Learning from patient safety incidents: can data mining help?

Candidate identifier: TMPC1

I hearby declare that the work presented in this thesis is my own.

Carl Jonathan Reynolds

Abstract

Contents

Acknowledgments

I am grateful to Sir Liam Donaldson and the National Patient Safety Agency for supporting and encouraging me through the early stages of this work. I owe a large debt to the open source software community for making, and documenting, the tools used in this thesis. Without the open source community this project would have been impossible. Finally, I am truly thankful to Ross Jones, my friend and business partner, who has been a rich source of ideas, guidance, and practical advice.

List of Tables

List of Figures

Part I

Background

"capturing and recording information on adverse events, and analysing them in the right way is an essential step to reducing risk to patients..."

Building a safer NHS for patients, Department of Health Report 2001

Chapter 1

Introduction

:

1.1 The National Patient Safety Agency

The National Patient Safety Agency (NPSA) was formed in 2001 following the publication of two landmark reports by the then Chief Medical Officer, Professor Sir Liam Donaldson. An organization with a memory[1] and Building a safer NHS[2] set out the urgent need for greater organizational learning from safety incidents to make the NHS safer for patients.[3]

1.1.1 Patient safety and incident reporting

Recognising that a 'blame and punish' culture can inhibit reporting of safety incidents the Department of Health and the NPSA placed emphasis on the role of systems, rather than individuals acting alone, in safety incidents. Drawing inspiration from risk management in aviation, where a safety culture and incident reporting are the norm, NHS employees are now actively encouraged to report safety incidents.

What is a safety/adverse incident? A safety incident, also called an adverse incident, is whenever something unexpected happens, or nearly happens, that would threaten patient safety.

How do health care professionals report safety incidents? Healthcare professionals report safety incidents to the National Reporting and Learning System (NRLS) using a standardized incident reporting form. This form, called an IR1 form, is usually paper based but may also be electronic.[4] The form serves as a means for the healthcare professional to inform his or her institution that an incident has, or has almost, occured.

What is contained on an incident report form? Incident reports record the following information:

1. The facts of the incident (including a free text description)
2. The perception of possible consequences (the potential or actual harm)
3. The perception of how the incident came to arise (the causes)

What happens to incident report forms? Incident report forms are reviewed, and investigated further if necessary, by the local clinical governance team. All incident forms are also submitted electronically to the NRLS for national level analysis.

1.1.2 The National Reporting and Learning Service (NRLS)

The National Reporting and Learning Service (NRLS) is the division of the National Patient Safety Agency (NPSA) concerned with the analysis of reports of patient safety incidents and safety information from other sources. On the basis of this analysis the NRLS develops and issues information products to NHS organizations.

Information products include:

- Patient safety alerts, including Rapid Response Reports
- Data reports, including Quarterly Data Summaries
- Guidance, including Seven Steps to Patient Safety guides
- Toolkits
- Signals: emerging issues from national review of serious patient safety incidents

A key premise of the NPSA's work is that incident reporting, the NRLS, and alerts can improve patient safety by allowing learning from incidents and near misses[5]. This model of incident reporting and learning has been successful in other high risk industries such as aviation where with time the number of incidents has increased but the total number of incidents causing death or serious harm has fallen.

Over one million incidents are reported to the NRLS per year, many more than would be humanly possible for the staff to review. Therefore, routine analysis, and learning, is limited in the main to incidents which are reported to have caused serious harm or death. A consequence of this is that there is potentially additional, as yet untapped learning, in the large number of incidents that are not analyzed centrally.

1.2 Data mining

1.2.1 What is data mining?

One can say that Knowledge Discovery in Databases (KDD) is the whole

eld of research of extracting new and interesting knowledge out of (masses of) infor- mation, which in addition might be very unstructured. The term KDD is often used equivalently in one context with the term of DM, w inc techniques

1.2.2 Techniques used in data mining

1.3 Patient safety incidents relating to clinical information systems

1.4 Data mining evaluation criteria

NPFIT, NRLS, Incident reports, descriptive stats on the NRLS data to do with IT.

Background: 1. What the National Reporting and Learning services (NRLS) is 2. What the current state of play with respect to health care information systems in the UK inc NPfIT is 3. It is known that health care information systems can introduce new, potentially dangerous, problems. Difficulties reported for current systems especially

around implementation. 4. Incidents reported to the NRLS relating to health care information systems have not been analyzed in depth before.

?more clearly define problem http://www.dissertationwriting.com/introduction-chapter-writing.shtml

?mention data science as an emerging field

Chapter 2

Review of the literature

2.0.1 Data mining techniques

2.0.2 Data mining techniques in patient safety

2.0.3 Data mining evaluation criteria

Part II

Method

Chapter 3

Data gathering

3.1 The “computer problem” extract

nrls excel sas plugin

Chapter 4

Data cleaning

4.1 Deduplication

refine

4.2 Establishing data quality

brewery

Chapter 5

Data analysis

5.1 Overview

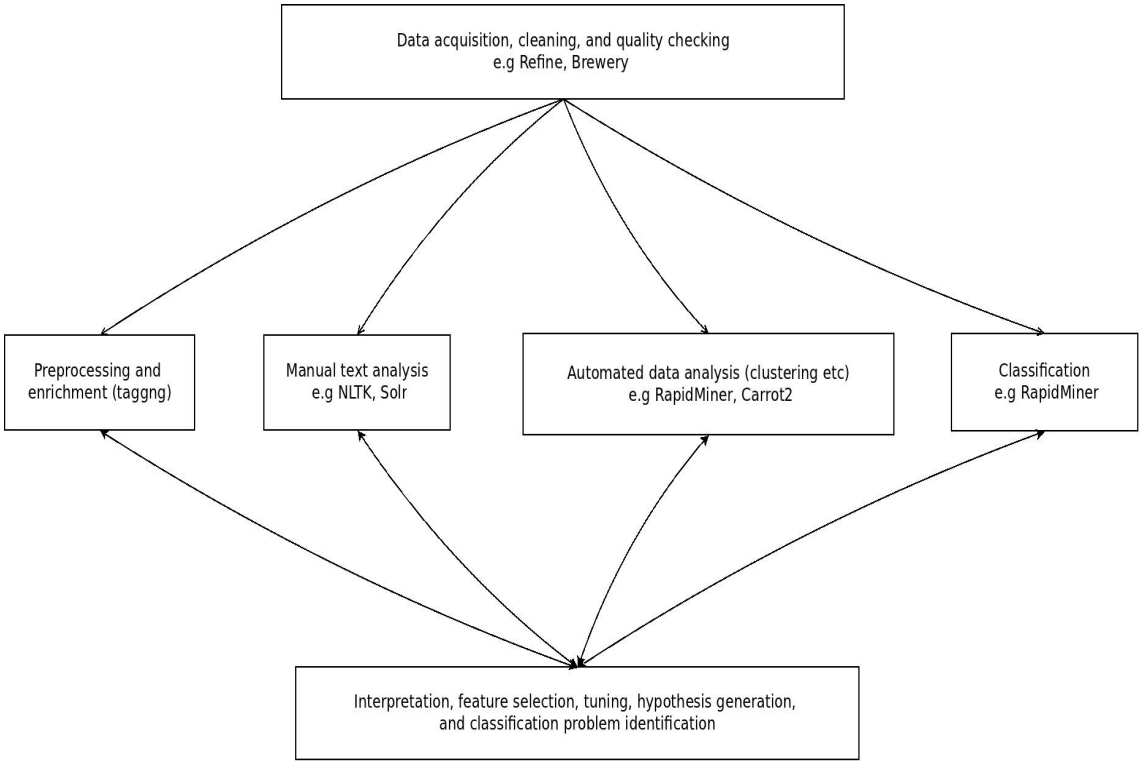


Figure 5.1: Overview of the data mining process. External corpora may be used to enrich data analysis. Visualisation is also usually performed

5.2 H

5.2.1 Apache Solr

Software bugs

Poor reliability

Ongoing problems

5.2.2 NLTK

People frequently report being unable to perform tasks because of computer systems failure. Using NLTK I created a frequncy distribution plot of the top 50 words which follow "unable to" in the text.

In addition, NLTK can identify collocations, pairs of words

5.3 Preprocessing techniques

5.3.1 Stemming

e.g porter stemmer http://tartarus.org/ martin/PorterStemmer/index.html

5.3.2 Tokenization

5.3.3 Tagging

Named entity recognition

http://uima.apache.org/

Parts of speech tagging

5.4 Linguistic analysis

5.4.1 Collocations

point wise mutual information

Top 10 point wise mutual information trigrams, filtering for those occuring less than 10 times:

[('First', 'Advice', 'Queue'), ('Abbott', 'Diagnostic', 'where'), ('regarding', 'poor', 'connectivity'), ('Ongoing', 'issues', 'regarding'), ('serial', 'no', 'C1600445'), ('by', 'Abbott', 'Diagnostic'), ('instrument', 'serial', 'no'), ('issues', 'regarding', 'poor'), ('Contributing', 'factors', ':'), ('Front', 'sheet', 'did')]

5.4.2 Regular expression matching

e.g using nltk

5.4.3 Term frequency-inverse document frequency

5.4.4 Cosine similarity

5.5 Association and item set mining

5.5.1 Association rule learning

5.6 Anomaly detection

5.7 Clustering

5.7.1 K-means clustering algorithm

5.7.2 Lingo clustering algorithm

5.8 Classification

5.8.1 Classification task

The example classification task chosen was to clasify the degre of harm of a clinical incident based on its description.

The training data set was composed of incident description and degree of harm fields. Each incident has an incident description and a human-assigned degree of harm

catergory.

Catergories used are:

- Death
- Severe
- Moderate
- Low
- No Harm

Classifier performance was measured by cross validation. The dataset is split into 10 partitions each containing 90% of the data, the training set, and tested against the remaining 10% of the data, the testing set. Partitions are created by random sampling of the dataset.

5.8.2 Rule induction

Single rule induction

5.8.3 Bayesian modelling

Naive bayes classifier

5.8.4 K nearest neighbour classifier

5.8.5 Support vector modelling

Chapter 6

Tuning steps, data visualisation and interpretation

6.1 Tuning steps

6.1.1 Preprocessing

6.1.2 Attribute selection

6.1.3 Alogrithm selection

6.1.4 Parameter optimization

Automated parameter optimization

6.1.5 Measuring performance

Cross validation

6.2 Data visualisation and interpretation

installation notes: solr - schema.xml + csv loading instructions from gmail

data source -¿ refine -¿ format for processing -¿ process -¿ visualization

Data source is SAS search for all computer incidents in NRLS = 7000 -¿ convert to csv

process for analysis platform

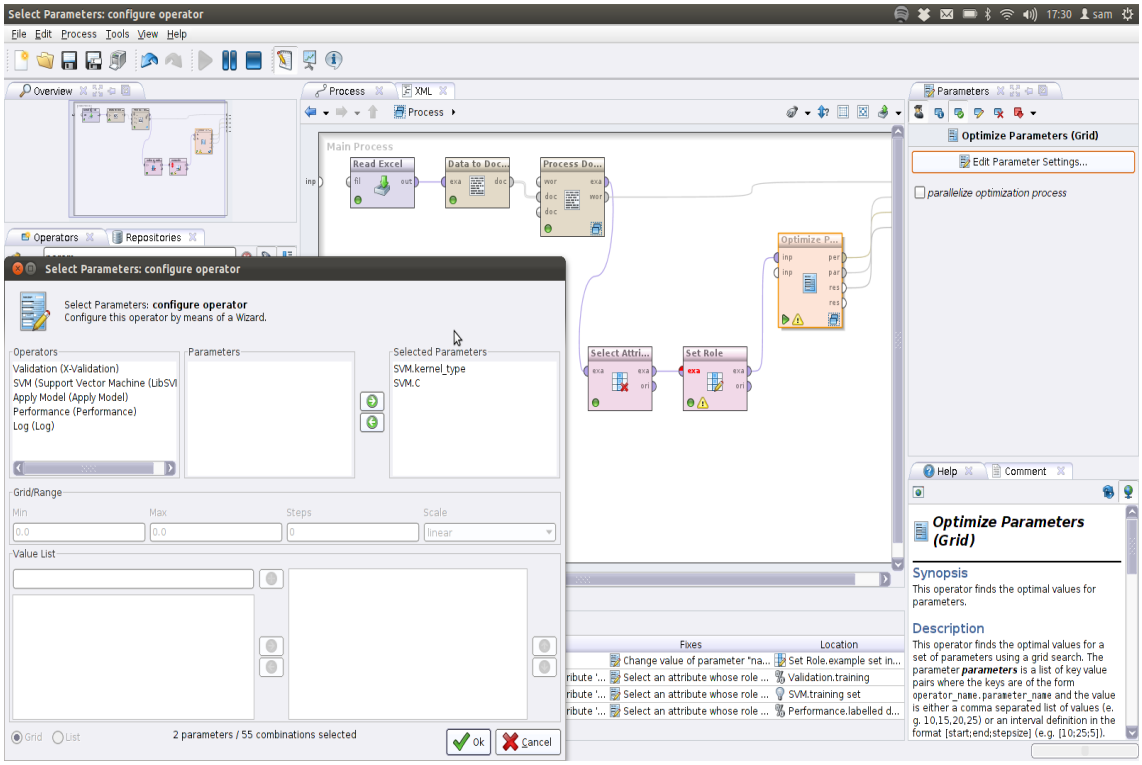


Figure 6.1: The optimize paramater process can be set to iterate through all possible parameter settings in a given range to identify the settings resulting in the best classifier performance

Two (Three? Four?) Example Algorithms, Why Chosen, How they work

Data mining techniques (to be informed by literature review of what has already been done) - clustering (generatecatergories/lables e.g with Carrot2) - classification (predefine catergories/labels e.g GATE, Weka) independant variable is free text dependant variable e.g harm - regression - summarization (e.g with NLTK)

1. The EHI Intelligence database was queried to identify examples of the major electronic medical record systems in primary and secondary care. 2. The NPSA NRLS database was queried using SAS *version* for all incidents coded as IT / telecommunications failure / overload incidents containing the words (Cerner, Lorenzo, Rio, EMIS, and PACS)

tf-idf We now combine the definitions of term frequency and inverse document frequency, to produce a composite weight for each term in each document. The tf-idf weighting scheme assigns to term a weight in document given by

(22)

In other words, assigns to term a weight in document that is

highest when occurs many times within a small number of documents (thus lending high discriminating power to those documents); lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);

lowest when the term occurs in virtually all documents.

Part III

Results

Chapter 7

Results

7.1 Descriptive statistics

7.1.1 "computer problem" extract

Data quality

The NRLS "computer problem" extract contains 7273 rows of data. Each row represents a clincal incident report. Each column represents one of 32 fields **??** which are extracted from the incident report form and cleaned to remove sensitive data by the NPSA. The dataset contained 159 duplicates **??** which were removed before further analysis. The completness of fields varied considerably and some fields are derived from other fields, for example, the age range field is populated by the patient age field.

Is the description of the incident a duplicate?	
False	7114
True	159

Table 7.1: Incidents with identical incident descriptions (duplicates)

Descriptive statsistics

What was the degree of harm?

Which specialties were involved?

Field name	**Field completness(%)**
Unique Incident Id	100.0
Age At Time Of Incident Date	30.0
Incident Received By Npsa	100.0
Type Of Device	6.0
Patent Age At Time Of Incident	30.0
Date Record Exported To Nrls	100.0
Date Of Incident	99.9
Location (lvl1)	100.0
Location (lvl2)	99.7
Location (lvl3)	55.6
Location - Free Text	7.5
Incident Category - Lvl1	100.0
Incident Category - Lvl2	100.0
Incident Category - Free Text	0.0
Free Text Description Of What Happened	100.0

Table 7.2: Field names in the dataset and percentage completness. Percentage field completness is calculated by dividing the number of non-blank fields for a given field name by the number of incident reports, after removing duplicates, and multiplying by 100.

Field name	Field completness(%)
Actions Preventing Reoccurrence	38.0
Apparent Causes	21.5
Med Process	1.6
Med Error Category	1.5
Approved Name (drug 1)	0.1
Proprietary Name (drug 1)	0.0
Patient Age Range	30.3
Patient Sex	55.5
Specialty - Lvl 1	100.0
Specialty - Lvl 2	63.3
Speciality - Free Text	18.8
Degree Of Harm (severity)	100.0
Paediatric Care	29.9
Source Of Notification	100.0
Care Setting Of Occurrence	100.0
Reason Exclusion	0.0

Table 7.3: Field names in the dataset and percentage completness. Percentage field completness is calculated by dividing the number of non-blank fields for a given field name by the number of incident reports, after removing duplicates, and multiplying by 100.

What was the degree of harm?	N
Death	7
Severe	65
Moderate	268
Low	792
No Harm	5982

Table 7.4: Degree of harm caused by incidents.

Specialty	N
Medical specialties	1104
Diagnostic services	1039
Not applicable	995
Obstetrics and gynaecology	695
Surgical specialties	673
Other	600
Unknown	526
Primary care / Community	508
Accident and Emergency	415
Other specialties	137
Mental health	123
PTS (Patient Transport Service)	98
Learning disabilitie s	78
Anaesthesia Pain Management and Critical Care	62
Dentistry - General and Community	60
Childrens Specialties	1

Table 7.5: Specialty involved in incidents, based on level one data.

Specialty	N
Acute / general hospital	5054
Ambulance service	121
Community and general dental service	34
Community nursing, medical and therapy service (incl. community hospital)	1614
General practice	92
Learning disabilities service	17
Mental health service	182

Table 7.6: Location of the incidents, based on care setting data.

Where did the incident happen?

7.2 Identification of selected themes using a search engine and NLTK

Poor system reliability Using NLTK it was established that the words "unable to" were followed by another word in 1398 instances. 256 unique words made up these instances.

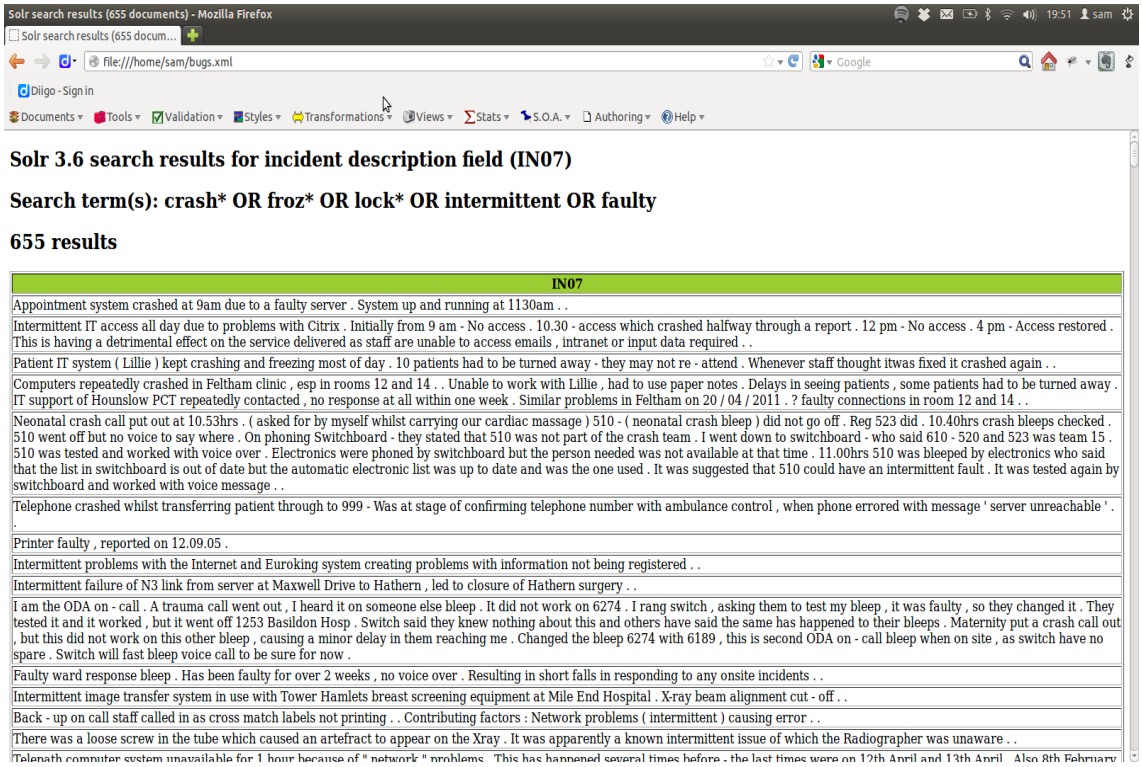


Figure 7.1: Search results for search on terms associated with poor system reliability.

Problems not fixed promptly

- "Yet again 400 + results , some a year old , have appeared today for signing off . This is a significant risk which has been highlighted repeatedly and I am not aware that any action has been taken / or even that a clear explanation has been given for why it occurs . (Examples of 4 sample patients listed on form .) ."
- "Computerised ' Carevue' system for charting vital sighs and ventilation on Nicu and Picu freezing over past 24hrs . ICT phoned x2 and team in Vienna investi-

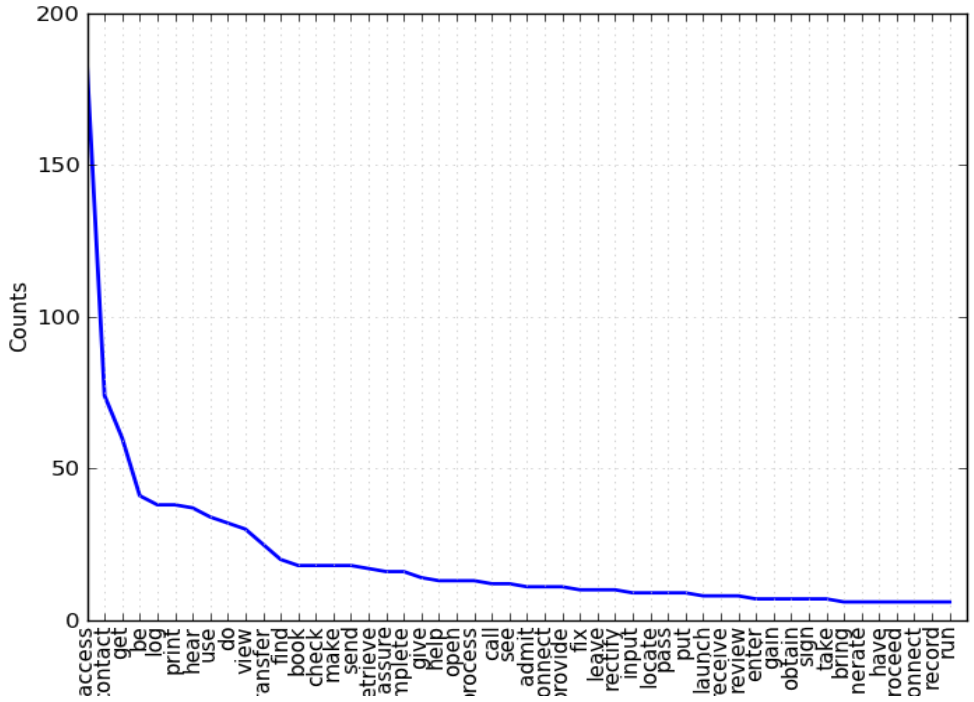


Figure 7.2: Frequency distribution plot of the 50 most common words to follow the words “unable to” in the text.

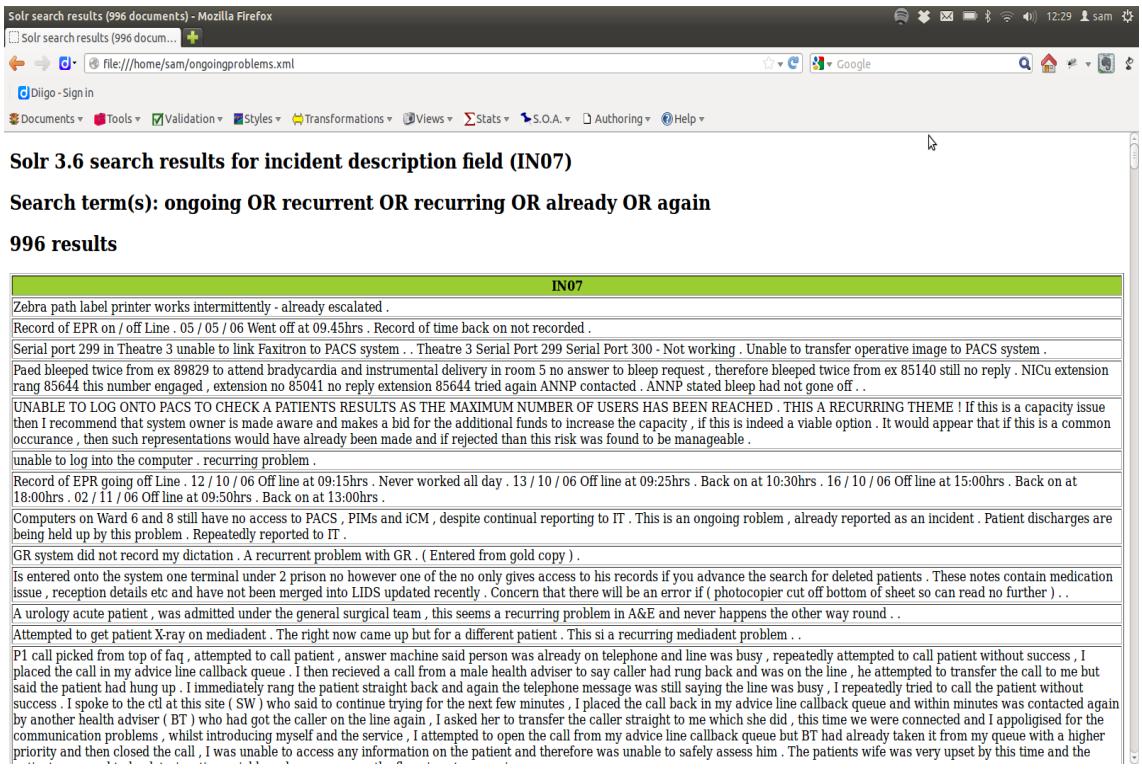


Figure 7.3: Searching the dataset for terms which might indicate an ongoing problem such as “ongoing OR recurrent OR recurring OR already OR again” yields 996 results.

gating . Was rebooted overnight . At 06.15 the carevue system again froze , this time completely , not allowing any entering or viewing of screens . This has an impact on patient safety as the day staff cannot review the patients and accuratly assess their clinical stability and plan treatment accordinglyly . Also , 2 new patients to NICU overnight , who clinical information cannot be accessed . ."

7.3 Linguistic analysis

7.3.1 Basic statistics

There were 408657 words in the text. 14018 words formed the vocabulary. The lexical diversity (number of words in text/vocabulary) was 29.

7.3.2 Collocations

A collocation is a sequence of words that occur together unusually often. Thus, "red wine" is a collocation, whereas "the wine" is not. Collocations are resistant to substitution, for example, "maroon wine" would sound very odd. They tend to be specific to a given text and so can convey useful information.[6]

Collocations identified in "computer problem" extract were:

- Staff Name
- cardiac arrest
- computer system
- Health Advisor
- warm transfer
- staff member
- Contributing factors
- labour ward
- switch board

- health advisor
- call back
- Staff member
- blood results
- NHS Direct
- several times
- help desk
- ambulance control
- File Closed
- ambulance service
- Staff name

7.3.3 Lexical dispersion plot

7.3.4 Cumulative frequency plot

7.3.5 Term frequency-inverse document frequency

7.3.6 Cosine similarity

7.4 Clustering

7.4.1 K-means clustering algorithm

7.4.2 Lingo clustering algorithm

Top 10 results by cluster size Table + excerpts

Top 10 results by score Table + excerpts

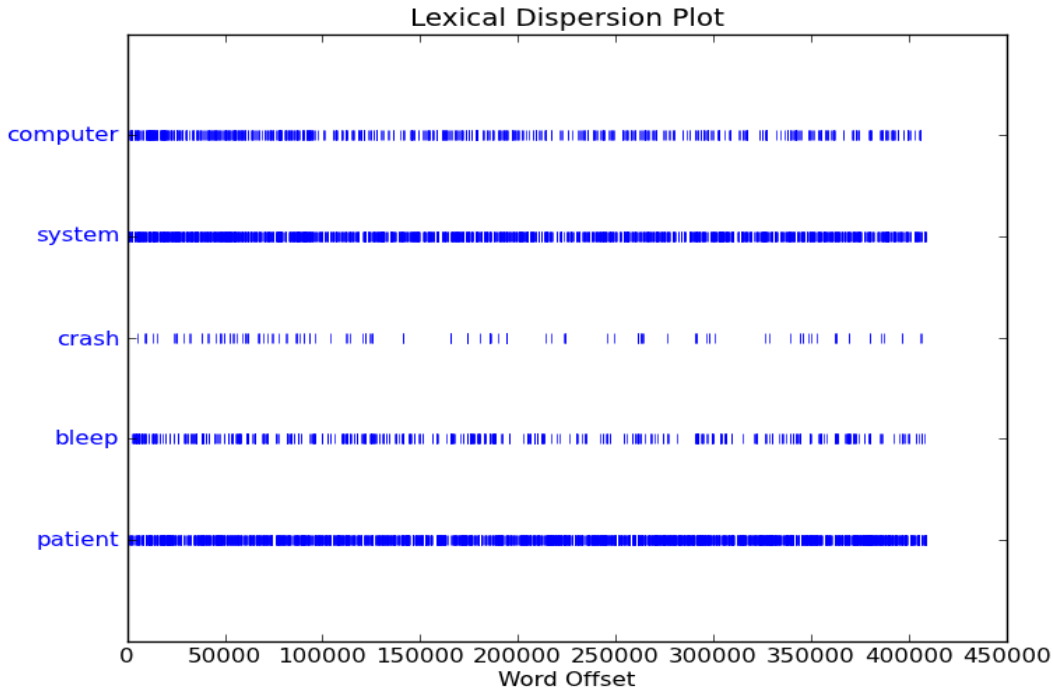


Figure 7.4: Each stripe represents an instance of a word and each row represents the entire text. If present, trends in the frequency of the use of words over time can be visualized in this way (the incident descriptions forming the text have been time ordered).

Cluster label	**Number of documents**
Computers Unable	1662
System Patient	759
Patient Called	756
Patients Unable	660
Patients and Staff	527
Bleep Problem	526
System Unable	520
Inform Patients	488
Time Patients	460
Patient Clinic	439

Table 7.7: Automatically generated cluster labels for the ten largest clusters identified by the lingo algorithm

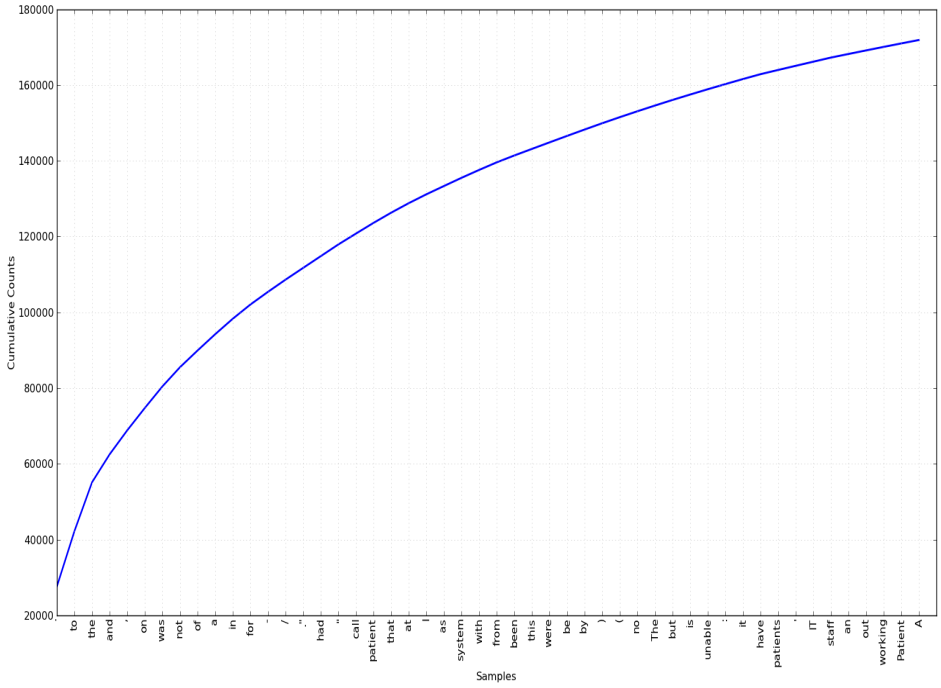


Figure 7.5: Cumulative frequency plot for the 50 most frequently used words in incident reports about computer problems.

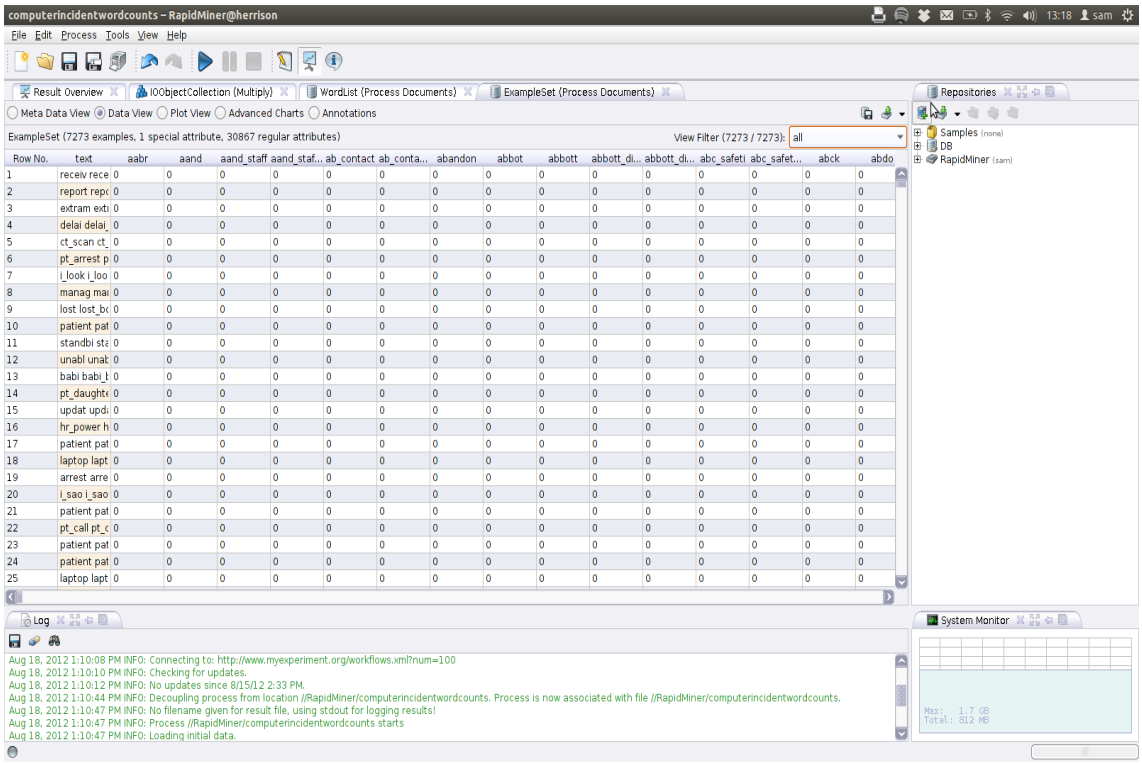


Figure 7.6: Each row represents a single incident report. Each column heading is a single term present in the collection of incidents.

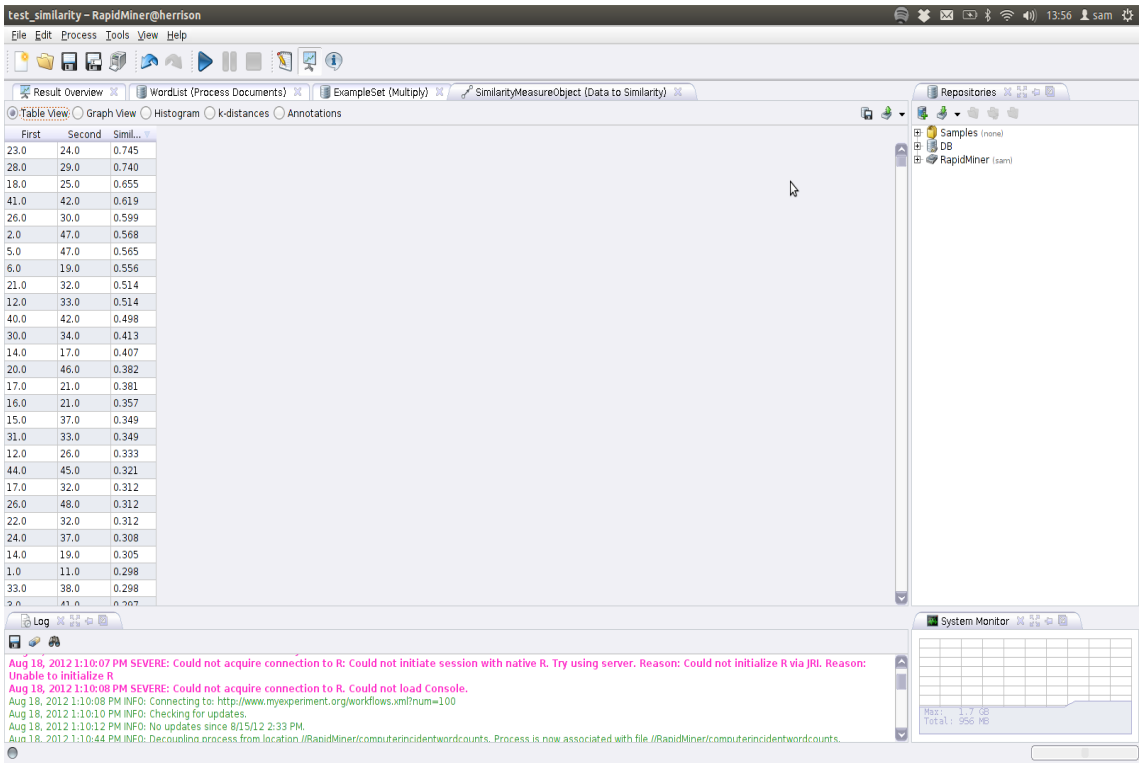


Figure 7.7: Cosine similarity for sample of 50 incidents.

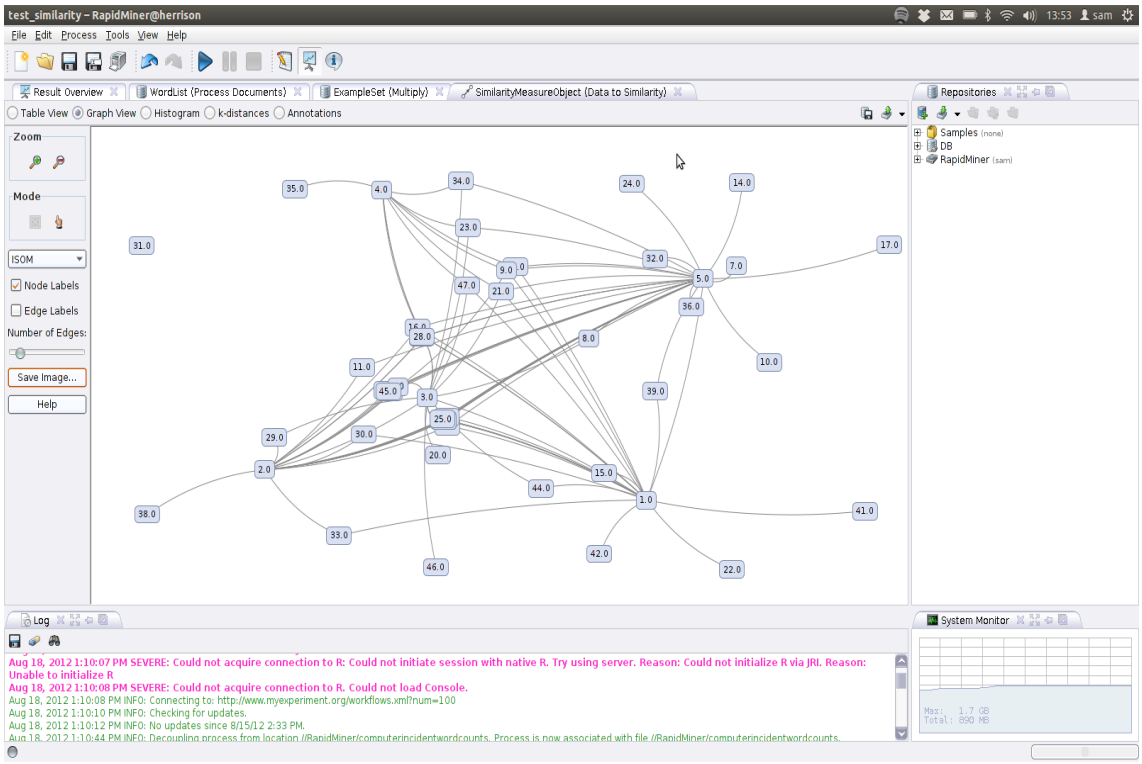


Figure 7.8: Cosine similarity for sample of 50 incidents, ISOM visulalisation

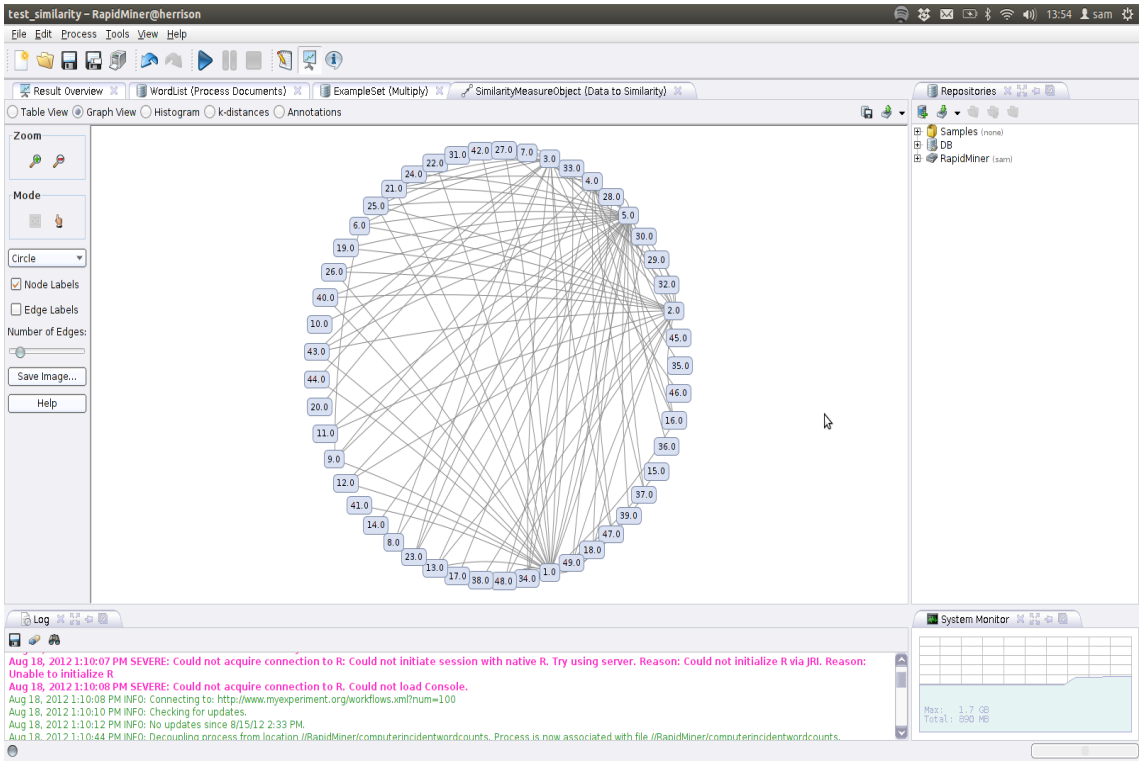


Figure 7.9: Cosine similarity for sample of 50 incidents, circle visulalisation

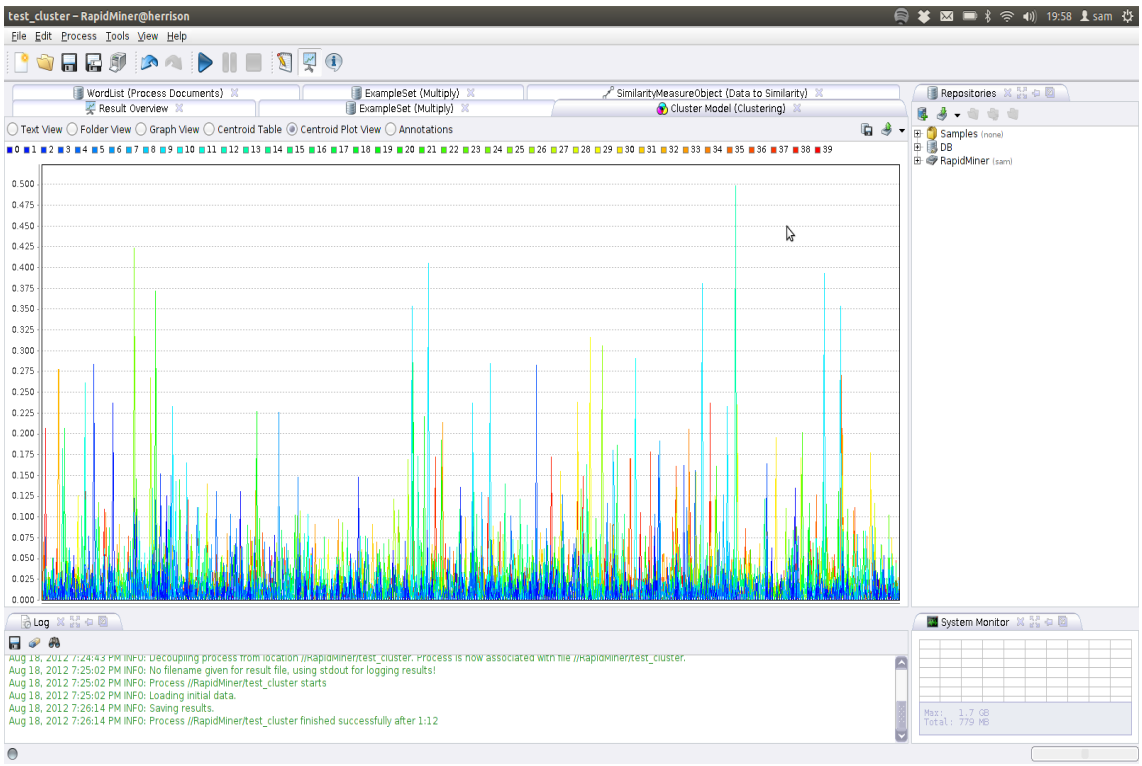


Figure 7.10: Centroid plot view of K means clustering of a sample of 500 incident reports.

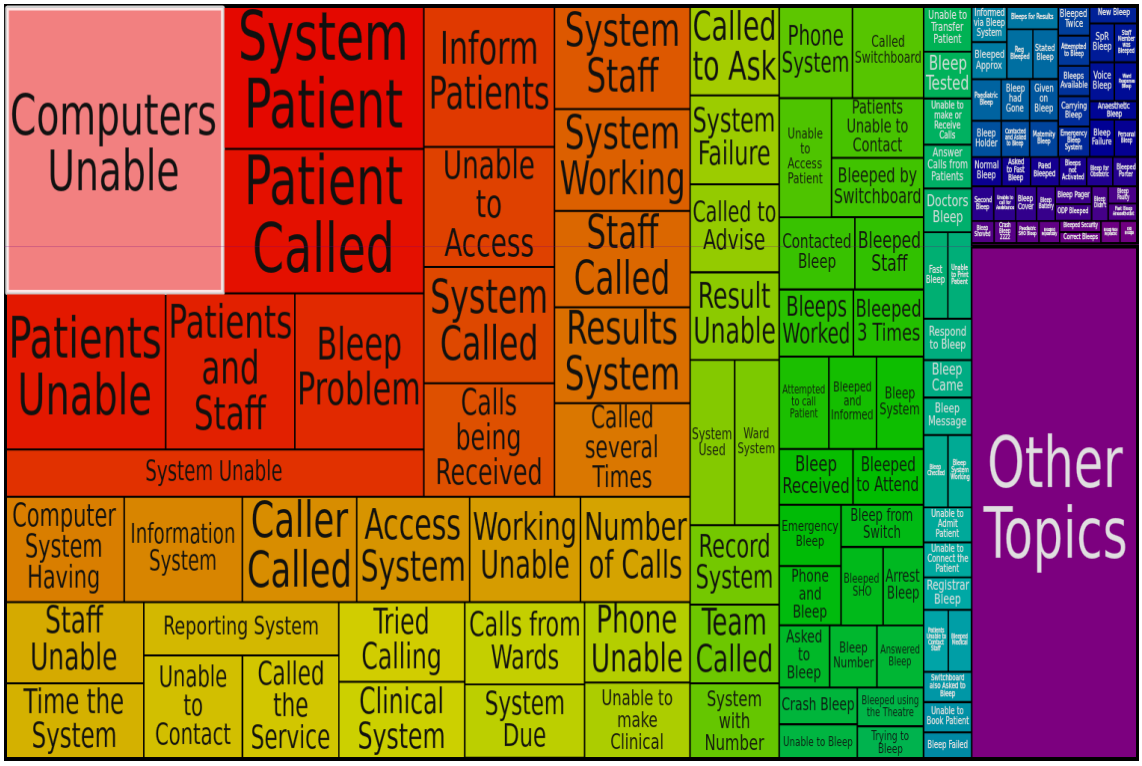


Figure 7.11: Each row represents a single incident report. Each column heading is a single term present in the collection of incidents.

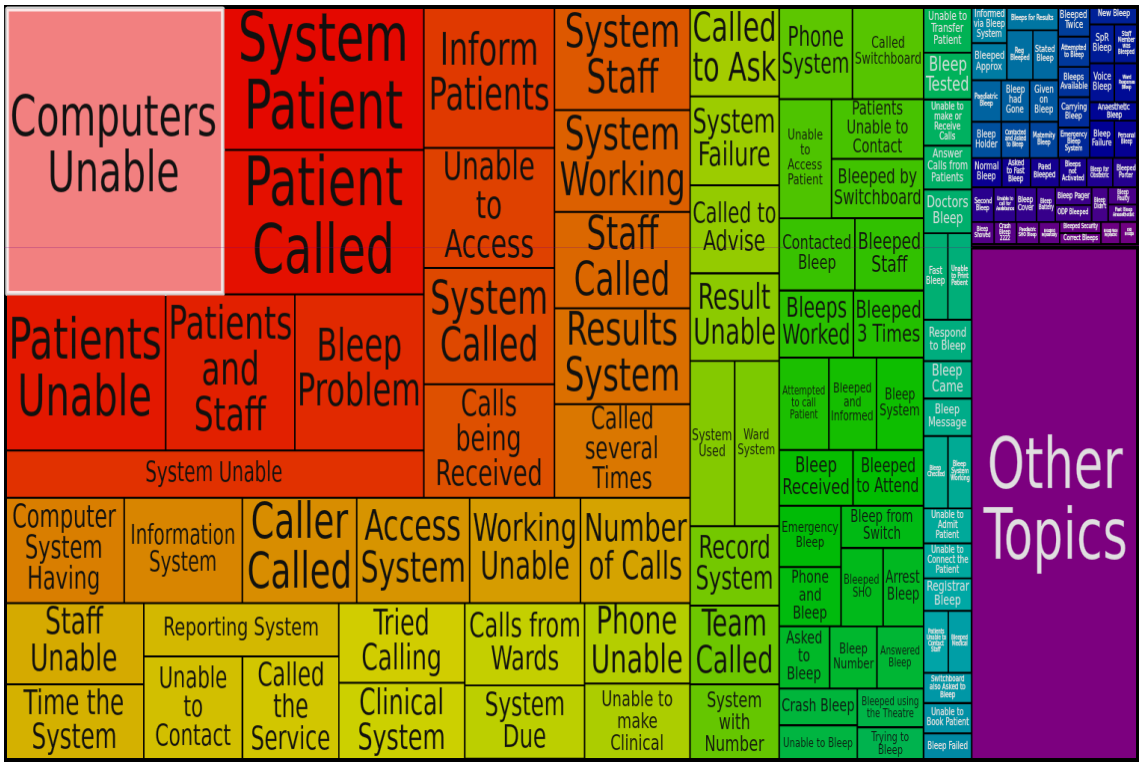


Figure 7.12: Each row represents a single incident report. Each column heading is a single term present in the collection of incidents.

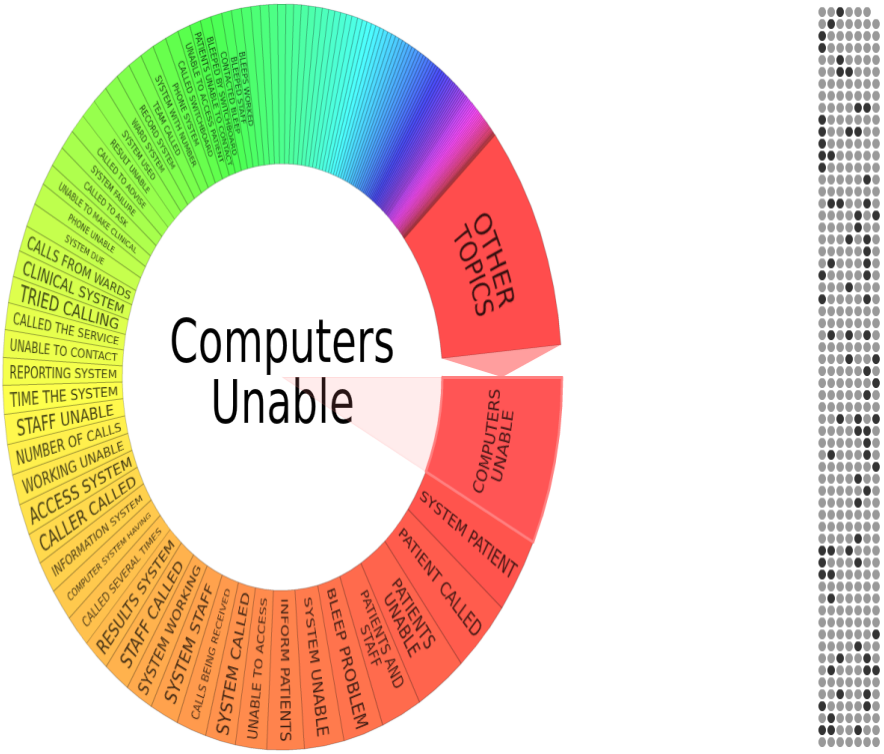


Figure 7.13: Each row represents a single incident report. Each column heading is a single term present in the collection of incidents.

Cluster label	Cluster reliability score
Bleep Problem	21.8
Unable to Access	12.4
System Staff	11.0
System Working	10.3
Calls being Received	9.6
Patients and Staff	9.3
Staff Called	7.9
Working Unable	7.6
Inform Patients	7.4
Called several Times	7.4

Table 7.8: Automatically generated cluster labels for the ten highest reliability scores identified by the lingo algorithm. The higher the reliability score the higher the reliability of the cluster content.

Bleep problems

- Bleep problem: "ODP bleeped 3 times giving no reply, rang main theatres who said ODP was in theatre 17. The bleep system does not work in this theatre, which is widely known, expect the ODP was unaware of this."
- Computer problem but not to do with bleep system: "[Manager] informed by bleep [Number] holder that [Trust name]PAS was down. No contingency plan available. Wards not aware that [Trust name]PAS being taken down for urgent updating/maintenance - poor communication by [Team name]. No ability to generate patient record num..."
- Unclear if bleep sytem problem or not: Pt daughter was feeding her. When a staff nurse was passing pt's daughter said that her mother was not well, having difficulty breathing. Went to see pt - breathing shallow. Oxygen was started. Bleeped Dr twice - no reply. Co-Ord informed. Co-Ord came immediately - card...

Cluster label	Number of documents	Cluster reliability score
Bleeps Worked	131	47.6
Bleeped 3 Times	127	44.5
Bleeped and Informed	117	27.8
Bleep System	115	50.2
Bleep Received	112	48.1
Bleeped to Attend	104	28.1
Emergency Bleep	103	49.8
Phone and Bleep	99	51.5
Bleep from Switch	91	41.8
Bleeped SHO	86	57.8

Table 7.9: Higher cluster reliability scores are achieved by applying lingo algorithm to incidents containing the word 'bleep'.

7.5 Association and item set mining

7.5.1 Association rule learning

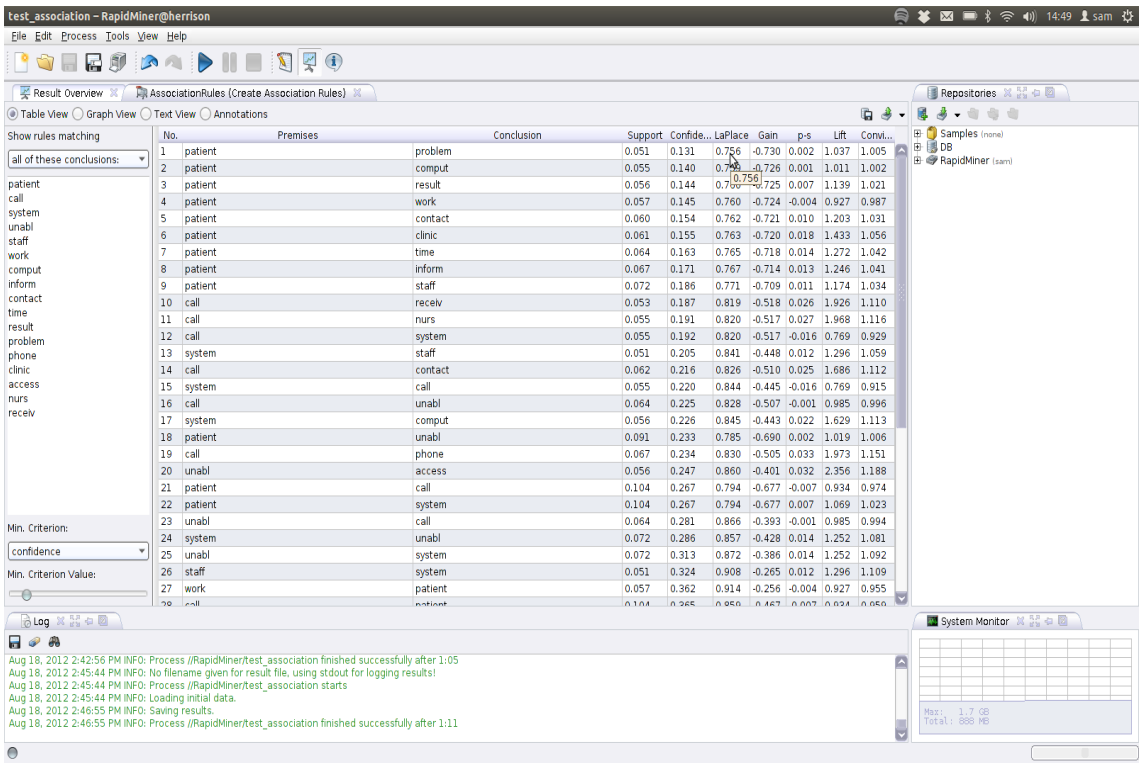


Figure 7.14: Unsupervised association rule learning results for all incidents.

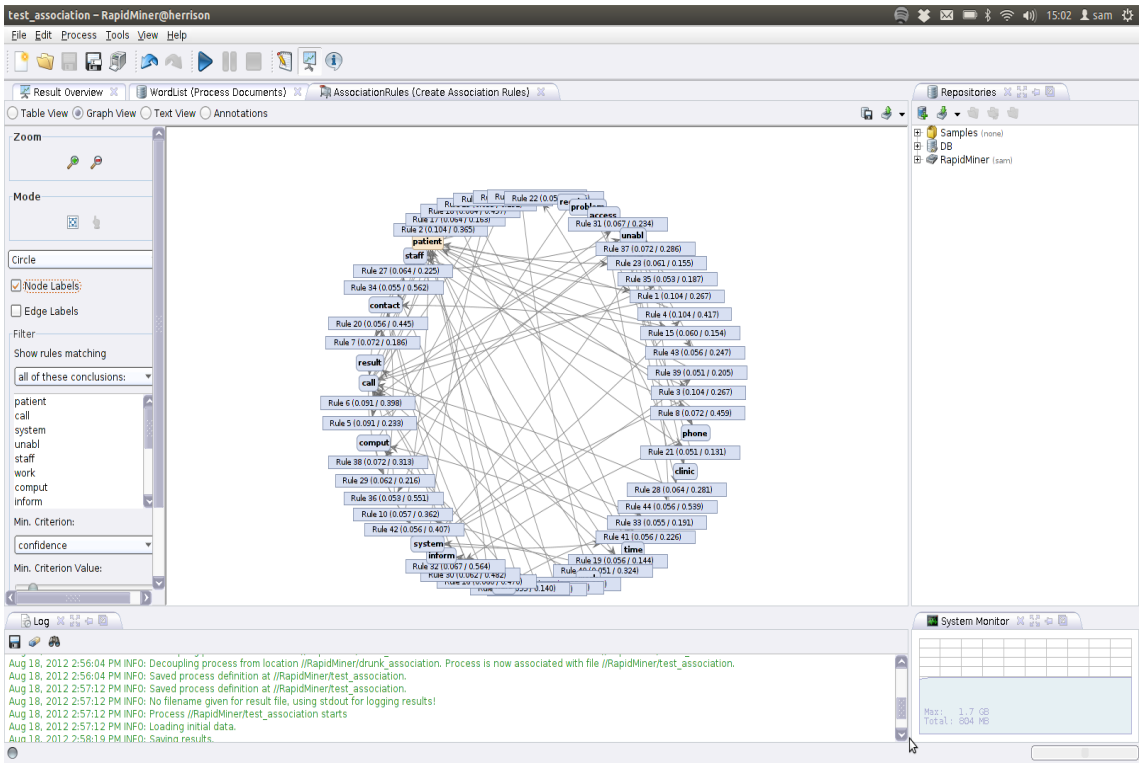


Figure 7.15: Unsupervised association rule learning results for all incidents, graph visualisation.

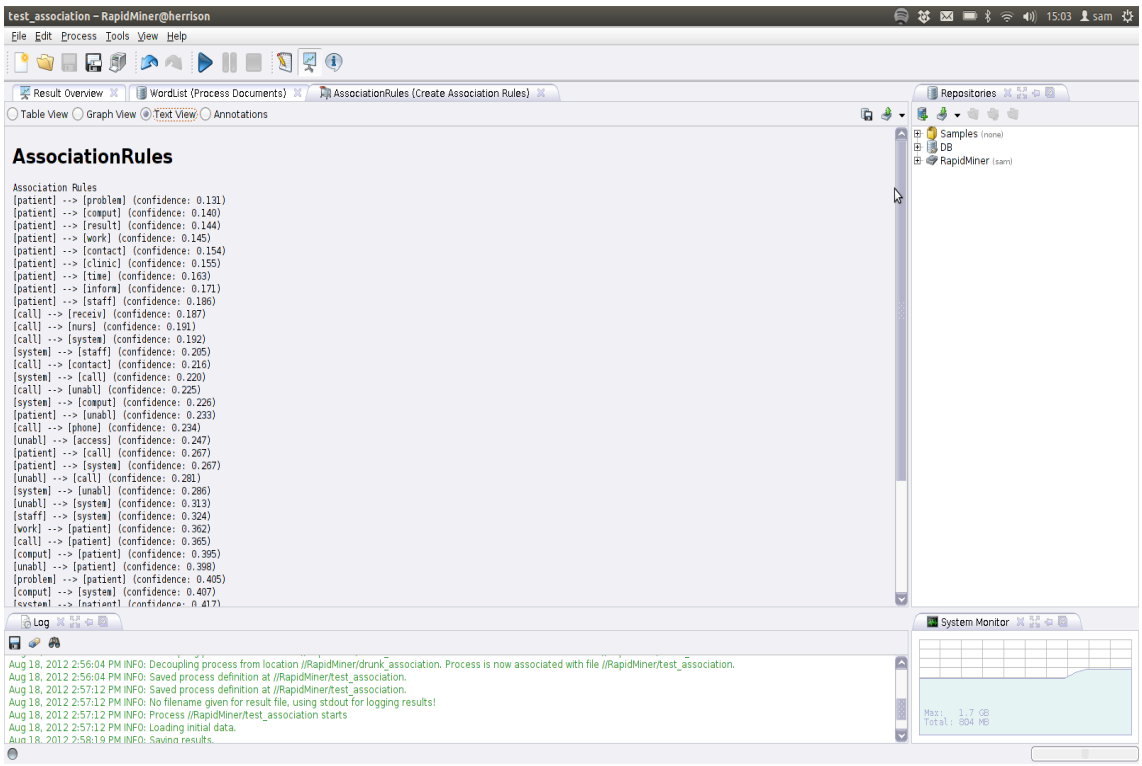


Figure 7.16: Unsupervised association rule learning results for all incidents, text summary.

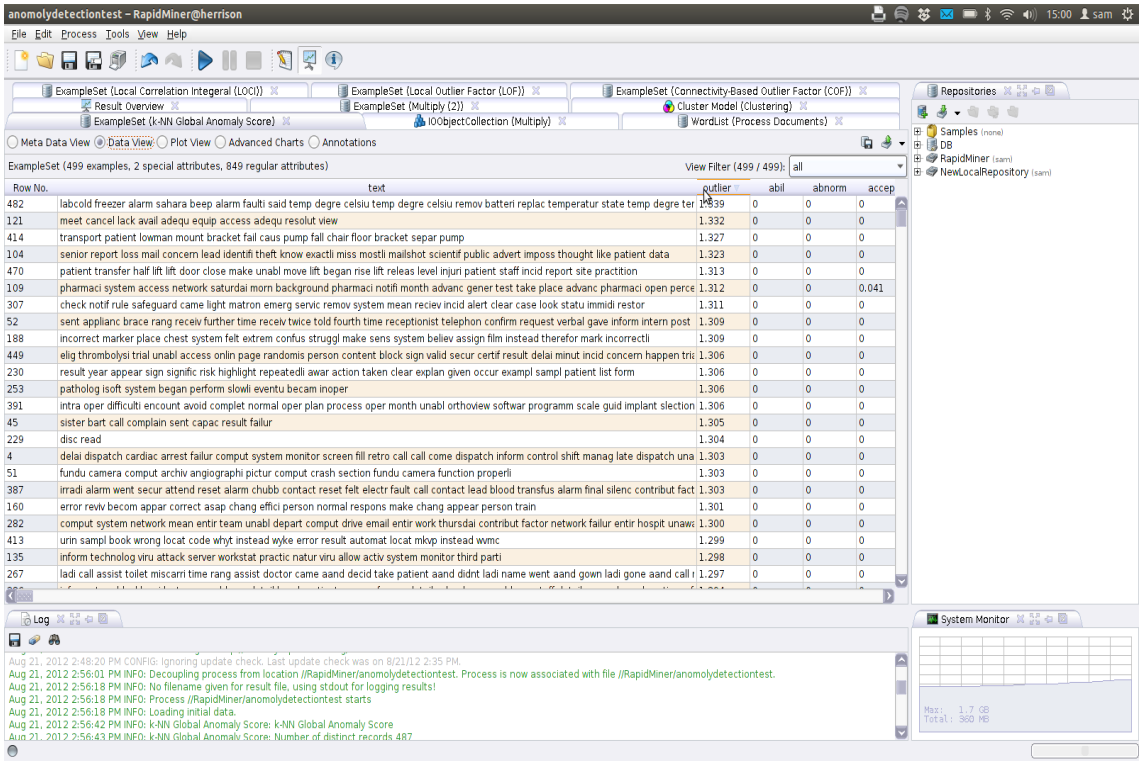


Figure 7.17: Centroid plot view of K means clustering of a sample of 500 incident reports.

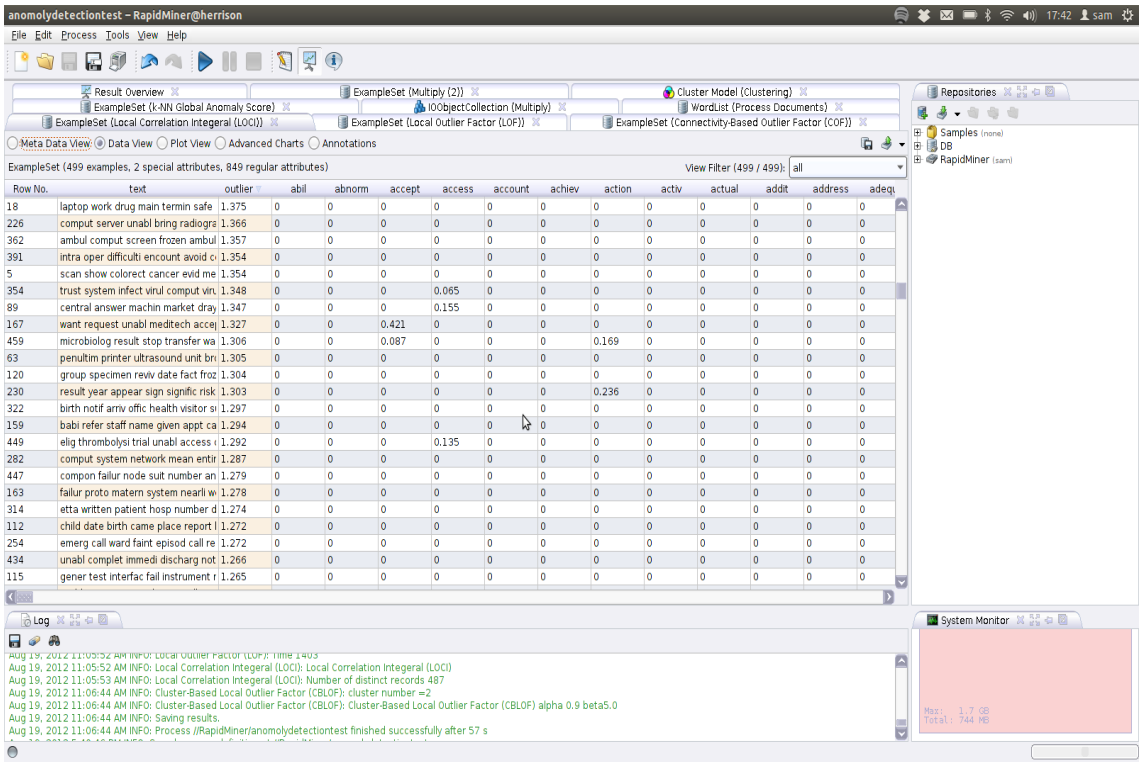
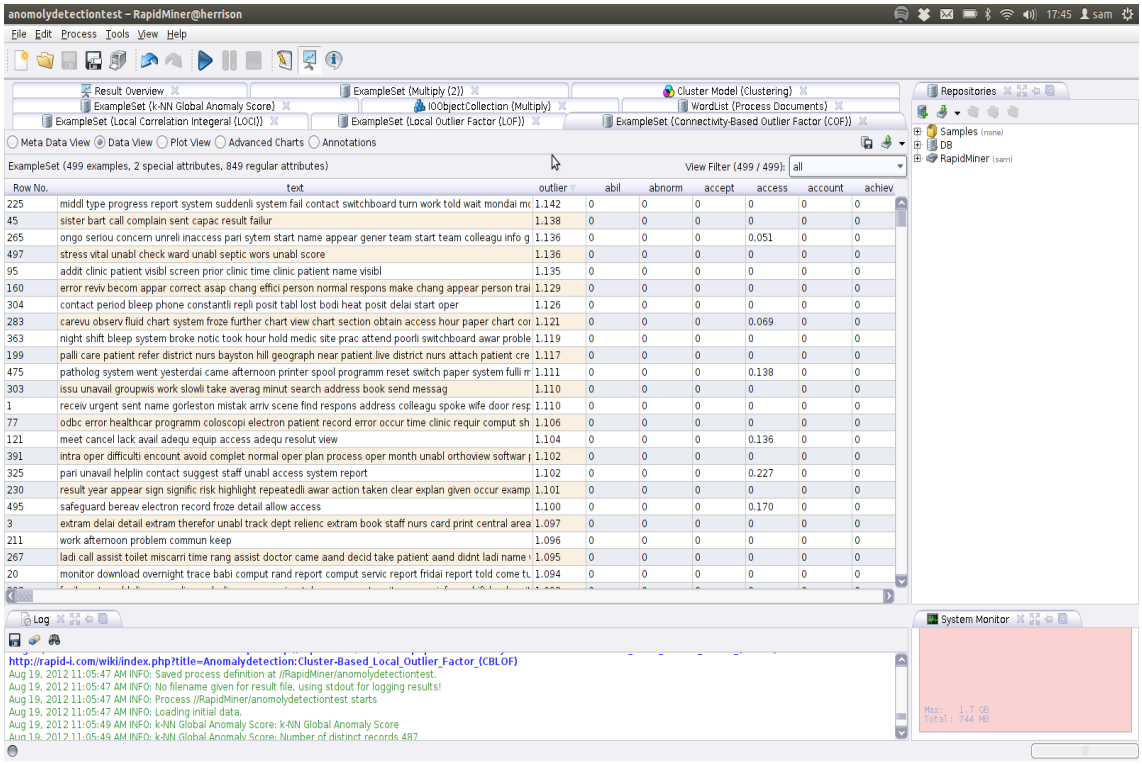
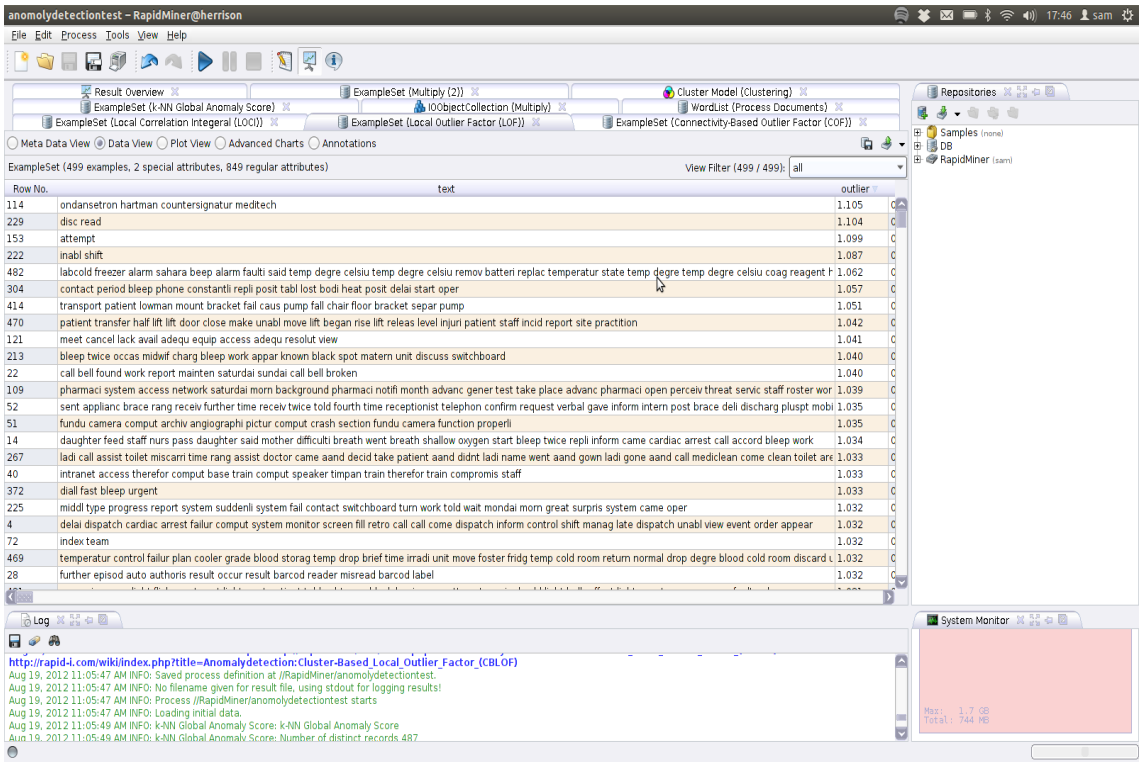


Figure 7.18: Centroid plot view of K means clustering of a sample of 500 incident reports.



Row No.	text	outlier	abil	abnorm	accept	access	account	achiev
225	middl type progress report system suddenli system fail contact switchboard turn work told wait mondai m	1.142	0	0	0	0	0	0
45	sister bart call complain sent capac result failur	1.138	0	0	0	0	0	0
265	ongo seriou concern unreli inaccess pari sytem start name appear gener team start team colleagu info g	1.136	0	0	0	0.051	0	0
497	stress vital unabl check ward unabl septic wors unabl score	1.136	0	0	0	0	0	0
95	addit clinic patient visibl screen prior clinic time clinic patient name visibl	1.135	0	0	0	0	0	0
160	error reviv becom appar correct asap chang effici person normal respons make chang appear person trai	1.129	0	0	0	0	0	0
304	contact period bleep phone constantli repli posit tabl lost bodi heat posit delai start oper	1.126	0	0	0	0	0	0
283	carevu observ fluid chart system froze further chart view chart section obtain access hour paper chart cor	1.121	0	0	0	0.069	0	0
363	night shift bleep system broke notic took hour hold medic site prac attend poorli switchboard awar proble	1.119	0	0	0	0	0	0
199	palli care patient refer district nurs bayston hill geograph near patient live district nurs attach patient cre	1.117	0	0	0	0	0	0
475	patholog system went yesterdai came afternoon printer spool programm reset switch paper system fulli r	1.111	0	0	0	0.138	0	0
303	issu unavail groupwis work slowli take averag minut search address book send messag	1.110	0	0	0	0	0	0
1	receiv urgent sent name gorleston mistak arriv scene find respons address colleagu spoke wife door resp	1.110	0	0	0	0	0	0
77	odbc error healthcar programm coloscopi electron patient record error occur time clinic requir comput sh	1.106	0	0	0	0	0	0
121	meet cancel lack avail adequ equip access adequ resolut view	1.104	0	0	0	0.136	0	0
391	intra oper difficulti encount avoid complet normal oper plan process oper month unabl orthoview softwar	1.102	0	0	0	0	0	0
325	pari unavail helplin contact suggest staff unabl access system report	1.102	0	0	0	0.227	0	0
230	result year appear sign signific risk highlight repeatedli awar action taken clear explan given occur examp	1.101	0	0	0	0	0	0
495	safeguard bereav electron record froze detail allow access	1.100	0	0	0	0.170	0	0
3	extram delai detail extram therefor unabl track dept relienc extram book staff nurs card print central area	1.097	0	0	0	0	0	0
211	work afternoon problem commun keep	1.096	0	0	0	0	0	0
267	ladi call assist toilet miscarri time rang assist doctor came aand decid take patient aand didnt ladi name	1.095	0	0	0	0	0	0
20	monitor download overnight trace babi comput rand report comput servic report fridai report told come tu	1.094	0	0	0	0	0	0

Figure 7.19: Centroid plot view of K means clustering of a sample of 500 incident reports.



Row No.	text	outlier
114	ondansetron hartman countersignatur meditech	1.105
229	disc read	1.104
153	attempt	1.099
222	inabl shift	1.087
482	labcold freezer alarm sahara beep alarm faulti said temp degre celsiu temp degre celsiu remov batteri replac temperatur state temp degre temp degre celsiu coag reagent h	1.062
304	contact period bleep phone constantli repli posit tabl lost bodi heat posit delai start oper	1.057
414	transport patient lowman mount bracket fail caus pump fall chair floor bracket separ pump	1.051
470	patient transfer half lift lift door close make unabl move lift began rise lift releas level injuri patient staff incid report site practition	1.042
121	meet cancel lack avail adequ equip access adequ resolut view	1.041
213	bleep twice occas midwif charg bleep work appar known black spot matern unit discuss switchboard	1.040
22	call bell found work report mainten saturdai sundai call bell broken	1.040
109	pharmaci system access network saturdai morn background pharmaci notifi month advanc gener test take place advanc pharmaci open perceiv threat servic staff roster wor	1.039
52	sent applianc brace rang receiv further time receiv twice told fourth time receptionist telephon confirm request verbal gave inform intern post brace deli discharg plusptt mobi	1.035
51	fundu camera comput archiv angiographi pictur comput crash section fundu camera function properli	1.035
14	daughter feed staff nurs pass daughter said mother difficulti breath went breath shallow oxygen start bleep twice repli inform came cardiac arrest call accord bleep work	1.034
267	ladi call assist toilet miscarri time rang assist doctor came aand decid take patient aand didnt ladi name went aand gown ladi gone aand call mediclean come clean toilet are	1.033
40	intranet access therefor comput base train comput speaker timpan train therefor train compromis staff	1.033
372	diall fast bleep urgent	1.033
225	middl type progress report system suddenli system fail contact switchboard turn work told wait mondai morn great surpris system came oper	1.032
4	delai dispatch cardiac arrest failur comput system monitor screen fill retro call call come dispatch inform control shift manag late dispatch unabl view event order appear	1.032
72	index team	1.032
469	temperatur control failur plan cooler grade blood storag temp drop brief time irradi unit move foster fridg temp cold room return normal drop degre blood cold room discard u	1.032
28	further episod auto authoris result occur result barcod reader misread barcod label	1.032

Figure 7.20: Centroid plot view of K means clustering of a sample of 500 incident reports.

7.6 Anomaly detection

7.6.1 K-nearest neighbour global anomaly score

7.6.2 Local correlation integeral

7.6.3 Connectivity-based outlier factor

7.6.4 Local outlier Factor

7.7 Classification

7.7.1 Rule induction

Single rule induction

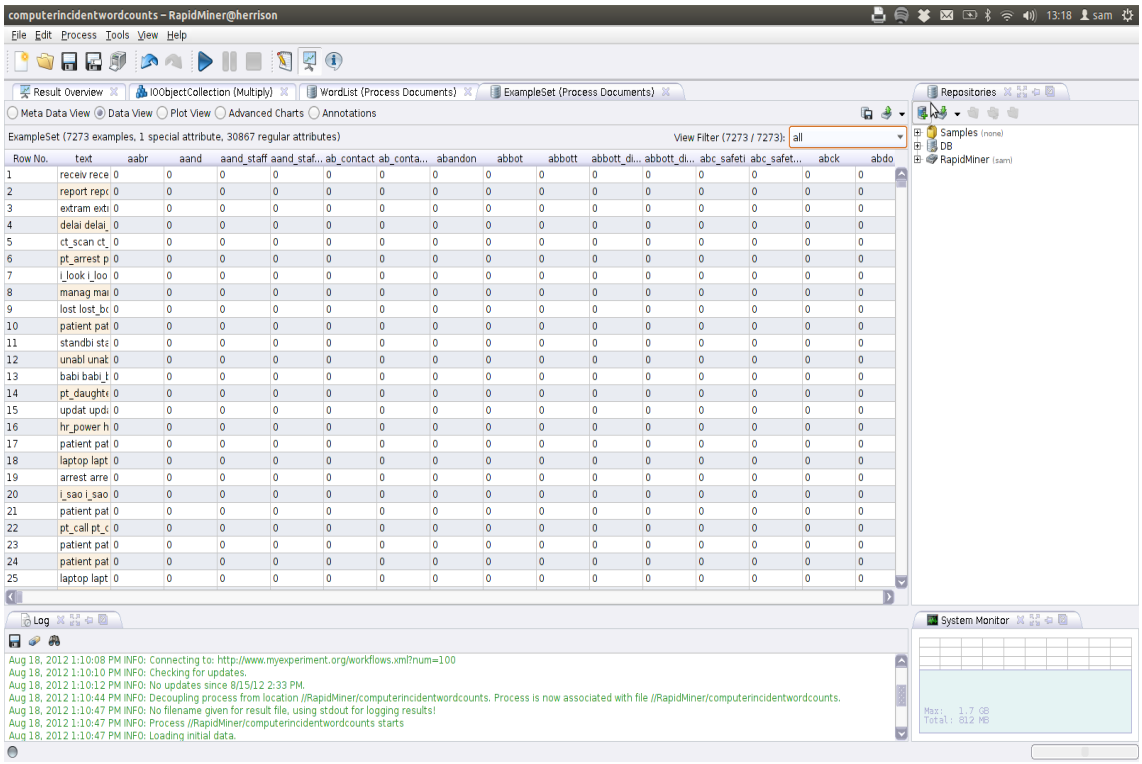
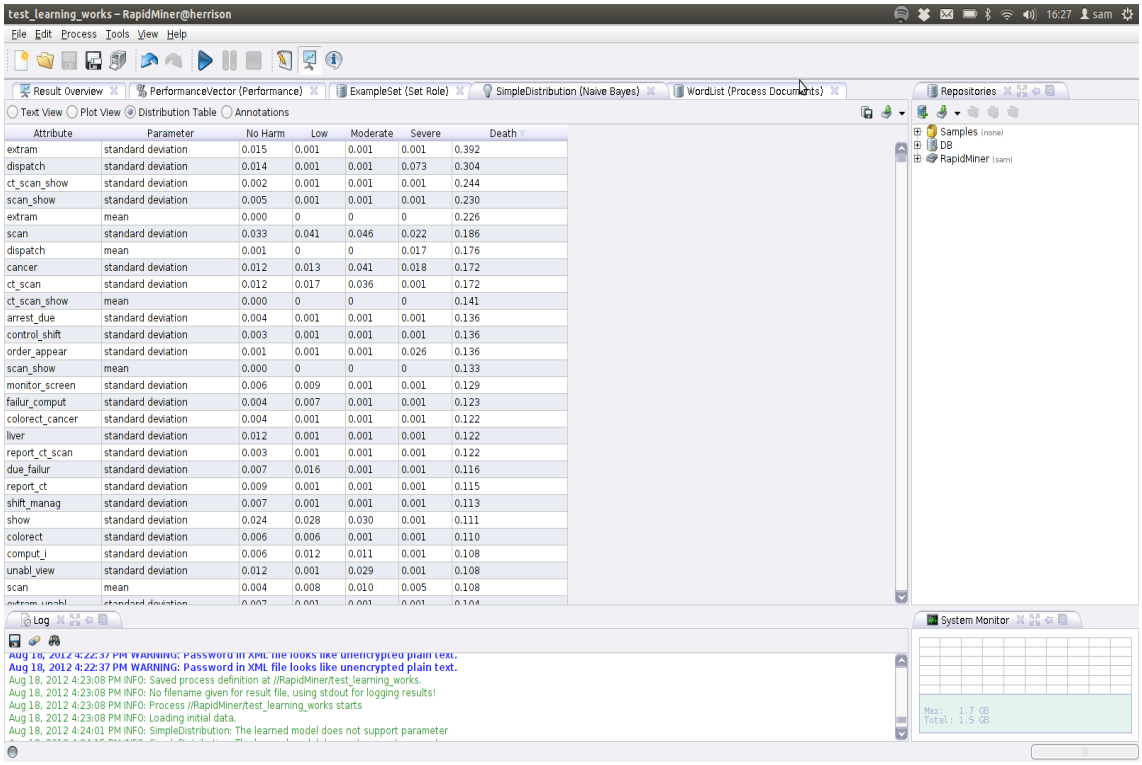


Figure 7.21: Each row represents a single incident report. Each column heading is a single term present in the collection of incidents.

7.7.2 Bayesian modelling

Naive bayes classifier



Attribute	Parameter	No Harm	Low	Moderate	Severe	Death
extram	standard deviation	0.015	0.001	0.001	0.001	0.392
dispatch	standard deviation	0.014	0.001	0.001	0.073	0.304
ct_scan_show	standard deviation	0.002	0.001	0.001	0.001	0.244
scan_show	standard deviation	0.005	0.001	0.001	0.001	0.230
extram	mean	0.000	0	0	0	0.226
scan	standard deviation	0.033	0.041	0.046	0.022	0.186
dispatch	mean	0.001	0	0	0.017	0.176
cancer	standard deviation	0.012	0.013	0.041	0.018	0.172
ct_scan	standard deviation	0.012	0.017	0.036	0.001	0.172
ct_scan_show	mean	0.000	0	0	0	0.141
arrest_due	standard deviation	0.004	0.001	0.001	0.001	0.136
control_shift	standard deviation	0.003	0.001	0.001	0.001	0.136
order_appear	standard deviation	0.001	0.001	0.001	0.026	0.136
scan_show	mean	0.000	0	0	0	0.133
monitor_screen	standard deviation	0.006	0.009	0.001	0.001	0.129
failur_comput	standard deviation	0.004	0.007	0.001	0.001	0.123
colorect_cancer	standard deviation	0.004	0.001	0.001	0.001	0.122
liver	standard deviation	0.012	0.001	0.001	0.001	0.122
report_ct_scan	standard deviation	0.003	0.001	0.001	0.001	0.122
due_failur	standard deviation	0.007	0.016	0.001	0.001	0.116
report_ct	standard deviation	0.009	0.001	0.001	0.001	0.115
shift_manag	standard deviation	0.007	0.001	0.001	0.001	0.113
show	standard deviation	0.024	0.028	0.030	0.001	0.111
colorect	standard deviation	0.006	0.006	0.001	0.001	0.110
comput_i	standard deviation	0.006	0.012	0.011	0.001	0.108
unabl_view	standard deviation	0.012	0.001	0.029	0.001	0.108
scan	mean	0.004	0.008	0.010	0.005	0.108

Figure 7.22: Top terms identified as predicting death by naive bayes classifier.

7.7.3 K nearest neighbour classifier

7.7.4 Support vector modelling

LibSVM

7.7.5 Neural networks

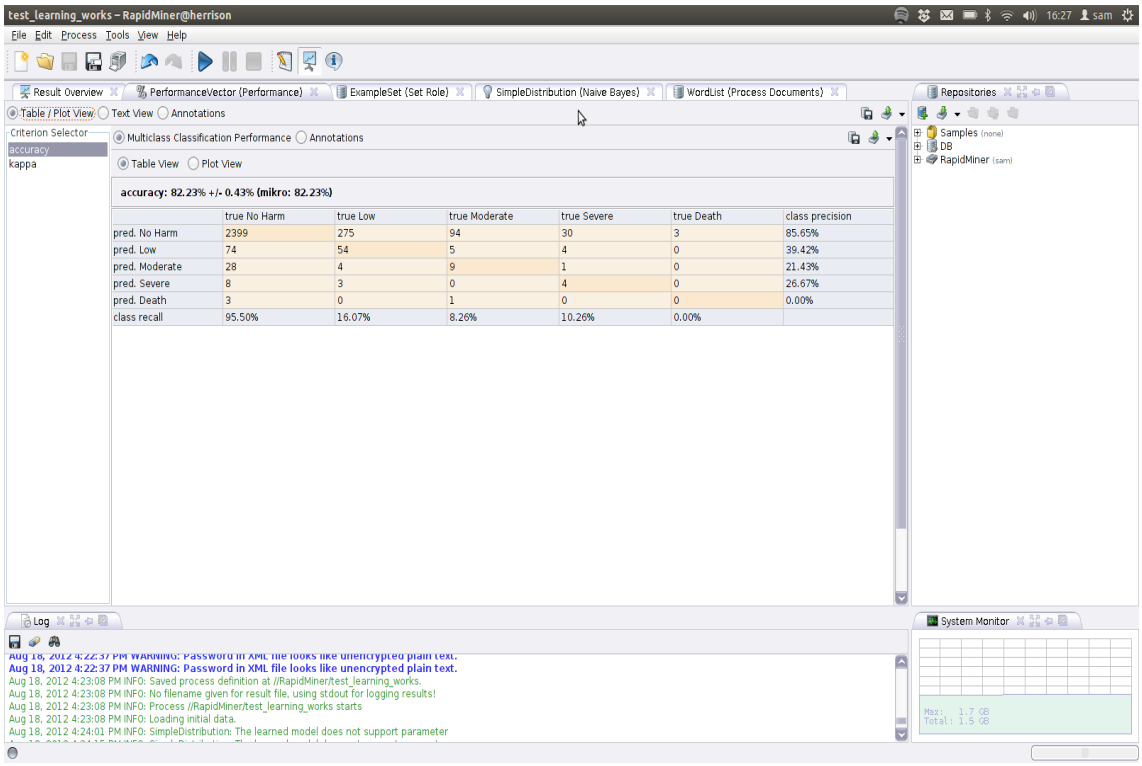


Figure 7.23: Performance of the naive bayes classifier with laplace correction for classifying degree of harm from incident description for a sample of 3000 reports.

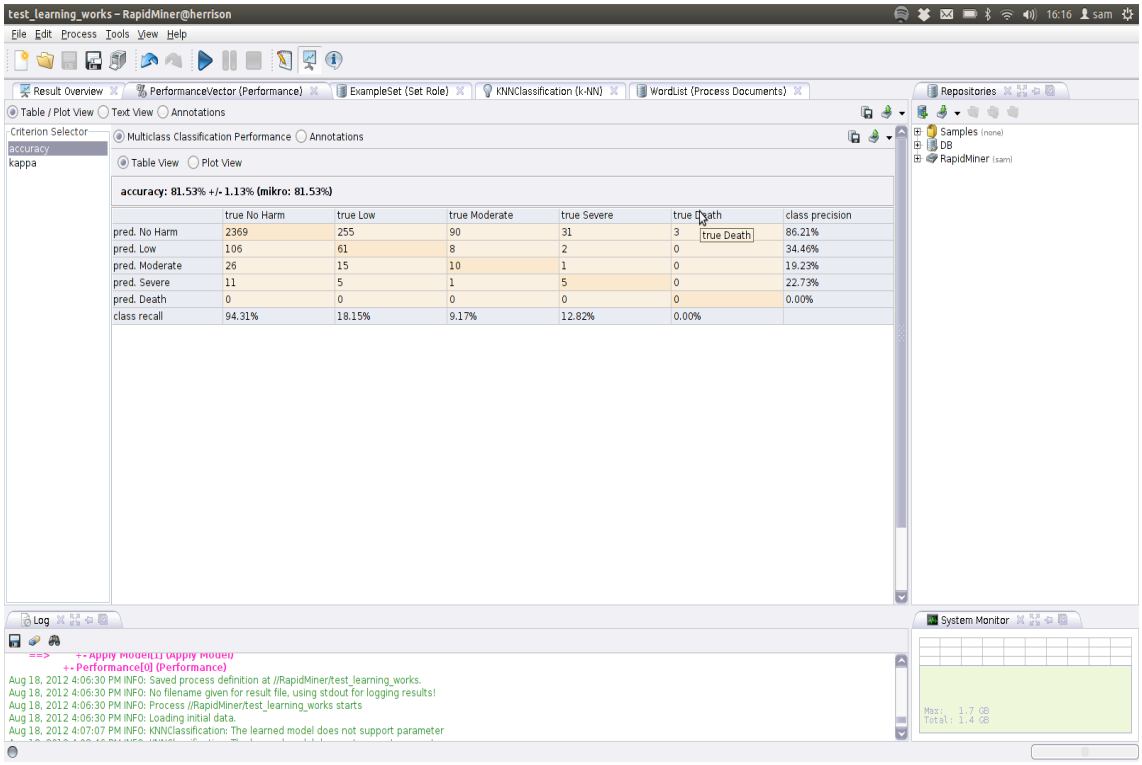


Figure 7.24: Performance of the K nearest neighbour classifier for classifiying degree of harm from incident description for a sample of 3000 reports. K=3, numerical measures with cosine similarity are used.

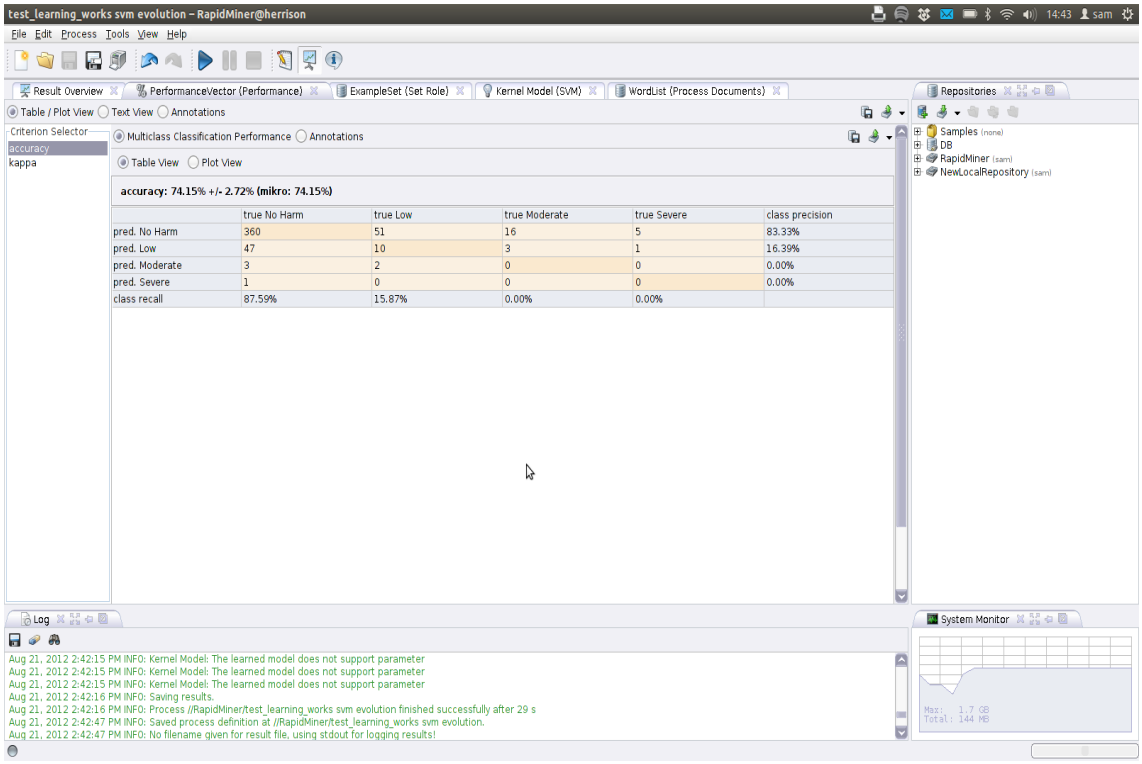


Figure 7.25: Support vector machine learner for classifiying degree of harm from incident description for a sample of 500 reports. C=1000.

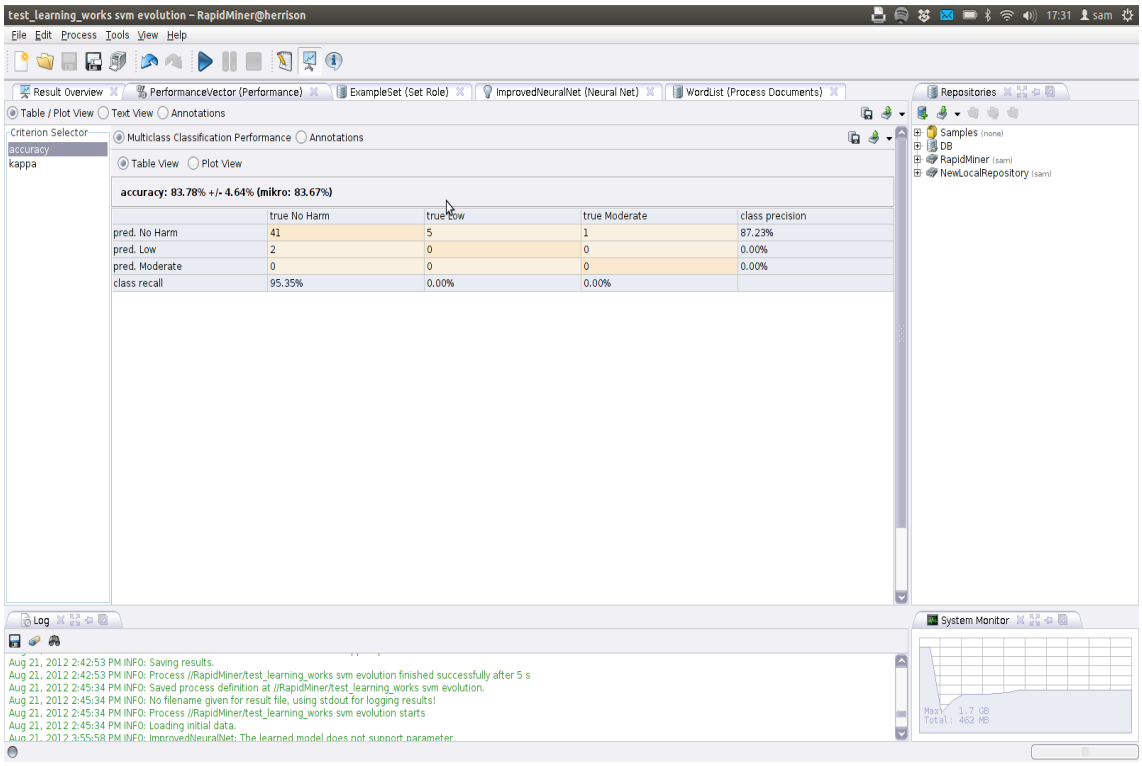


Figure 7.26: Neural network machine learner for classifiying degree of harm from incident description for a sample of 50 reports (there were major performance issues).

Chapter 8

Discussion

8.1 Results

8.2 Limitations

8.3 Evaluation of data mining

8.4 Future applications

poss applications

Discussion

lingo algorithm identifies (some) npsa alerts (e.g bleep)

Part IV

Conclusion

Chapter 9

Conclusion

9.1 Can data mining help?

9.1.1 Yes for problem topic discovery

9.1.2 Yes for association discovery

9.1.3 Yes for specific classification tasks

Part V

Appendix

computer a contrived acronym (aca) aca

Glossary

aca a contrived acronym. 60

computer is a programmable machine that receives input, stores and manipulates data, and provides output in a useful format. 60

Appendix A

Tools used

tools used SAS Excel plugin Python scripts inc brewery Carrot2 Rapid miner with plugins Solr XLS Refine LaTeX

Appendix B

Data supplement

B.1 XML files

file:///home/sam/computer-problems-clusters.xml file:///home/sam/bleep-prob.xml file:///home/sam/b
file:///home/sam/crash-prob.xml file:///home/sam/cerner-clusters.xml file:///home/sam/systemone-
clusters.xml file:///home/sam/ongoingproblems.xml

B.2 XML style sheets

bleep-prob.xsl carrot2-clusters.xsl computers.xsl example.xsl bleep.xsl computer-problems.xsl
crash-prob.xsl solrresults.xsl

Appendix C

Rapid miner files and example screen shots

C.1 Files

C.2 Screen shots

C.2.1 Process

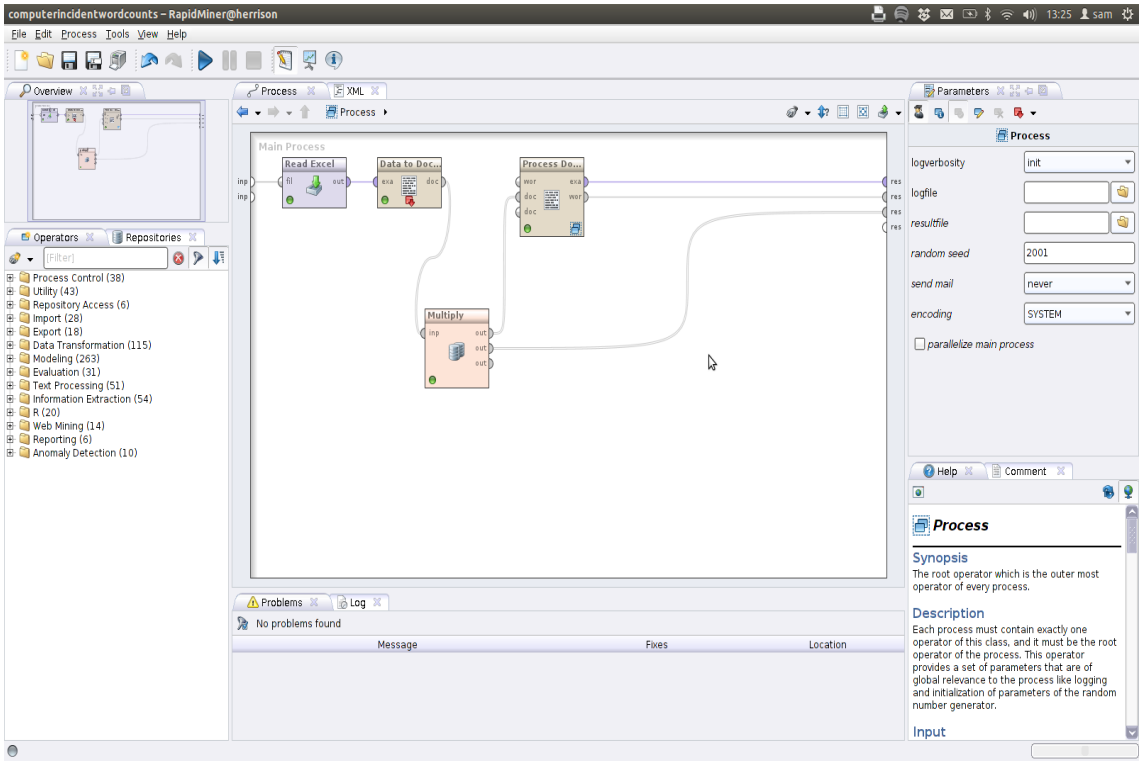


Figure C.1: Overview of Rapid Miner process to create term-frequency inverse document frequency table.

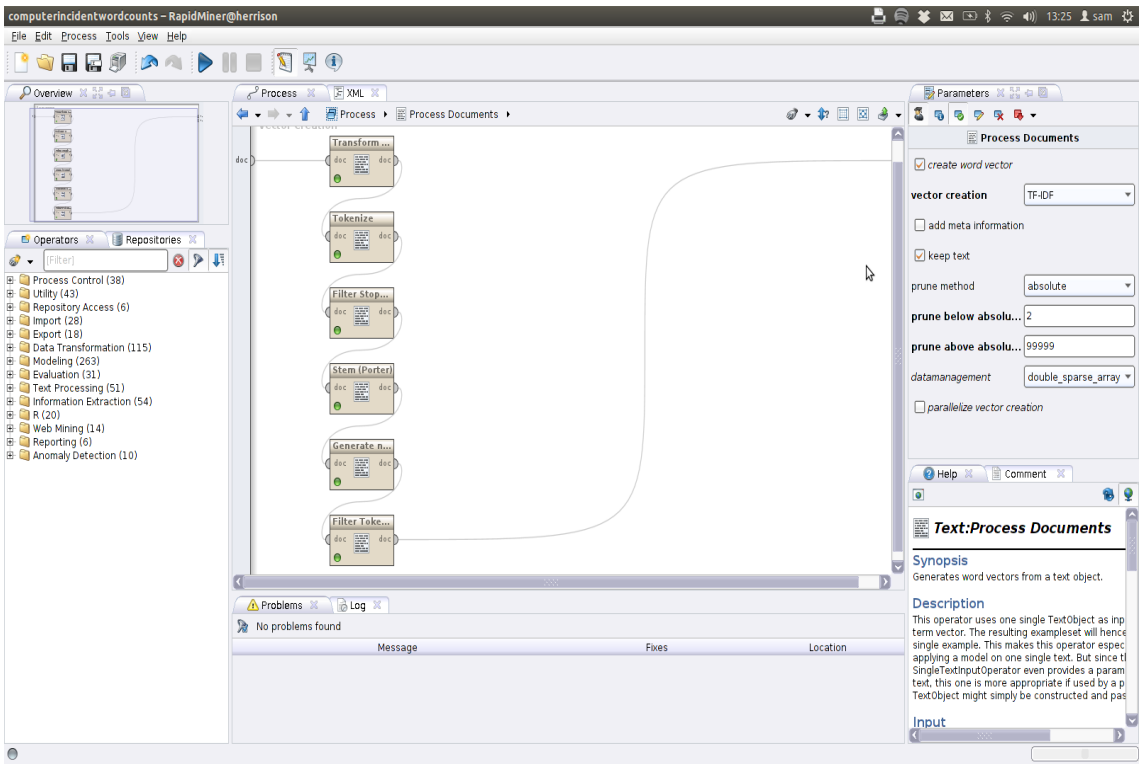


Figure C.2: Detail of Rapid Miner Document processing stages.

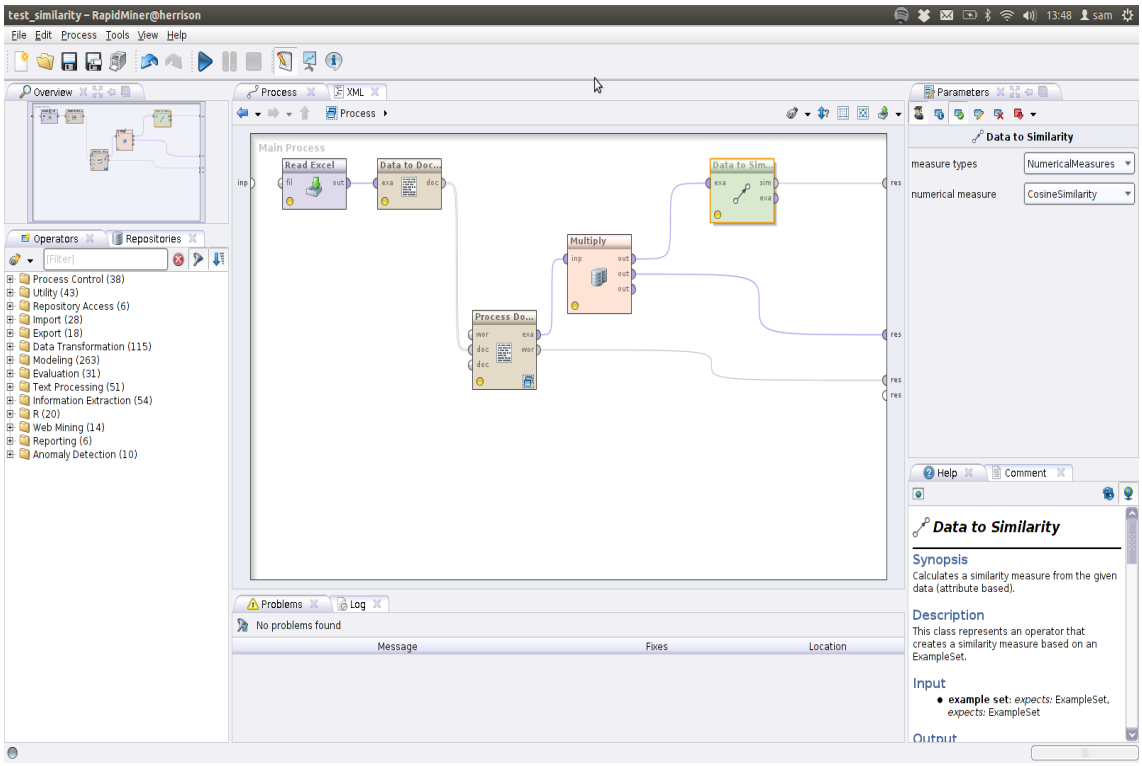


Figure C.3: Cosine similarity for sample of 50 incidents, process diagram.

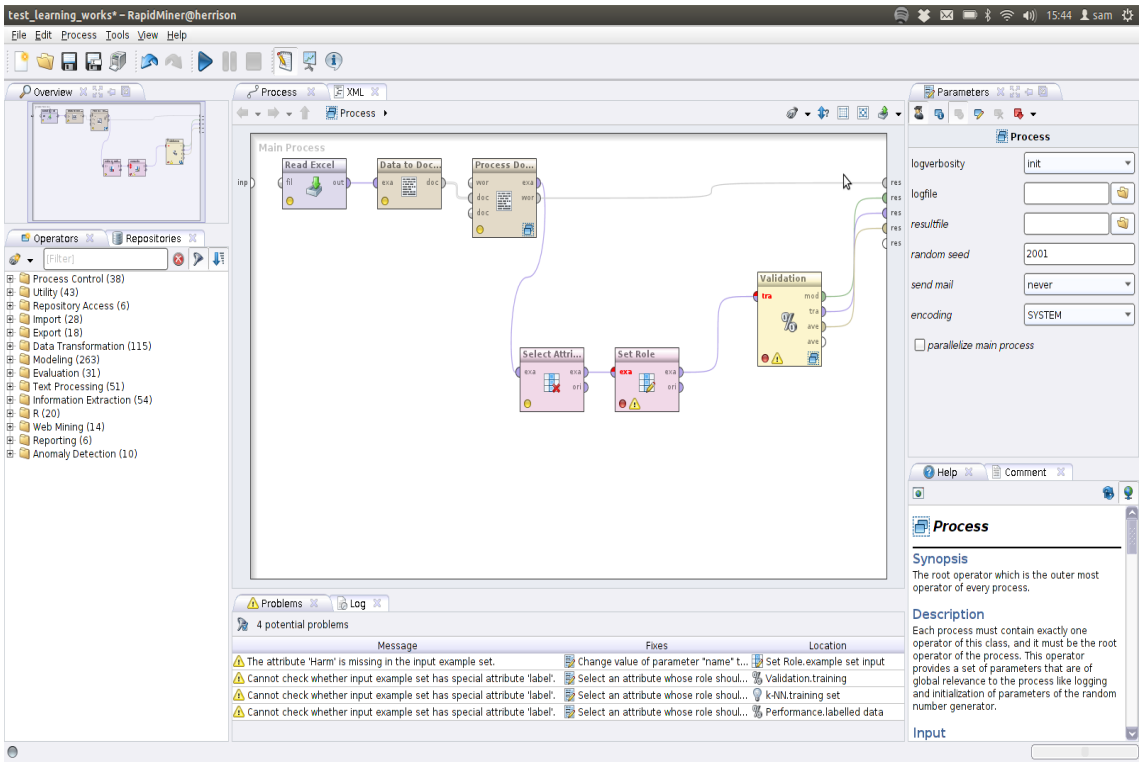


Figure C.4: Overview of Rapid Miner process for rule learning

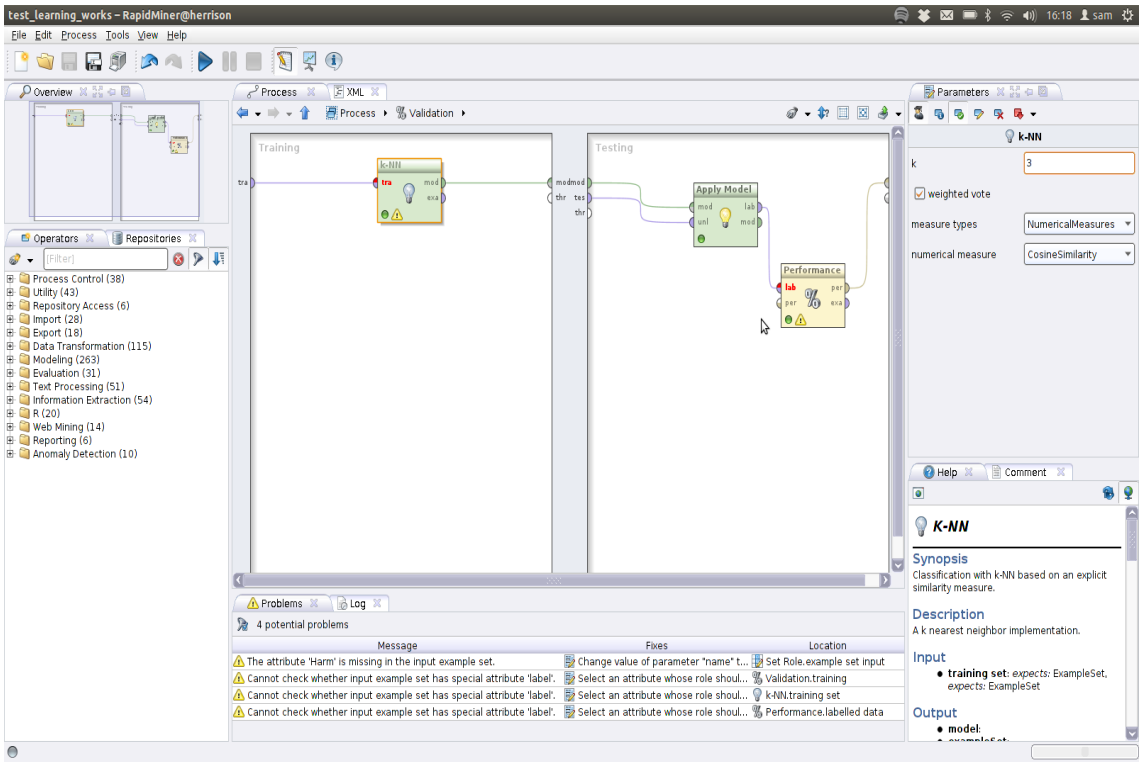


Figure C.5: Details of the Rapid Miner process for rule learning

Appendix D

Source code

D.1 Python code

D.1.1 Lexical analysis with NLTK

lexicalanalysis.py unableanalysis.py

D.1.2 File preparation for Carrot2

csvtoxml.py

Appendix E

Solr configuration files

xml inc solr config xls csv python json

Bibliography

1 CMO. An organization with a memory. *Department of Health*, 2000. URL http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4065086.pdf.

2 CMO. Building a safer nhs for patients, implementing an organization with a memory. *Department of Health*, 2001. URL http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_098565.pdf.

3 T. Stephenson. The national patient safety agency. *Archives of disease in childhood*, 90(3):226–228, 2005.

4 Report an incident. URL http://www.npsa.nhs.uk/pleaseask/experience/reportanincidentcontentid4618/.

5 A. F. Smith and R. P. Mahajan. National critical incident reporting: improving patient safety. *Br J Anaesth*, 103(5):623–625, Nov 2009. doi: 10.1093/bja/aep273. URL http://dx.doi.org/10.1093/bja/aep273.

6 S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. O'Reilly Media, 2009.