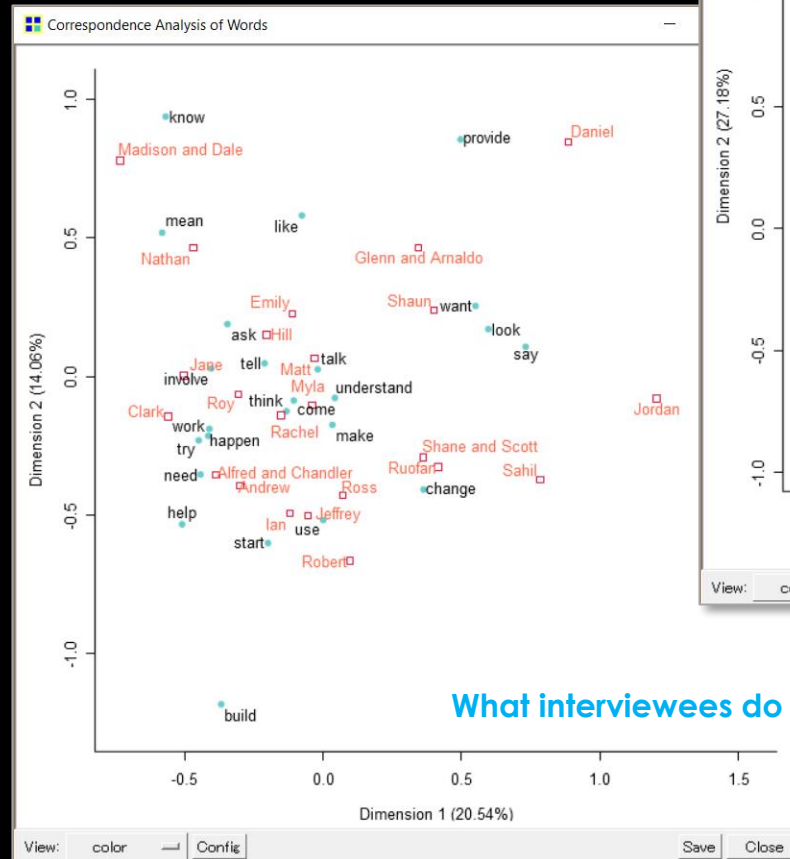
CORRESPONDENCE ANALYSIS

Cross-tabulation of documents vs. term frequency allows calculation of Chi-Squared distance between terms, which can be used in singular value decomposition to identify planar coordinates of terms and documents.

Spatial proximity of terms in 2D provides an intuition of their similarity.

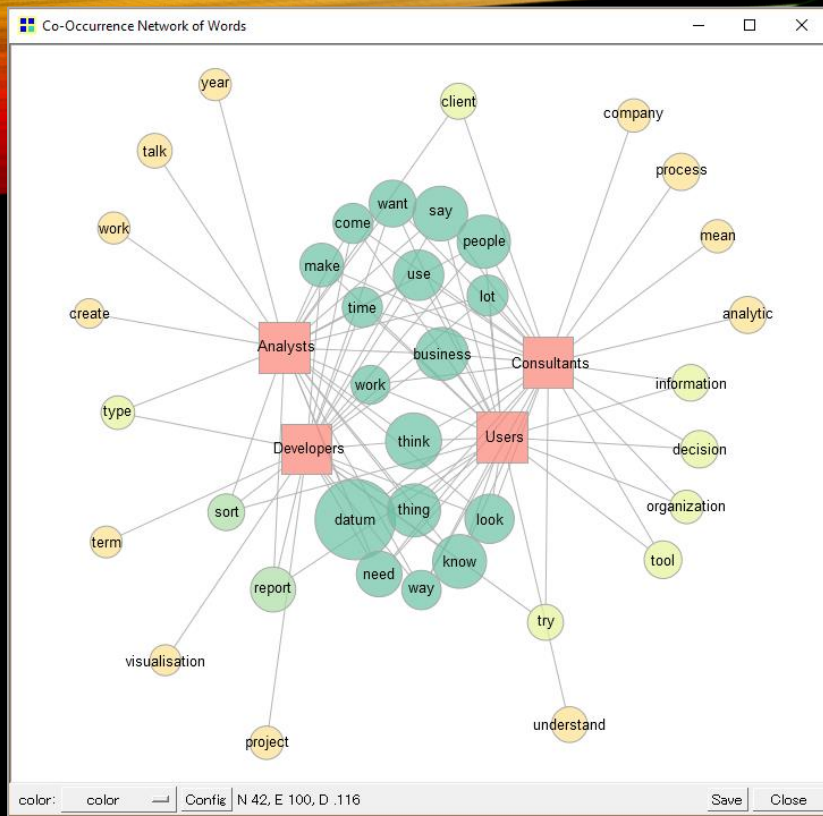
Clusters of spatially close terms give an intuition of terms forming larger topics.

Note that 2D projection simplifies term relationships and its quality is assessed as a % of the total inertia.



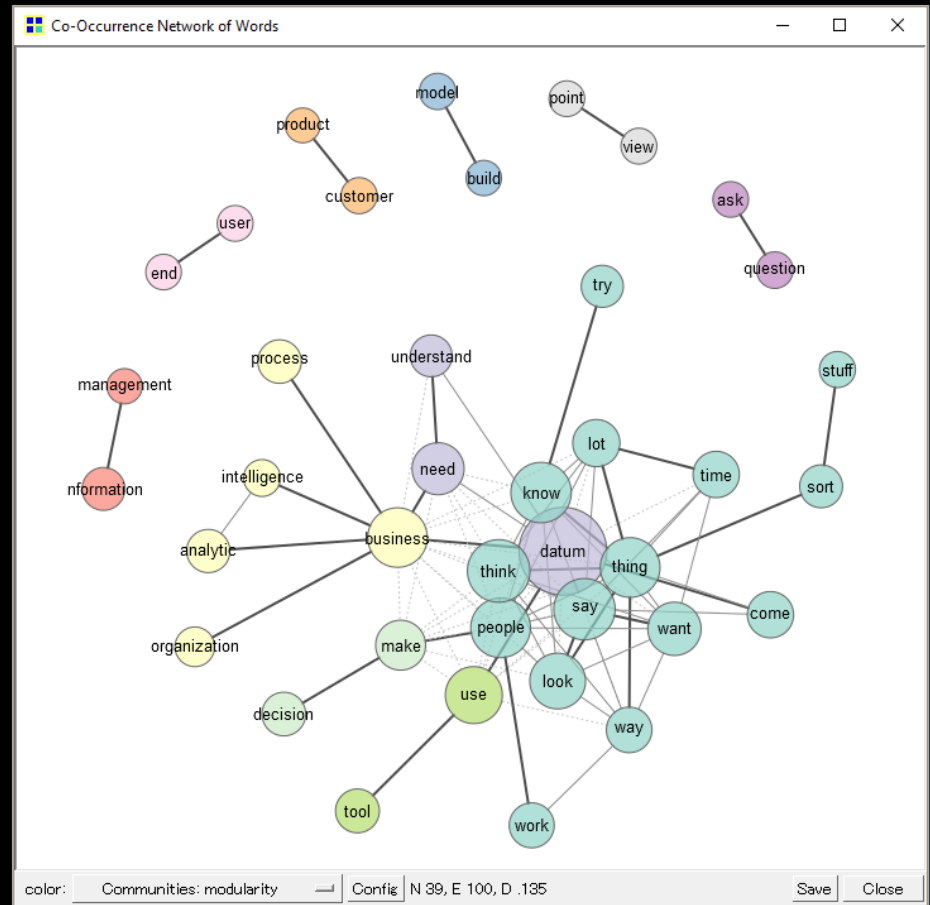
The same approach can be used to assess similarity of terms (black), as well as, documents or their groups (red).

TERM CO-OCCURRENCE



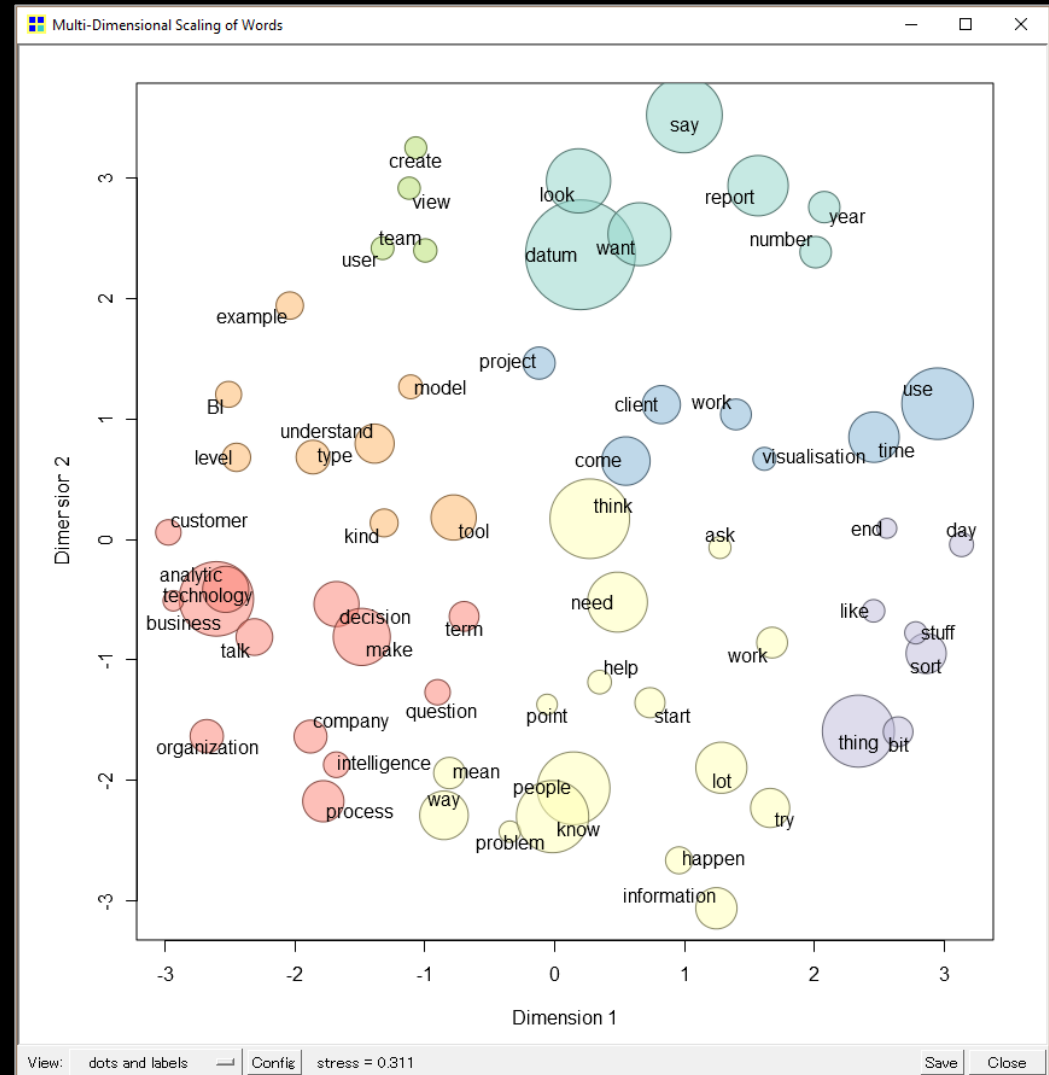
Term co-occurrence in a sentence or paragraph, provides a unique insight into their direct relationship and creates a visual narrative capturing in a very succinct form pages and pages of text.

This chart is produced by using a method of force directed graph drawing, which defines attraction and repelling forces between data points, which aim to iteratively reposition the points to minimize the network energy.



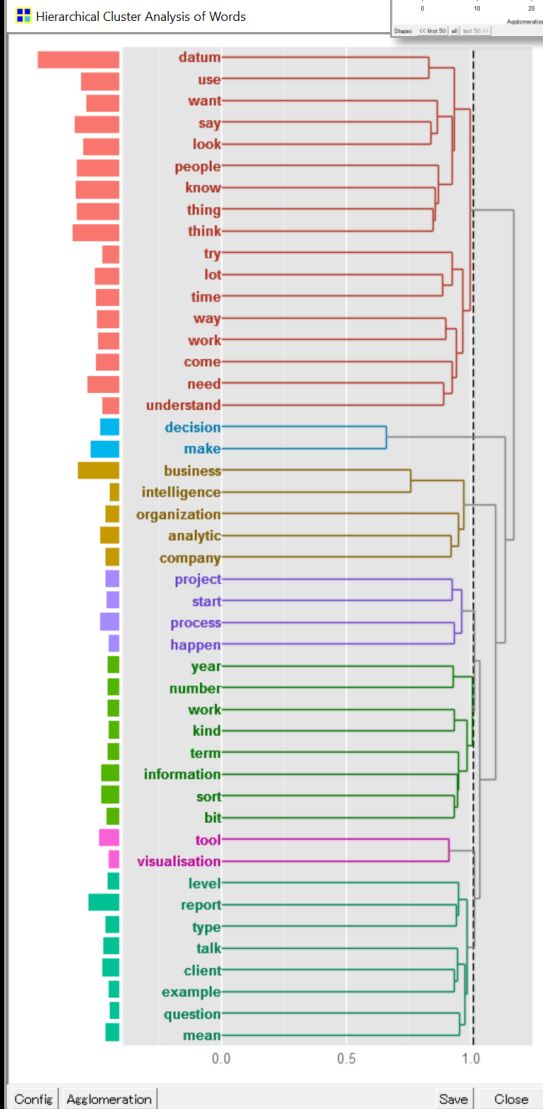
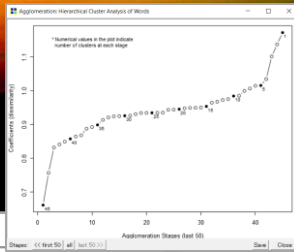
MULTIDIMENSIONAL SCALING

- Unlike correspondence and co-occurrence analysis, multidimensional scaling aims to reduce dimensionality of geometric spaces.
- The method uses an algorithm that tries to re-arrange all points in multi-dimensional space with an aim to fit (squeeze) them into a required number of dimensions.
- The distances between all points measure their similarity or dissimilarity and is given using some well-known metrics (e.g. Euclidean, Cosine or Jaccard).
- The best configuration is trying to approximate the original distances between data points in such a way as to minimize a “stress” function (e.g. metric, Kruskal’s non-metric or Sammon mapping method).
- If the target dimension is 1, 2 or 3, we can visualize the resulting scaled-down space.
- Data clustering could be used for chart colouring.

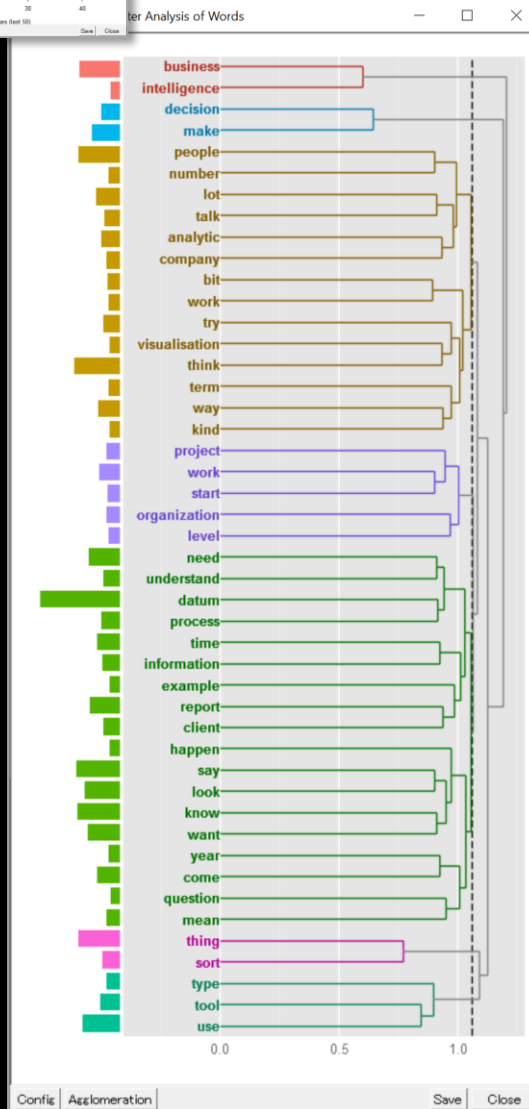


Clustering Effectiveness

Clustering 1 using Jaccard measures



Clustering 2 using Cosine measures



CLUSTER ANALYSIS

Data clustering can be used very effectively to:

- reduce data dimensionality, and
- understand relationships between terms and/or documents

There are many different approaches to creating term (or document) clusters, which result in hierarchical or flat group structures.

The advantage of hierarchical clusters is that they provide a navigable visual structure of the subject domain.

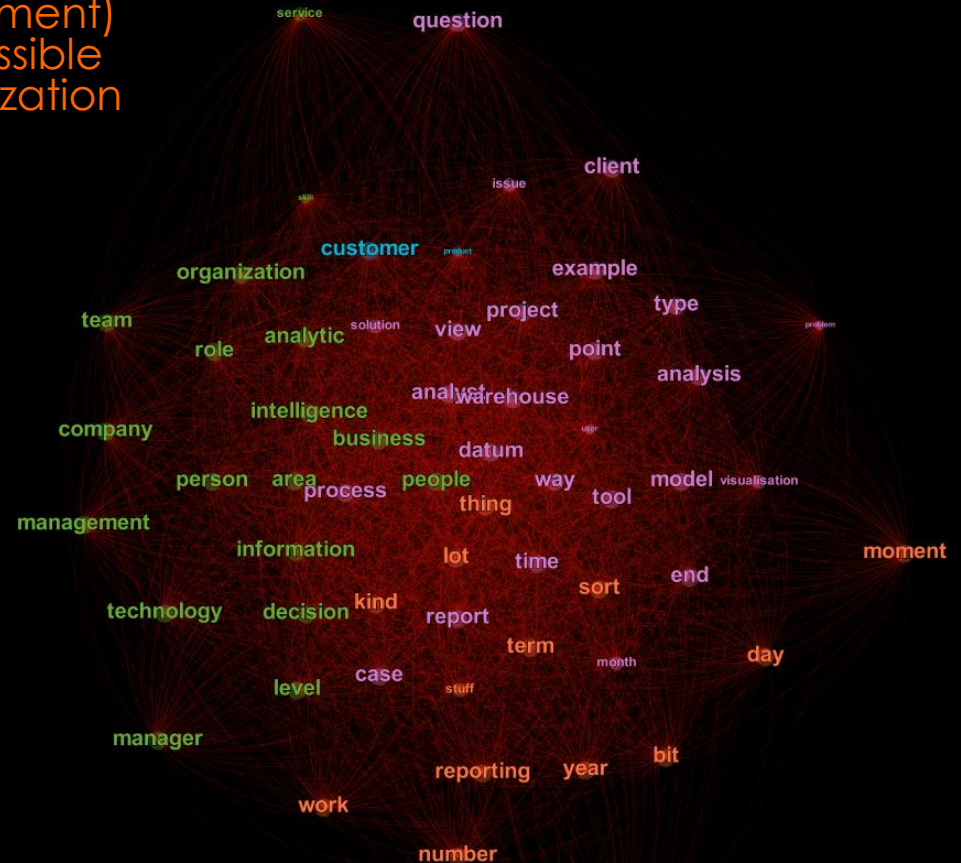
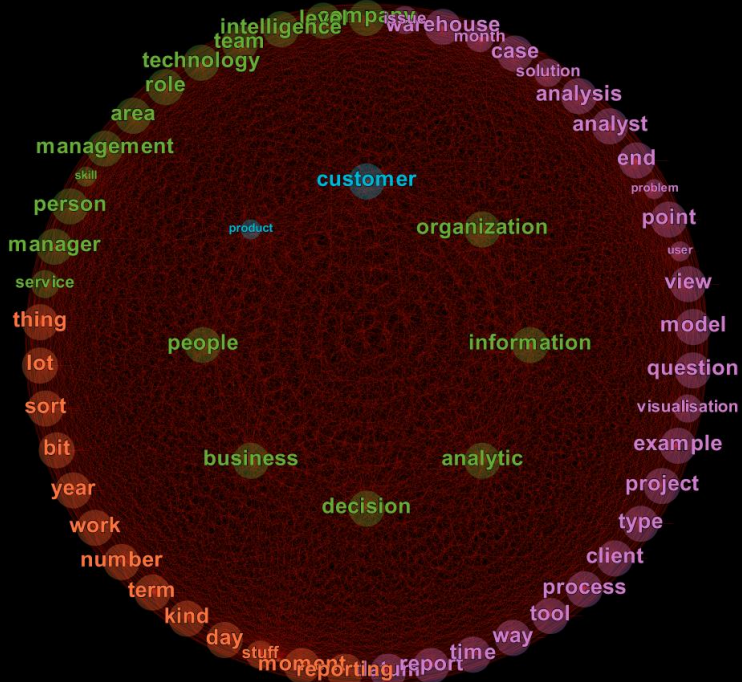
The disadvantage of all clustering methods is that they are sensitive to clustering parameters, such as the metric in use or the target number of clusters.

GEPHI VISUALISATION

When a matrix of term-to-term (or document) measures have been calculated it is possible to use it within a specialized data visualization software, such as Gephi.

Gephi provides functionality to apply multiple layouts to networks of nodes inter-linked with edges.

Nodes and edges can be displayed with numerous visual attributes.



The following charts have been produced by applying a Circular (left) and ForceAtlas (above) layouts. Each layout provides a set of parameters to alter the chart construction and its appearance.

ANSWER & REFLECTION

How can data mining and data visualization assist analysis of interview transcripts?

- Analysis of interview transcripts relies primarily on data mining techniques, such as:
 - text parsing,
 - identification of terms of interests,
 - vector representation of documents, and
 - analysis of relationships between documents and terms.
- Visualisation of results is based on various mathematical methods of dimensionality reduction.
- It assists understanding of vast amount of text data in succinct form.
- However, it is important to be aware that visualization simplifies text data and in the processes may omit or even falsify some data.

QUESTION

How can data mining and data visualization assist analysis of text fields stored in a database of structured records?

RapidMiner Studio
SAS Enterprise Miner

CASE 1: WIKILEAKS AFGHAN WARS

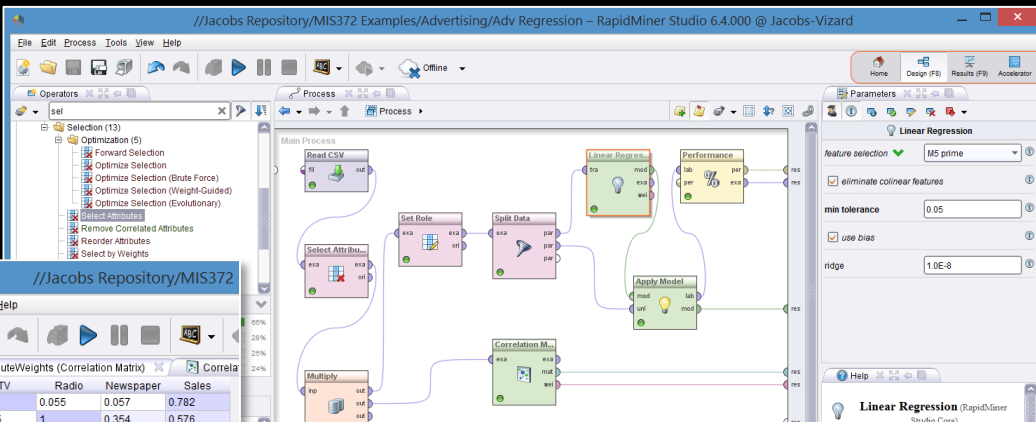
- On 25th July 2010 WikiLeaks released over 91,000 reports called the *Afghan War Diary*, covering the war in Afghanistan from 2004 to 2010
- The reported data describes US military actions, intelligence information, meetings with political figures, and other details
- We will use a subset of these documents (75,000) to predict likelihood of civilian casualties from events' textual description, whether completed and reported or just planned
- Wikileaks documents are in the open data domain. They have been redacted and contain no sensitive information
- The first step in this endeavour is to process textual descriptions in each of the Wikileaks records and understand what the body of all documents describe.

Report key: C592135C-1BFF-4AEC-B469-0A495FDA78D9
Date: Jan 1, 2004 12:00:00 AM
Type: Friendly Action
Category: Cache Found/Cleared
Tracking number: 2007-033-004738-0185
Title: CACHE FOUND/CLEARED Other
Summary: USSF FINDS CACHE IN VILLAGE OF WALU TANGAY: USSF CONDUCTED A MEET AND GREET IN THE VILLAGE OF WALU TANGAY. USSF MEMBERS WERE APPROACHED BY A LOCAL BOY WHO SPOKE OF A CACHE IN A CAVE ON A NEARBY HILL. USSF MEMBERS INVESTIGATED AND FOUND A CACHE CONSISTING OF THIRTEEN 82MM MORTAR ROUNDS, SIXTY RPG ROUNDS, FIFTEEN BOXES 12.7X108MM AMMO (85 ROUNDS PER BOX), FIVE BOXES NON-DISINTEGRATING 12.7X108MM LINK, AND ONE DSHK BARREL LOCATED IN A CAVE AT 350107.26N 0705513.00E. USSF CONFISCATED THE AMMO. THE REST WAS BLOWN IN PLACE.
Region: RC EAST
Attack on: FRIEND
Complex attack: FALSE
Reporting unit: OTHER
Unit name: OTHER
Type of unit: Coalition
Friendly WIA: 0
Friendly KIA: 0
Host nation WIA: 0
Host nation KIA: 0
Civilian WIA: 0
Civilian KIA: 0
Enemy WIA: 0
Enemy KIA: 0
Enemy detained: 0
MGRS: 42SXD7520076792
Latitude: 35.01860809
Longitude: 70.92027283
Originator group: UNKNOWN
Updated by group: UNKNOWN
Affiliation: FRIEND
D color: BLUE
Classification: SECRET

Sample record
from the Afghan
Wars document
corpus

RAPIDMINER STUDIO

Process



Correlation

AttributesWeights (Correlation Matrix)

Attributes	TV	Radio	Newspaper	Sales
TV	1	0.055	0.057	0.782
Radio	0.055	1	0.354	0.576
Newspaper	0.057	0.354	1	0.228
Sales	0.782	0.576	0.228	1

Performance

Performance

Description

Annotation

PerformanceVector

PerformanceVector:
 root_mean_squared_error: 1.783 +/- 0.000
 squared_correlation: 0.887

Prediction

ExampleSet (100 examples, 3 special attributes, 3 regular attributes)

Row No.	Id	Sales	prediction(S...	TV	Radio	Newspaper
1	1	22.100	20.407	230.100	37.800	69.200
2	2	10.400	12.115	44.500	39.300	45.100
3	5	12.900	13.428	180.800	10.800	58.400
4	7	11.800	11.581	57.500	32.800	23.500
5	8	13.200	12.169	120.200	19.600	11.600
6	9	4.800	3.976	8.600	2.100	1
7	11	8.600	7.271	66.100	5.800	24.200
8	12	214.700	24	4		
9	13	195.400	47.700	52.900		
10	14	67.800	36.600	114		
11	15	147.300	23.900	19.100		
12	16	62.300	12.600	18.300		
13	17	240.100	18.700	22.000		

RapidMiner

LinearRegression

Description

Annotation

0.046 * TV
 + 0.174 * Radio
 + 3.213

Regression Formula

AttributeWeights (Correlation Matrix)

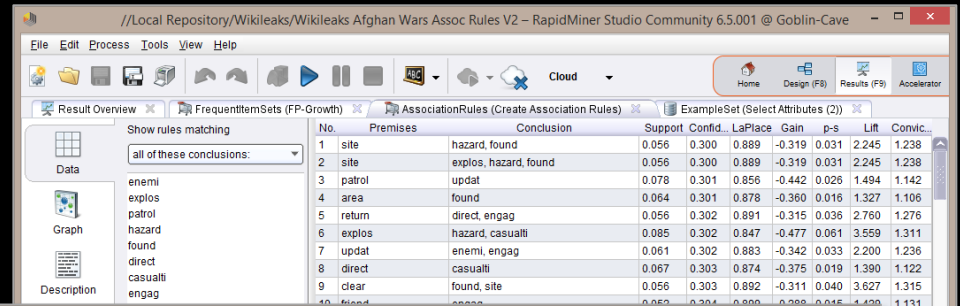
Attribute	Coefficient	Std. Error	Std. Coeffi...	Tolerance	t-Stat	p-Value	Code
TV	0.046	0.002	0.789	0.999	25.546	0	****
Radio	0.174	0.011	0.500	0.999	16.191	0	****
(Intercept)	3.213	0.370	?	?	8.685	0	****

Coefficients

DOMAIN EXPLORATION - ASSOCIATION RULES

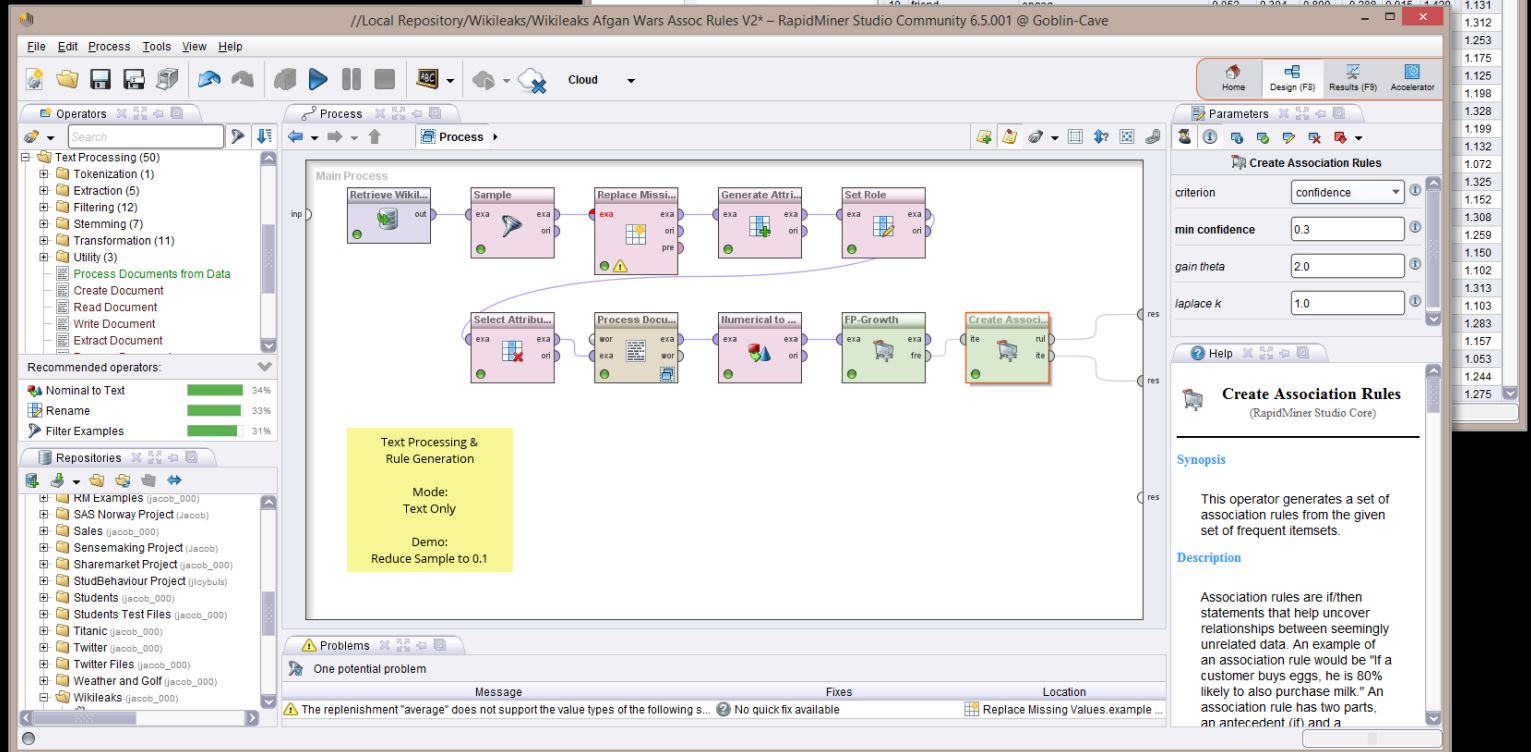
RapidMiner has some great tools for text analysis. However, it has no tools for text visualization.

Instead we will use association rules (commonly used in Market Basket Analysis) to identify co-occurring terms and generate associations between these terms. We will visualize these associations to gain new domain insights.



The screenshot shows the 'AssociationRules (Create Association Rules)' operator output in RapidMiner Studio. The table displays various metrics for different association rules, including premises, conclusions, support, confidence, lift, and gain.

No.	Premises	Conclusion	Support	Confid.	LaPlace	Gain	p-s	Lift	Conv...
1	site	hazard, found	0.056	0.300	0.889	-0.319	0.031	2.245	1.238
2	site	explos, hazard, found	0.056	0.300	0.889	-0.319	0.031	2.245	1.238
3	patrol	updat	0.078	0.301	0.856	-0.442	0.026	1.494	1.142
4	area	found	0.064	0.301	0.878	-0.360	0.016	1.327	1.106
5	return	direct, engag	0.056	0.302	0.891	-0.315	0.036	2.760	1.276
6	explos	hazard, casualti	0.085	0.302	0.847	-0.477	0.061	3.559	1.311
7	updat	enemi, engag	0.061	0.302	0.883	-0.342	0.033	2.200	1.236
8	direct	casualti	0.067	0.303	0.874	-0.375	0.019	1.390	1.122
9	clear	found, site	0.056	0.303	0.892	-0.311	0.040	3.627	1.315



The screenshot shows the main workflow in RapidMiner Studio. The process starts with 'Retrieve Wikil...', followed by 'Sample', 'Replace Missing Values', 'Generate Attributes', and 'Set Role'. The data then flows through 'Select Attributes', 'Process Documents from Text', 'Numerical to Text', 'FP-Growth', and finally 'Create Association Rules'. A yellow box highlights the 'Text Processing & Rule Generation' section with the following details:

- Mode: Text Only
- Demo: Reduce Sample to 0.1

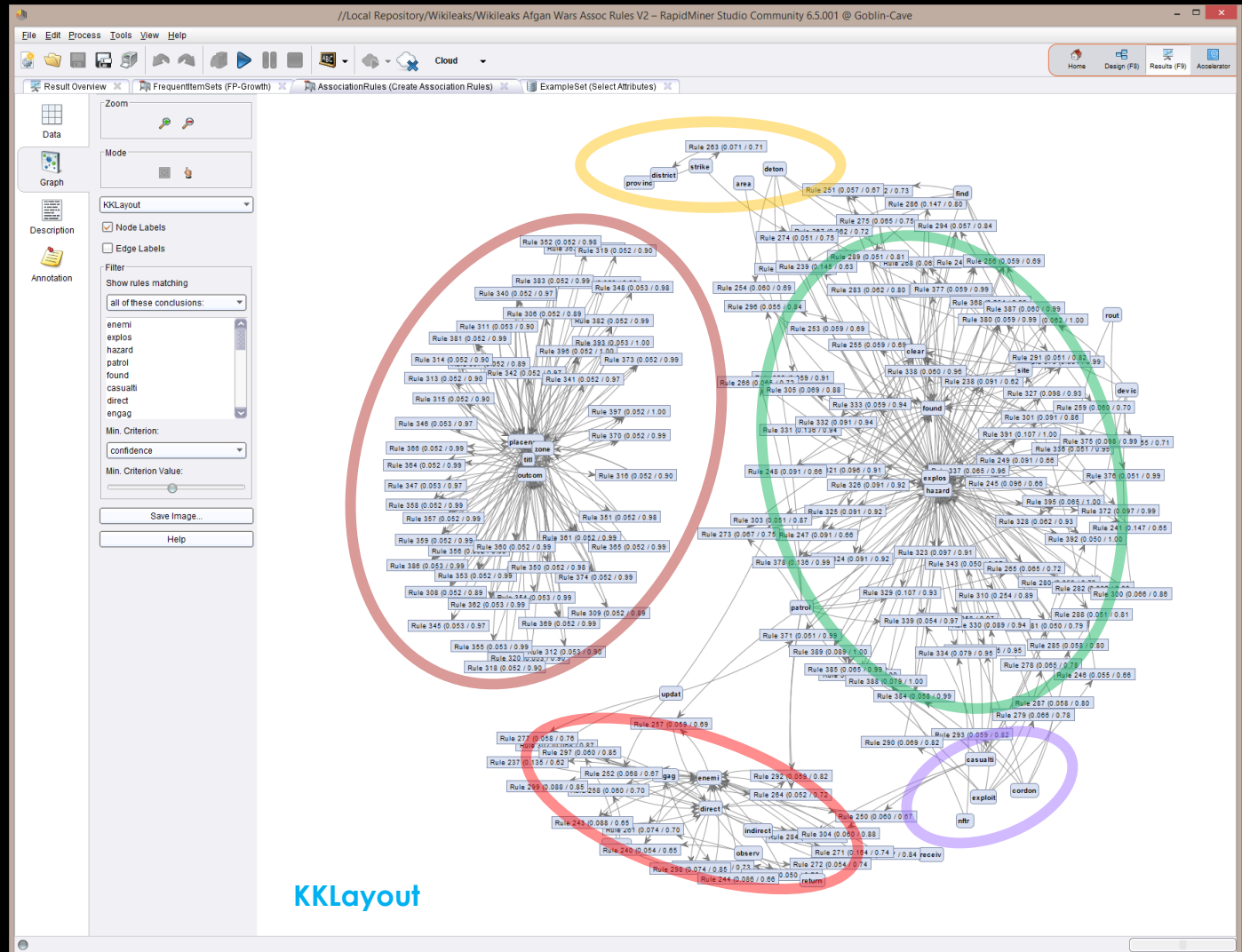
The 'Create Association Rules' operator parameters are shown on the right:

- criterion: confidence
- min confidence: 0.3
- gain theta: 2.0
- laplace k: 1.0

The 'Synopsis' section states: "This operator generates a set of association rules from the given set of frequent itemsets." The 'Description' section explains: "Association rules are if/then statements that help uncover relationships between seemingly unrelated data. An example of an association rule would be 'If a customer buys eggs, he is 80% likely to also purchase milk.' An association rule has two parts, an antecedent (if) and a consequent (then)." A warning message at the bottom indicates: "The replenishment 'average' does not support the value types of the following s... No quick fix available" with a location pointing to 'Replace Missing Values example...'.

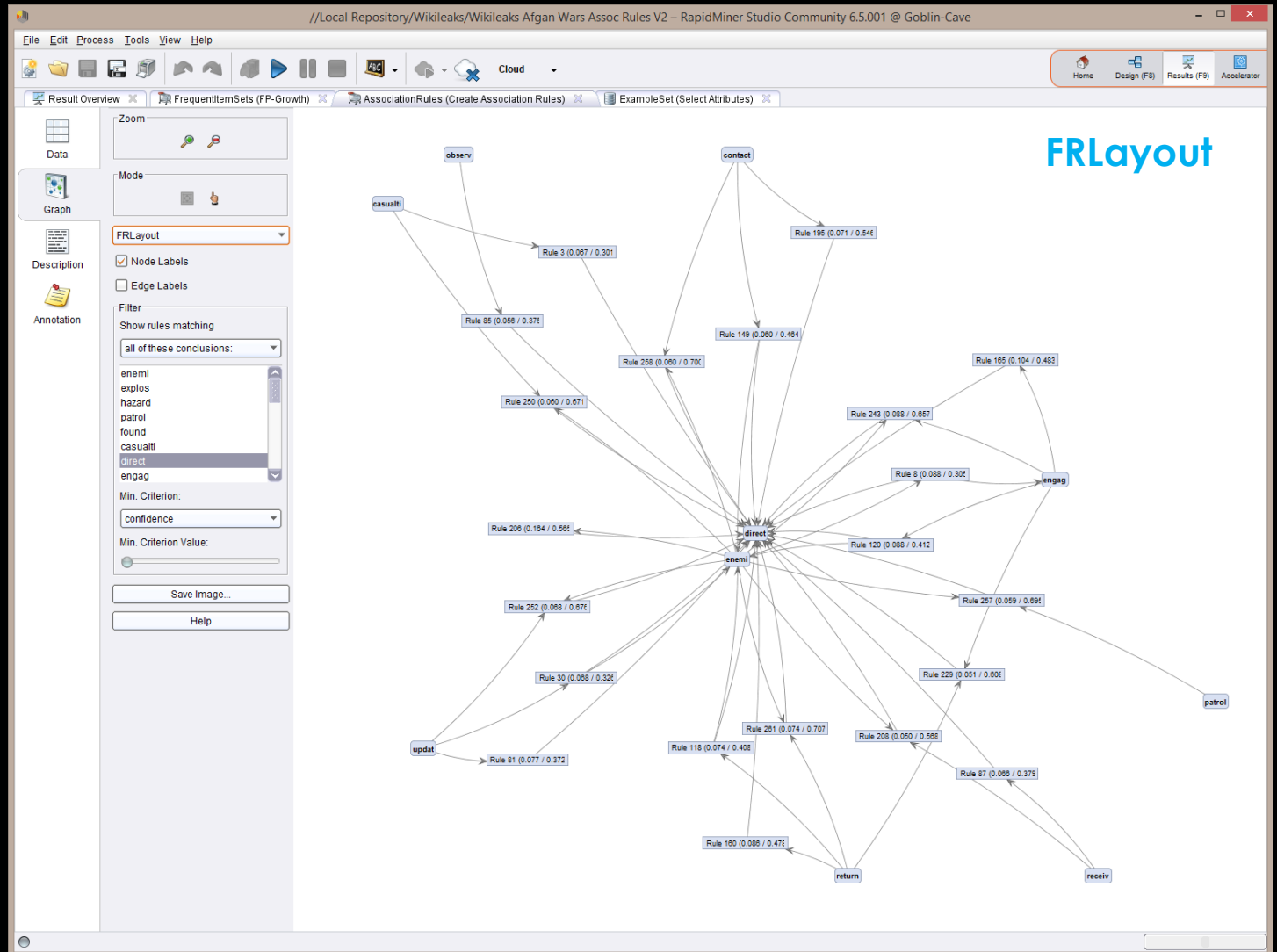
RECONSTRUCTING EVENTS

Topics: planning with district officials (yellow), preparation of the area (blue), engagement with the enemy (red), clearing explosives (green), dealing with casualties (purple).



DRILLING INTO DETAILS

Exploring finer details: direct contact with the enemy



We can further extend the analytic process to create a predictive model of civilian casualties.

CASE 2: WORKERS COMPENSATION CLAIMS

- This case aims at predicting the possibility of a legal case to recover money (subrogation) from negligent employers when workers suffered injury at work
- A data source of previous workers compensation claims have been provided
- Part of the data is structured and part of it is in text , i.e. unstructured form
- Both types of data need to be used when creating a predictive model
- A file of new claims has also been given for scoring

Obs #	Claim Number	Adjustor Notes	Body Part	Nature of Injury	Cause of Injury	Vehicle Flag...	Subrogation...	Fraud Flag ...
	1000004487308	Strained neck trying to catch falling product.	Neck	Sprain/Strain	Slip/Fall	0	1	0
	2000309831108	Fingers caught in machine.	Finger	Contusion	Caught in Machine	0	0	0
	3001301185908	Claimant caught left hand between two machine sound enclosu...	Hand	Laceration	Equipment/Machi...	0	0	0
	4001716965808	Claimant states that while he and coworker were driving a deliv...	Multiple	Contusion	Struck Object	1	1	0
	5001924817308	Smashed right second finger, was using a drill press and smas...	Finger	Contusion	Struck Object	0	0	0
	6002500385808	Claimant alleges that he injured his right knee. Three weeks sin...	Knee	Sprain/Strain	Unknown	0	1	0
	7002525865808	Left ankle pain due to getting in and out of a truck repeatedly.	Ankle	Repetitive Motion	Repetitive Motion	1	0	0
	8002601381908	While trying to avoid hitting a car out of control, came to comple...	Neck	Contusion	MVA	1	1	0
	9002613478908	Fell in blast freezer, injured back and side.	Back	Contusion	Struck Object	0	0	0
	10002614936508	Employee was struck by automobile --- contusion to knee.	Knee	Contusion	MVA	1	1	0
	11002701592908	Employee alleges while letting a machine down into basement ...	Shoulder	Contusion	Struck Object	0	1	0
	12002714742208	Employee failed to yield and was hit by an oncoming vehicle.	Multiple	Contusion	MVA	1	1	0
	13002829018508	Claimant states he was loading a patio door onto a truck and he...	Knee	Sprain/Strain	Lifting	1	0	0
	14004026028708	Right ring finger laceration, transporting patient to hospital, takin...	Finger	Laceration	Struck Object	0	0	0
	15004726761108	Cart lurched to right and grabbed copier to keep it from falling a...	Hand	Fracture	Slip/Fall	0	0	0
	16004819819208	Employee states he was in his vehicle and was hit on the driver...	Back	Sprain/Strain	MVA	1	1	0
	17004926810208	Claimant rearended by another vehicle.	Neck	Sprain/Strain	MVA	1	1	0
	18005822830908	Claimant states she slipped on concrete floor due to unknown g...	Back	Sprain/Strain	Slip/Fall	0	0	0
	19006503221708	Operating bender, felt twinge in arm.	Arm	Sprain/Strain	Contact with Obj...	0	0	0
	20006928514708	Employee was lifting 200 pound plate at work and has mid back...	Back	Sprain/Strain	Lifting	0	1	0
	21007207609408	Employee was involved in a MVA.	Multiple	Contusion	MVA	1	1	0
	22007408194208	Vehicle was rearended while stopped at red traffic light.	Neck	Sprain/Strain	MVA	1	1	0
	23007429454808	Employee alleges from heavy typing, filing and phones a repetiti...	Wrist	Repetitive Motion	Repetitive Motion	0	1	0
	24007907289608	After employee stepped over pallet her right heel hurt.	Foot	Sprain/Strain	Unusual Body Mo...	0	1	0
	25008503290208	Neck strain, hit as backing out of parking space.	Back	Sprain/Strain	MVA	1	1	0
	26008513773908	Slipped on ladder and cut right shin.	Leg	Laceration	Struck Object	0	1	0
	27008903914208	Was in a car accident.	Multiple	Contusion	Struck Object	0	1	0
	28009602377208	Machine closed on left thumb.	Finger	Contusion	Caught in Machine	0	0	0
	29010009593908	While driving to a customer site hit by another vehicle.	Back	Contusion	MVA	1	1	0
	30010103504608	Was grinding carrots finger abrasion.	Finger	Abrasion	Struck Object	0	0	0
	31010410050208	Riding in city vehicle/rearended by a van.	Neck	Sprain/Strain	MVA	1	1	0

Sample records from the Workers
Compensation Claims

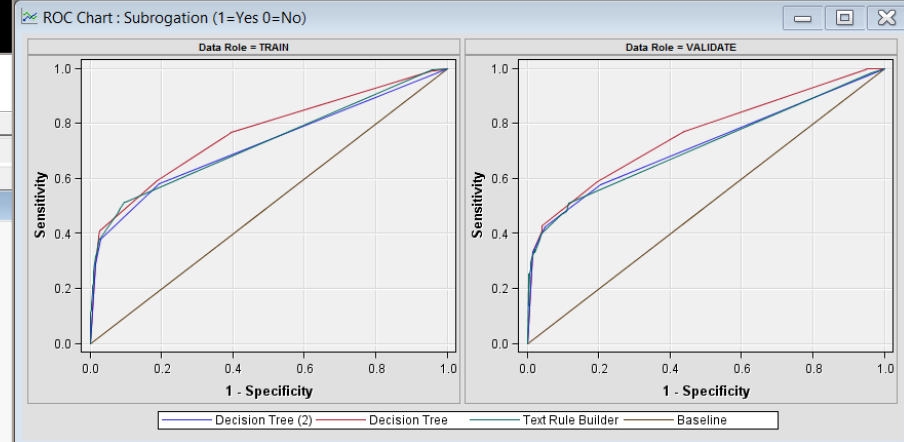
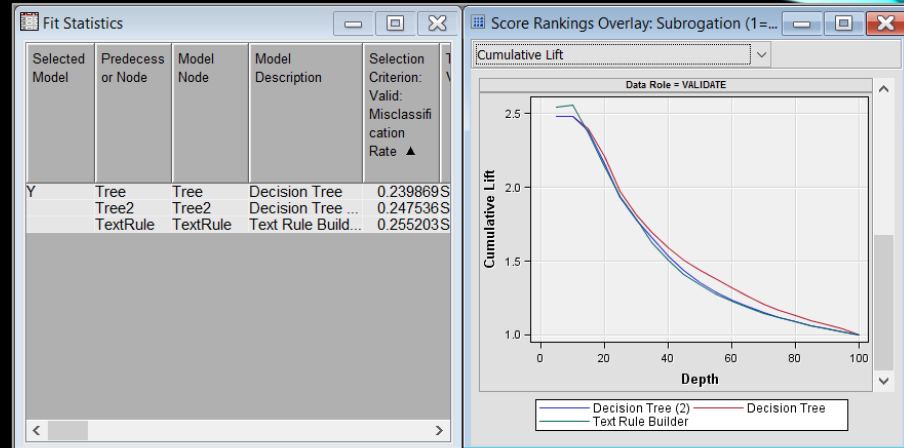
MODEL DEVELOPMENT AND EVALUATION

Data Role=VALIDATE Target Variable=SubroFlag Target Label=Subrogation (1=Yes 0=No)

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage	False Neg
0	0	75.3210	92.3077	528	57.8313	
1	0	24.6790	50.7331	173	18.9485	
0	1	20.7547	7.6923	44	4.8193	
1	1	79.2453	49.2669	168	18.4009	Text

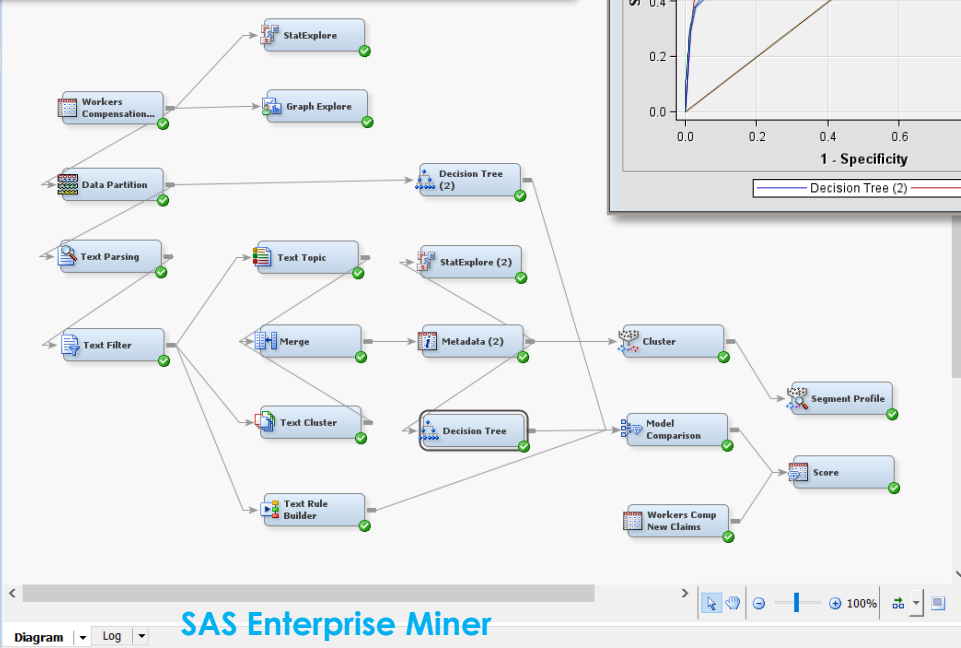
Data Role=VALIDATE Target Variable=SubroFlag Target Label=Subrogation (1=Yes 0=No)

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage	False Neg
0	0	73.3784	94.9301	543	59.4743	
1	0	26.6216	57.7713	197	21.5772	
0	1	16.7630	5.0699	29	3.1763	
1	1	83.2370	42.2287	144	15.7722	No Text



Enterprise Miner interface showing Property Value table and General properties.

Property	Value
General	
Node ID	Tree
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Leaf Size	5
Number of Rules	5
General	
General Properties	



- Process text to create new variables derived from document terms.
- Use these new variables to create a number of classification models for workers compensation claims.
- Compare the models based on their validation results.

SAS Enterprise Miner

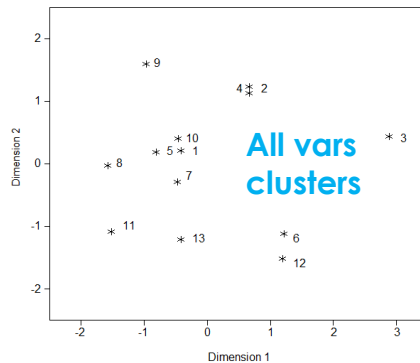
Run completed

© jacob_000 as jacob_000 Connected to Jacobs-Vizard

Descriptive Terms

- 1+vehicle +hand back lifting +strain
- 1walking unloading +forklift neck truck
- 2lifting +pain +hand +machine +door
- 2+vehicle accident truck car +ankle
- 3fell +back back +strain lifting
- 4fell +back back +strain lifting
- 5fell +back back +strain lifting
- 7lifting +machine +arm +ladder +pain
- 7+knee +foot +shoulder +hand fell
- 8+machine back +ladder +finger +door
- 8+arm +eye +leg felt forearm
- 13+arm +leg head accident auto
- 13+ey

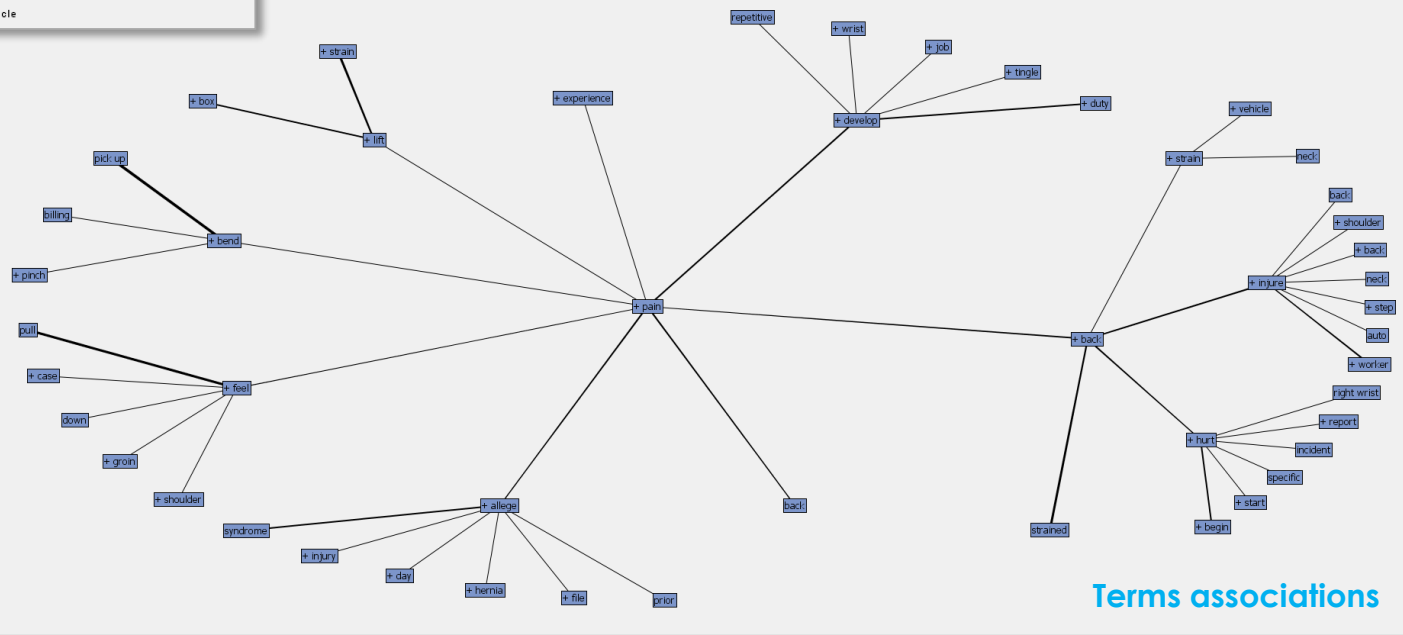
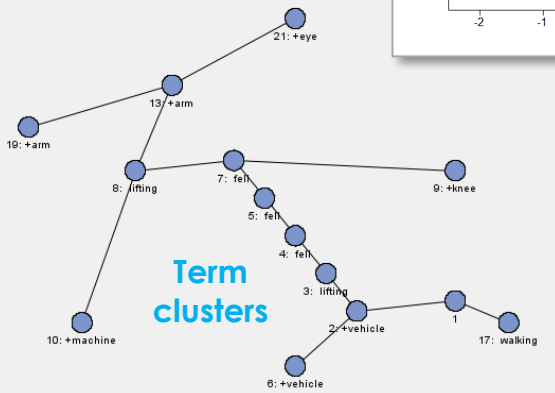
Cluster Proximities



ANALYSE TEXT ASSOCIATIONS & CLUSTERS

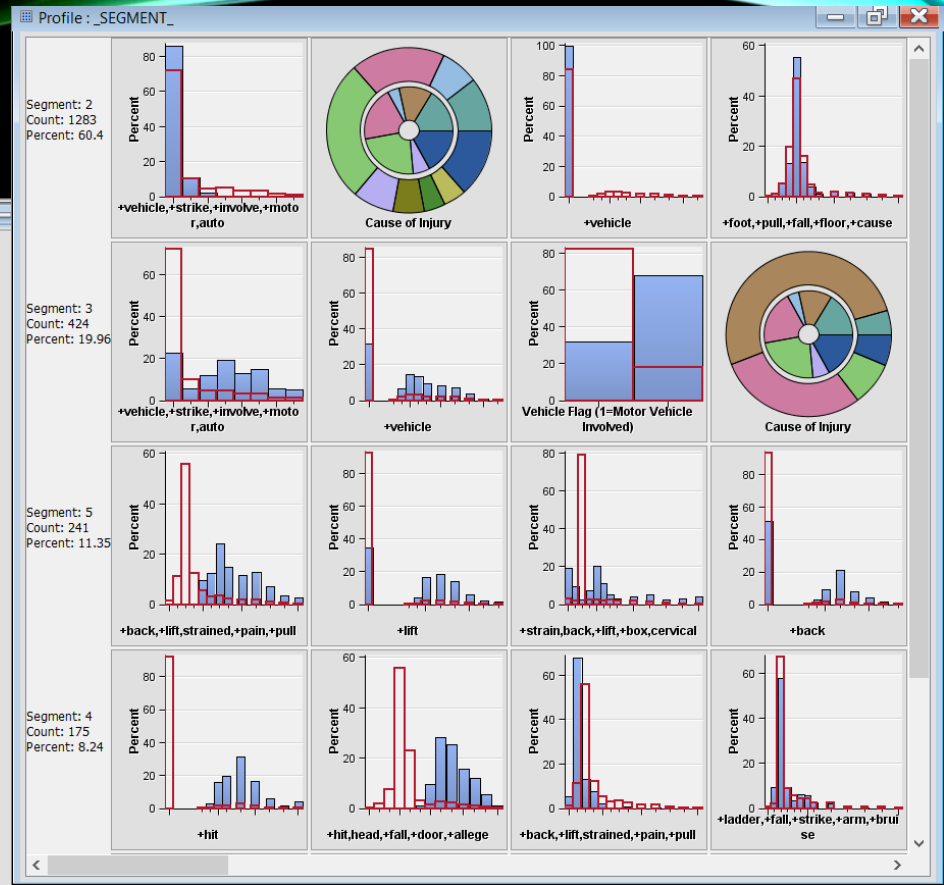
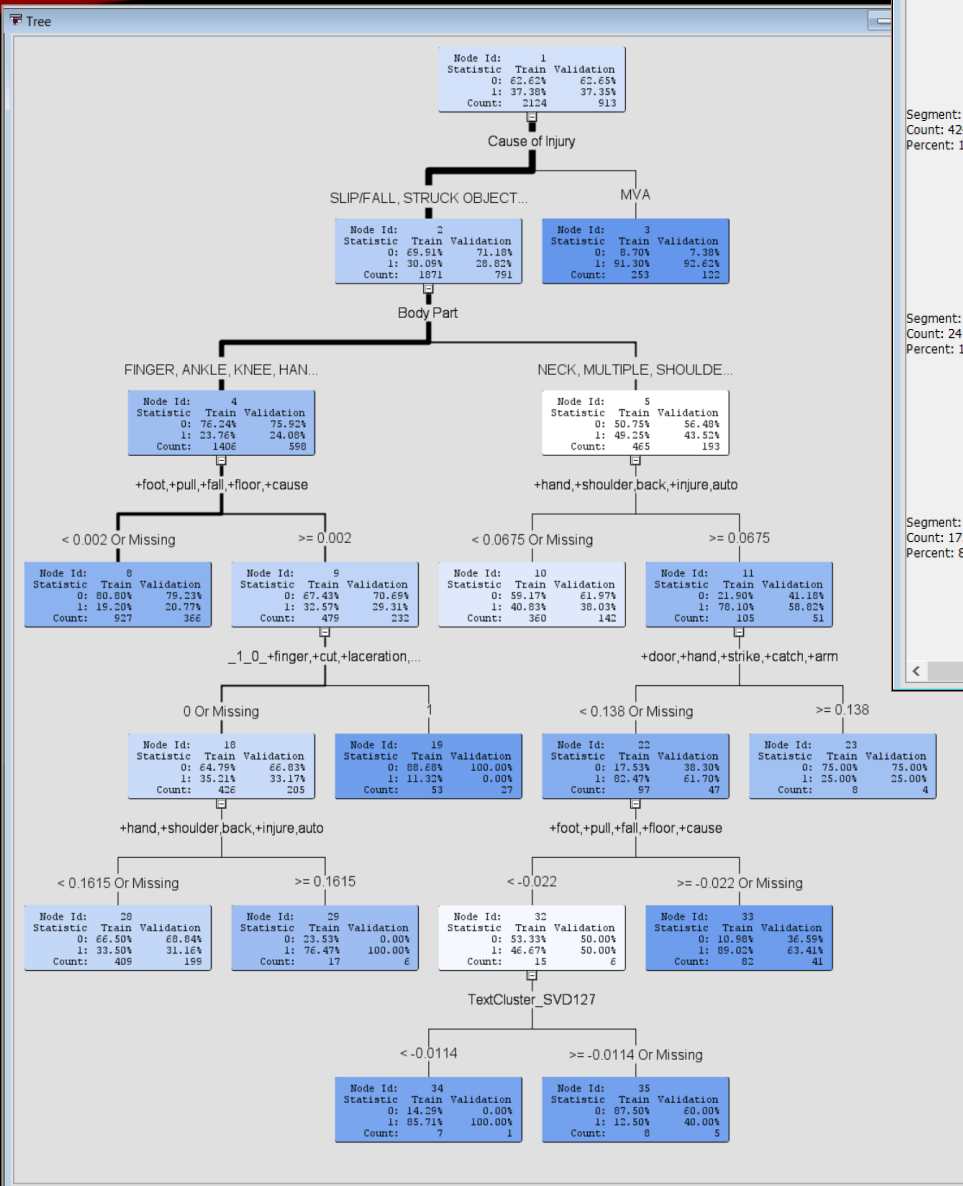
We can now explore text clusters, analyse term associations, inspect text topics, evaluate a model based on structured and text data, and use it in prediction.

Term clusters



Terms associations

IMPACT OF TEXT VARIABLES



- As can be seen from the previously shown results, the best model was a decision tree using a combination of structured data and text.
- Data clustering and segmentation identify three variables that are the most important in isolating two most important clusters (also prediction target) are in fact all derived from text.

ANSWER & REFLECTION

How can data mining and data visualization assist analysis of text fields stored in a database of structured records?

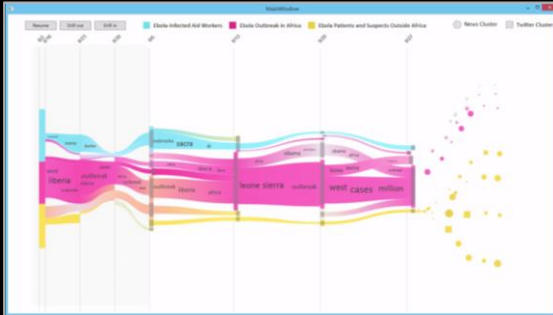
- Analysis of text accompanying structured information (stored in databases) aims primarily to:
 - convert text into variables derived from terms,
 - create models using a combination of all variables,
 - visualise these models with a view to understand data, and
 - use the created models for other analytic tasks.
- As text is a very rich medium, models created with text often perform better than those relying on structured data only.
- Visualisation of textual data is used commonly to improve the creation of new text-based variables.
- Visualization of data that incorporates text and structured information assist improvement of predictive models, as well as, understanding of the subject domain and prediction results.

Analysis of dynamic text in complex social contexts

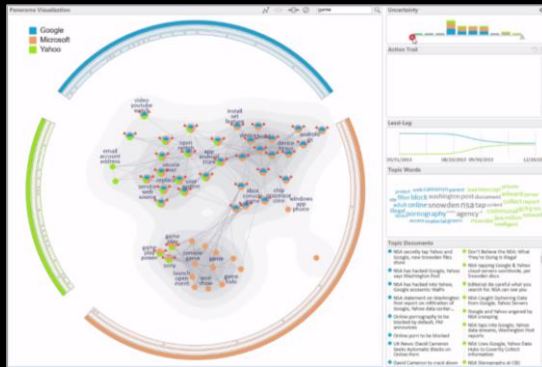
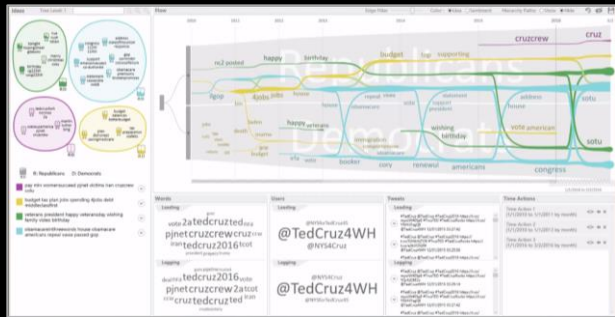
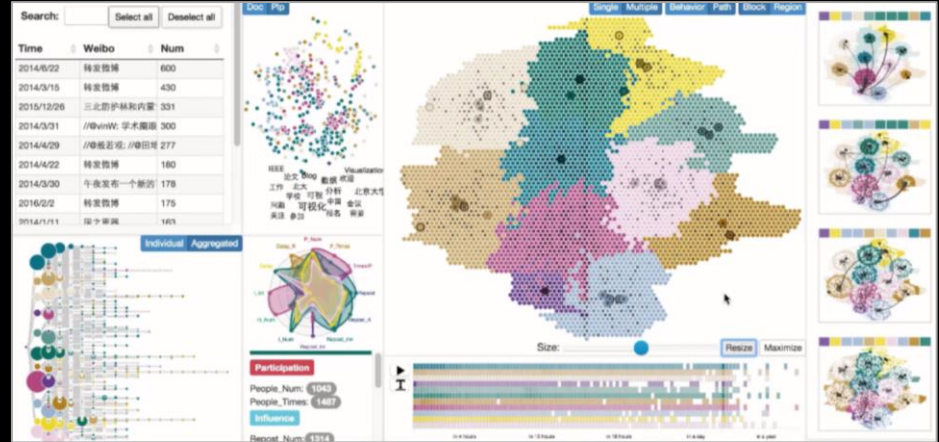
FUTURE TRENDS

PROJECTIONS FROM IEEE VIS'2016

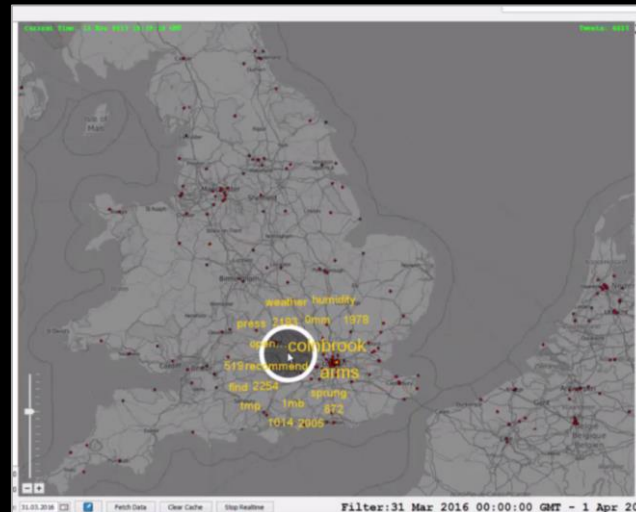
Text streaming and flowing (e.g. Twitter) by Shixia Liu and students, Tsinghua University



Information diffusion and propagation by Siming Chen et al, Peking University



Topic panoramas by Shixia Liu and students, Tsinghua University



Documents Compass by Florian Heimerl, Markus John, Qi Han and Steffen Koch Universität Stuttgart

SUMMARY & REFLECTION

The main purpose of text visualisation is to make sense of the domain of discourse the text represents.

The first tasks in the process of text visualization is in its parsing and representation in a structured form – usually as a space of term vectors.

The main problem of this representation is in its very high dimensionality (often over 20,000).

To make this information useful for both visualization and analytics various dimension reduction techniques are applied.

For the purpose of text visualization the most commonly used methods include correspondence analysis, multi-dimensional scaling, co-occurrence analysis, as well as, cluster analysis.

There are many algorithmic methods of data visualization, such as force-directed graphs, which are based on iterative optimization techniques.

Data visualization requires a very significant effort in understanding and representing data, developing analytic solutions and then creating a visual form.

There are many deep theoretical questions related to data (including text) visualization, some related to data representation, some to mathematical methods, others in the area of human cognition, which need to be pursued.

Recent projects

Teaching Data Analytics (OLT)
Sensemaking and Legitimation (ICAA)
Larger study of IVA sensemaking
Study comparing 2D and 3D IVA
Collaborative analytics

Current and future work

Virtual reality analytics
Dynamic data and text

References

Cybulski, J.L., Keller, S., Nguyen, L. and Saundage, D. (2015). Creative Problem Solving in Digital Space Using Visual Analytics. *Computers in Human Behavior* 42: 20–35.

Cybulski, J.L., Keller, S. and Saundage, D. (2015). Interactive Exploration of Data with Visual Metaphors. *International Journal of Software Engineering and Knowledge Engineering* 25, no. 02, 231–52.

Visual Analyst 3D with Accidents Data - Gender balance

APPENDIX: INTERACTIVE VISUAL ANALYTICS

IVA serves different purposes for different audiences

