

# **Learning from patient safety incidents: can data mining help?**

Candidate identifier: TMPC1

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

Carl Jonathan Reynolds

# **Abstract**

# Contents

<b>Acknowledgments</b>	<b>8</b>
<b>List of Tables</b>	<b>9</b>
<b>List of Figures</b>	<b>10</b>
<b>I Background</b>	<b>13</b>
<b>1 Introduction</b>	<b>15</b>
1.1 The National Patient Safety Agency . . . . .	15
1.1.1 Patient safety and incident reporting . . . . .	15
1.1.2 The National Reporting and Learning Service (NRLS) . . . .	16
1.2 Data mining . . . . .	17
1.2.1 What is data mining? . . . . .	17
1.2.2 Techniques used in data mining . . . . .	17
1.3 Patient safety incidents relating to clinical information systems . . . .	17
1.4 Data mining evaluation criteria . . . . .	17
<b>2 Review of the literature</b>	<b>19</b>
2.0.1 Data mining techniques . . . . .	19
2.0.2 Data mining techniques in patient safety . . . . .	19
2.0.3 Data mining evaluation criteria . . . . .	19
<b>II Method</b>	<b>20</b>
<b>3 Data gathering</b>	<b>21</b>

3.1	The “computer problem” extract . . . . .	21
<b>4</b>	<b>Data cleaning</b>	<b>22</b>
4.1	Deduplication . . . . .	22
4.2	Establishing data quality . . . . .	22
<b>5</b>	<b>Data analysis</b>	<b>23</b>
5.1	Overview . . . . .	23
5.2	H . . . . .	24
5.2.1	Apache Solr . . . . .	24
5.2.2	NLTK . . . . .	24
5.3	Preprocessing techniques . . . . .	24
5.3.1	Stemming . . . . .	24
5.3.2	Tokenization . . . . .	24
5.3.3	Tagging . . . . .	24
5.4	Linguistic analysis . . . . .	24
5.4.1	Collocations . . . . .	24
5.4.2	Regular expression matching . . . . .	25
5.4.3	Term frequency-inverse document frequency . . . . .	25
5.4.4	Cosine similarity . . . . .	25
5.5	Association and item set mining . . . . .	25
5.5.1	Association rule learning . . . . .	25
5.6	Anomaly detection . . . . .	25
5.7	Clustering . . . . .	25
5.7.1	K-means clustering algorithm . . . . .	25
5.7.2	Lingo clustering algorithm . . . . .	25
5.8	Classification . . . . .	25
5.8.1	Classification task . . . . .	25
5.8.2	Rule induction . . . . .	26
5.8.3	Bayesian modelling . . . . .	26
5.8.4	K nearest neighbour classifier . . . . .	26
5.8.5	Support vector modelling . . . . .	26

<b>6 Tuning steps, data visualisation and interpretation</b>	<b>27</b>
6.1 Tuning steps . . . . .	27
6.1.1 Preprocessing . . . . .	27
6.1.2 Attribute selection . . . . .	27
6.1.3 Alogrithm selection . . . . .	27
6.1.4 Parameter optimization . . . . .	27
6.1.5 Measuring performance . . . . .	27
6.2 Data visualisation and interpretation . . . . .	27
<b>III Results</b>	<b>30</b>
<b>7 Results</b>	<b>31</b>
7.1 Descriptive statistics . . . . .	31
7.1.1 “computer problem” extract . . . . .	31
7.2 Identification of selected themes using a search engine and NLTK . .	35
7.3 Linguistic analysis . . . . .	37
7.3.1 Basic statistics . . . . .	37
7.3.2 Collocations . . . . .	37
7.3.3 Lexical dispersion plot . . . . .	38
7.3.4 Cumulative frequency plot . . . . .	38
7.3.5 Term frequency-inverse document frequency . . . . .	38
7.3.6 Cosine similarity . . . . .	38
7.4 Clustering . . . . .	38
7.4.1 K-means clustering algorithm . . . . .	38
7.4.2 Lingo clustering algorithm . . . . .	38
7.5 Association and item set mining . . . . .	46
7.5.1 Association rule learning . . . . .	46
7.6 Anomaly detection . . . . .	50
7.6.1 K-nearest neighbour global anomaly score . . . . .	50
7.6.2 Local correlation integeral . . . . .	50
7.6.3 Connectivity-based outlier factor . . . . .	50
7.6.4 Local outlier Factor . . . . .	50

7.7	Classification . . . . .	50
7.7.1	Rule induction . . . . .	50
7.7.2	Bayesian modelling . . . . .	51
7.7.3	K nearest neighbour classifier . . . . .	51
7.7.4	Support vector modelling . . . . .	51
7.7.5	Neural networks . . . . .	51
<b>8</b>	<b>Discussion</b>	<b>55</b>
8.1	Results . . . . .	55
8.2	Limitations . . . . .	55
8.3	Evaluation of data mining . . . . .	55
8.4	Future applications . . . . .	55
<b>IV</b>	<b>Conclusion</b>	<b>56</b>
<b>9</b>	<b>Conclusion</b>	<b>57</b>
9.1	Can data mining help? . . . . .	57
9.1.1	Yes for problem topic discovery . . . . .	57
9.1.2	Yes for association discovery . . . . .	57
9.1.3	Yes for specific classification tasks . . . . .	57
<b>V</b>	<b>Appendix</b>	<b>58</b>
<b>Glossary</b>		<b>60</b>
<b>A</b>	<b>Tools used</b>	<b>61</b>
<b>B</b>	<b>Data supplement</b>	<b>62</b>
B.1	XML files . . . . .	62
B.2	XML style sheets . . . . .	62
<b>C</b>	<b>Rapid miner files and example screen shots</b>	<b>63</b>
C.1	Files . . . . .	63
C.2	Screen shots . . . . .	63

C.2.1	Process	63
<b>D</b>	<b>Source code</b>	<b>67</b>
D.1	Python code	67
D.1.1	Lexical analysis with NLTK	67
D.1.2	File preparation for Carrot2	67
<b>E</b>	<b>Solr configuration files</b>	<b>68</b>

## 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

7.1	Incidents with identical incident descriptions (duplicates) . . . . .	31
7.2	Field names in the dataset and percentage completeness. Percentage field completeness 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. . . . .	32
7.3	Field names in the dataset and percentage completeness. Percentage field completeness 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. . . . .	33
7.4	Degree of harm caused by incidents. . . . .	33
7.5	Specialty involved in incidents, based on level one data. . . . .	34
7.6	Location of the incidents, based on care setting data. . . . .	34
7.7	Automatically generated cluster labels for the ten largest clusters identified by the lingo algorithm . . . . .	39
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. . . . .	44
7.9	Higher cluster reliability scores are achieved by applying lingo algorithm to incidents containing the word ‘bleep’ . . . . .	45

# List of Figures

5.1	Overview of the data mining process. External corpora may be used to enrich data analysis. Visualisation is also usually performed . . . . .	23
6.1	The optimize parameter 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 . . . . .	28
7.1	Search results for search on terms associated with poor system reliability.	35
7.2	Frequency distribution plot of the 50 most common words to follow the words “unable to” in the text. . . . .	36
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. . . . .	36
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). . . . .	39
7.5	Cumulative frequency plot for the 50 most frequently used words in incident reports about computer problems. . . . .	40
7.6	Each row represents a single incident report. Each column heading is a single term present in the collection of incidents. . . . .	40
7.7	Cosine similarity for sample of 50 incidents. . . . .	41
7.8	Cosine similarity for sample of 50 incidents, ISOM visualisation . . .	41
7.9	Cosine similarity for sample of 50 incidents, circle visualisation . . .	42
7.10	Centroid plot view of K means clustering of a sample of 500 incident reports. . . . .	42

7.11	Each row represents a single incident report. Each column heading is a single term present in the collection of incidents. . . . .	43
7.12	Each row represents a single incident report. Each column heading is a single term present in the collection of incidents. . . . .	43
7.13	Each row represents a single incident report. Each column heading is a single term present in the collection of incidents. . . . .	44
7.14	Unsupervised association rule learning results for all incidents. . . . .	46
7.15	Unsupervised association rule learning results for all incidents, graph visualisation. . . . .	47
7.16	Unsupervised association rule learning results for all incidents, text summary. . . . .	47
7.17	Centroid plot view of K means clustering of a sample of 500 incident reports. . . . .	48
7.18	Centroid plot view of K means clustering of a sample of 500 incident reports. . . . .	48
7.19	Centroid plot view of K means clustering of a sample of 500 incident reports. . . . .	49
7.20	Centroid plot view of K means clustering of a sample of 500 incident reports. . . . .	49
7.21	Each row represents a single incident report. Each column heading is a single term present in the collection of incidents. . . . .	50
7.22	Top terms identified as predicting death by naive bayes classifier. . . . .	51
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. . . . .	52
7.24	Performance of the K nearest neighbour classifier for classifying degree of harm from incident description for a sample of 3000 reports. K=3, numerical measures with cosine similarity are used. . . . .	53
7.25	Support vector machine learner for classifying degree of harm from incident description for a sample of 500 reports. C=1000. . . . .	54

---

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). . . . .	54
C.1 Overview of Rapid Miner process to create term-frequency inverse document frequency table. . . . .	64
C.2 Detail of Rapid Miner Document processing stages. . . . .	64
C.3 Cosine similarity for sample of 50 incidents, process diagram. . . . .	65
C.4 Overview of Rapid Miner process for rule learning . . . . .	65
C.5 Details of the Rapid Miner process for rule learning . . . . .	66

# **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<sup>1</sup> and Building a safer NHS<sup>2</sup> set out the urgent need for greater organizational learning from safety incidents to make the NHS safer for patients.<sup>3</sup>

#### **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.<sup>4</sup> The form serves as a means for the healthcare professional to inform his or her institution that an incident has, or has almost, occurred.

**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<sup>5</sup>. 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 field of research of extracting new and interesting knowledge out of (masses of) information, 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

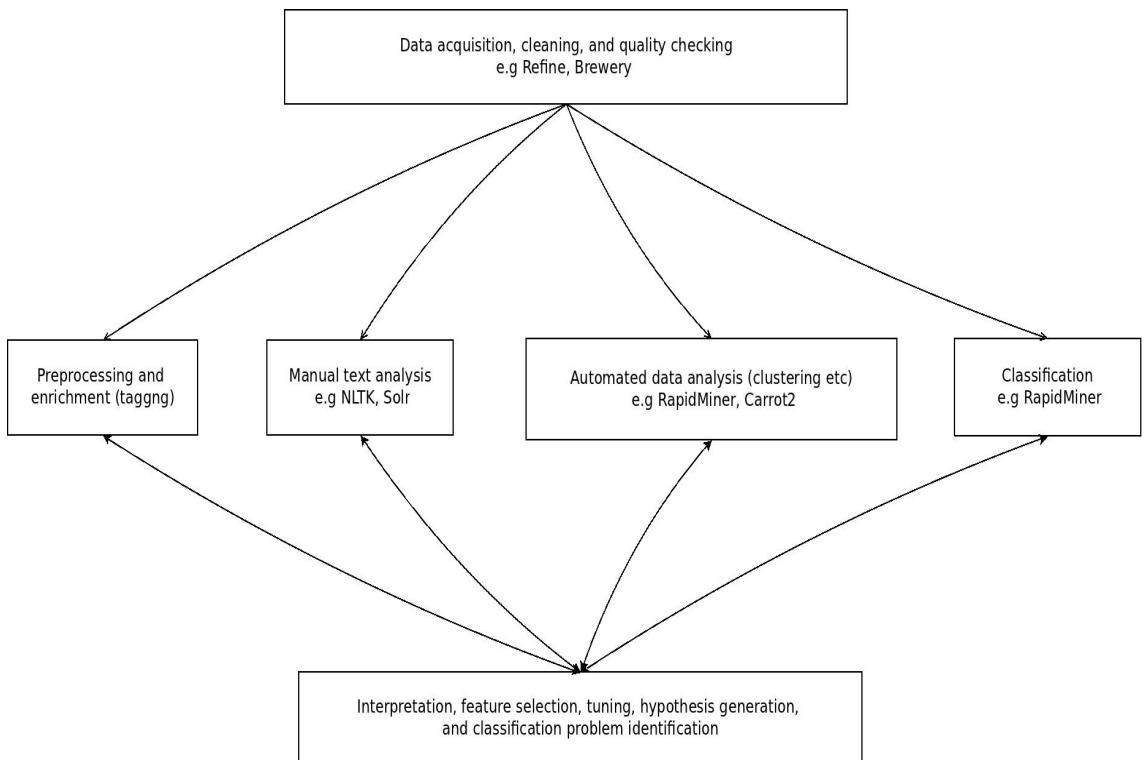


Figure 5.1: Overview of the data mining process. External corpora may be used to enrich data analysis. Visualisation is also usually performed

## 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 frequency 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 occurring 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 classify the degree 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 Algorithm 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 -*c* refine -*c* format for processing -*c* process -*c* visualization  
Data source is SAS search for all computer incidents in NRLS = 7000 -*c* convert  
to csv  
process for analysis platform

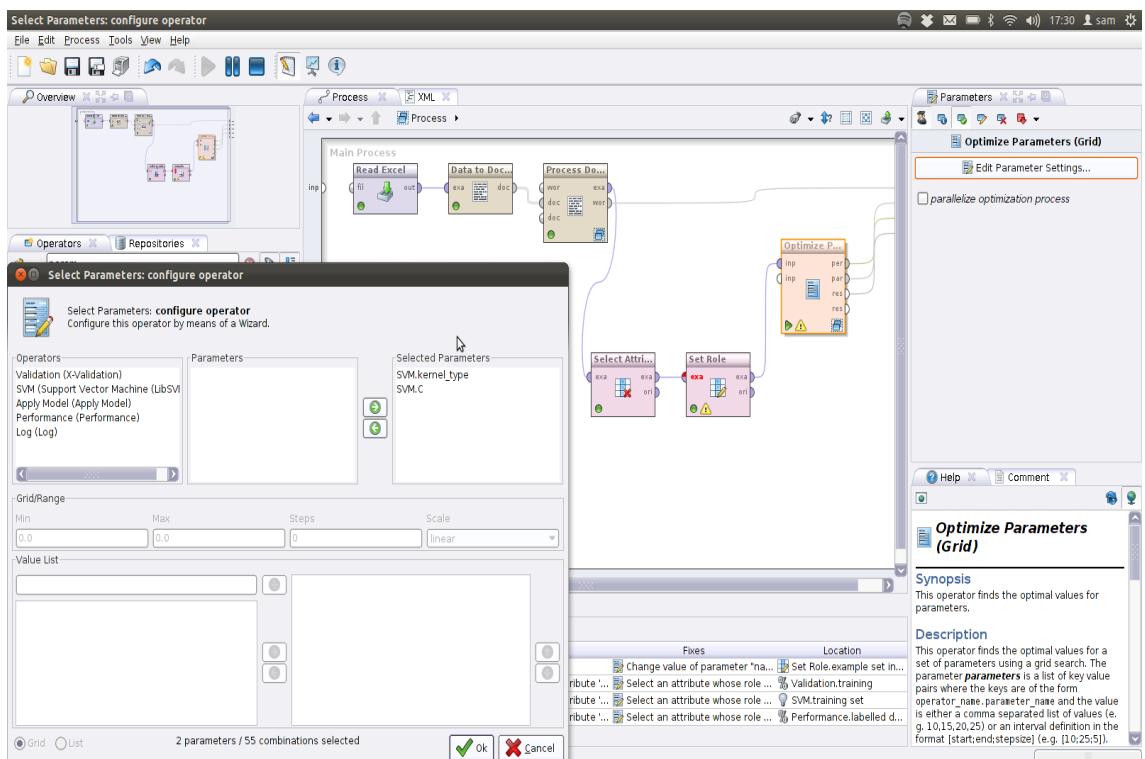


Figure 6.1: The optimize parameter process can be set to iterate through all possible parameter settings in a given range to identify the settings resulting in the best classifier performance

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 (generate categories/labels e.g with Carrot2) - classification (define categories/labels e.g GATE, Weka) independent variable is free text dependent 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 clinical 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 completeness 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 statistics

##### What was the degree of harm?

##### Which specialties were involved?

<b>Field name</b>	<b>Field completeness(%)</b>
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 Nrsls	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 completeness. Percentage field completeness 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.

<b>Field name</b>	<b>Field completeness(%)</b>
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 completeness. Percentage field completeness 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.

<b>What was the degree of harm?</b>	<b>N</b>
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 disabilities	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.

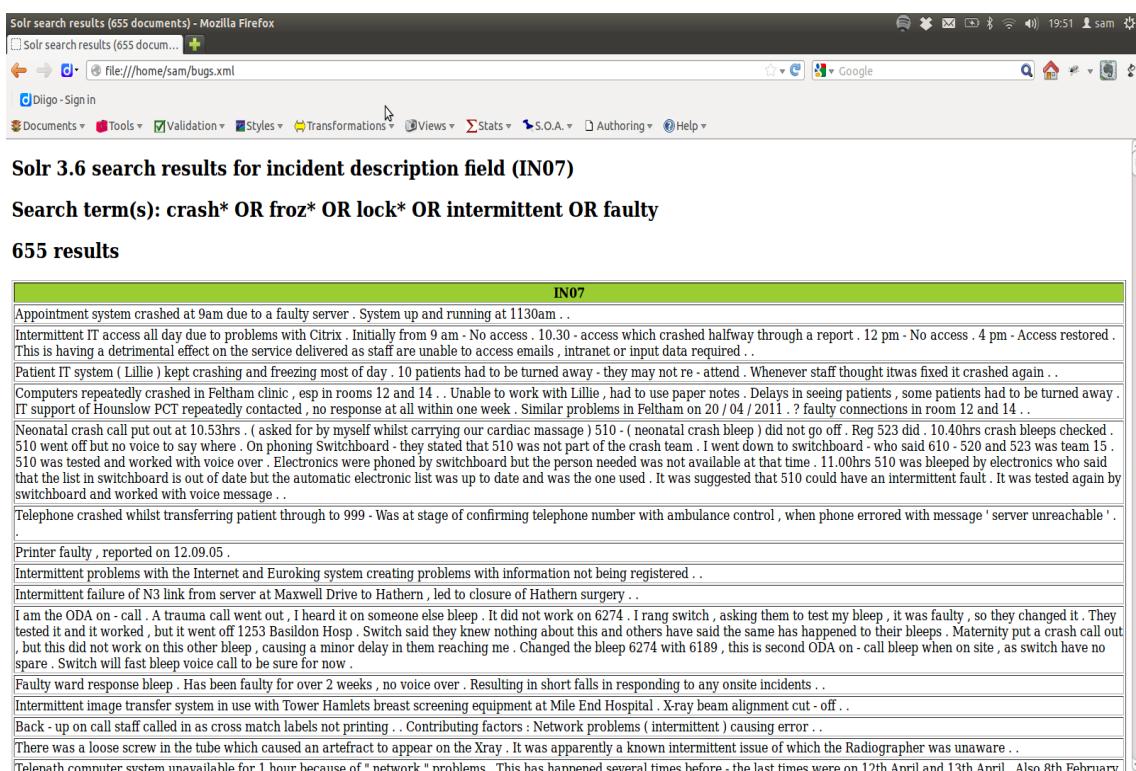
Specialty	N
Acute / general hospital	5054
Ambulance service	121
Community and general dental service	34
Community nursing, medical and therapy service (incl. community hospital)	1614
General practice	92
Learning disabilities service	17
Mental health service	182

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.



**Solr 3.6 search results for incident description field (IN07)**

**Search term(s):** crash\* OR froz\* OR lock\* OR intermittent OR faulty

**655 results**

IN07
Appointment system crashed at 9am due to a faulty server . System up and running at 1130am ..
Intermittent IT access all day due to problems with Citrix . Initially from 9 am - No access . 10.30 - access which crashed halfway through a report . 12 pm - No access . 4 pm - Access restored . This is having a detrimental effect on the service delivered as staff are unable to access emails , intranet or input data required ..
Patient IT system ( Lillie ) kept crashing and freezing most of day . 10 patients had to be turned away - they may not re - attend . Whenever staff thought it was fixed it crashed again ..
Computers repeatedly crashed in Feltham clinic , esp in rooms 12 and 14 . Unable to work with Lillie , had to use paper notes . Delays in seeing patients , some patients had to be turned away . IT support of Hounslow PCT repeatedly contacted , no response at all within one week . Similar problems in Feltham on 20 / 04 / 2011 . ? faulty connections in room 12 and 14 ..
Neonatal crash call put out at 10.53hrs . ( asked for by myself whilst carrying our cardiac massage ) 510 - ( neonatal crash bleep ) did not go off . Reg 523 did . 10.40hrs crash bleeps checked . 510 went off but no voice to say where . On phoning Switchboard - they stated that 510 was not part of the crash team . I went down to switchboard - who said 610 - 520 and 523 was team 15 . 510 was tested and worked with voice over . Electronics were phoned by switchboard but the person needed was not available at that time . 11.00hrs 510 was bleeped by electronics who said that the list in switchboard is out of date but the automatic electronic list was up to date and was the one used . It was suggested that 510 could have an intermittent fault . It was tested again by switchboard and worked with voice message ..
Telephone crashed whilst transferring patient through to 999 . Was at stage of confirming telephone number with ambulance control , when phone errored with message ' server unreachable ' .
Printer faulty , reported on 12.09.05 .
Intermittent problems with the Internet and Euroking system creating problems with information not being registered ..
Intermittent failure of N3 link from server at Maxwell Drive to Hatherne , led to closure of Hatherne surgery ..
I am the ODA on - call . A trauma call went out . I heard it on someone else bleep . It did not work on 6274 . I rang switch , asking them to test my bleep , it was faulty , so they changed it . They tested it and it worked , but it went off 1253 Basildon Hosp . Switch said they knew nothing about this and others have said the same has happened to their bleeps . Maternity put a crash call out , but this did not work on this other bleep , causing a minor delay in them reaching me . Changed the bleep 6274 with 6189 , this is second ODA on - call bleep when on site , as switch have no spare . Switch will fast bleep voice call to be sure for now .
Faulty ward response bleep . Has been faulty for over 2 weeks , no voice over . Resulting in short falls in responding to any onsite incidents ..
Intermittent image transfer system in use with Tower Hamlets breast screening equipment at Mile End Hospital . X-ray beam alignment cut - off ..
Back - up on call staff called in as cross match labels not printing .. Contributing factors : Network problems ( intermittent ) causing error ..
There was a loose screw in the tube which caused an artefact to appear on the Xray . It was apparently a known intermittent issue of which the Radiographer was unaware ..
Telomath computer system unavailable for 1 hour because of " network " problems . This has happened several times before . The last times were on 17th April and 13th April . Also 8th February ..

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 signs and ventilation on Nicu and Picu freezing over past 24hrs . ICT phoned x2 and team in Vienna investi-

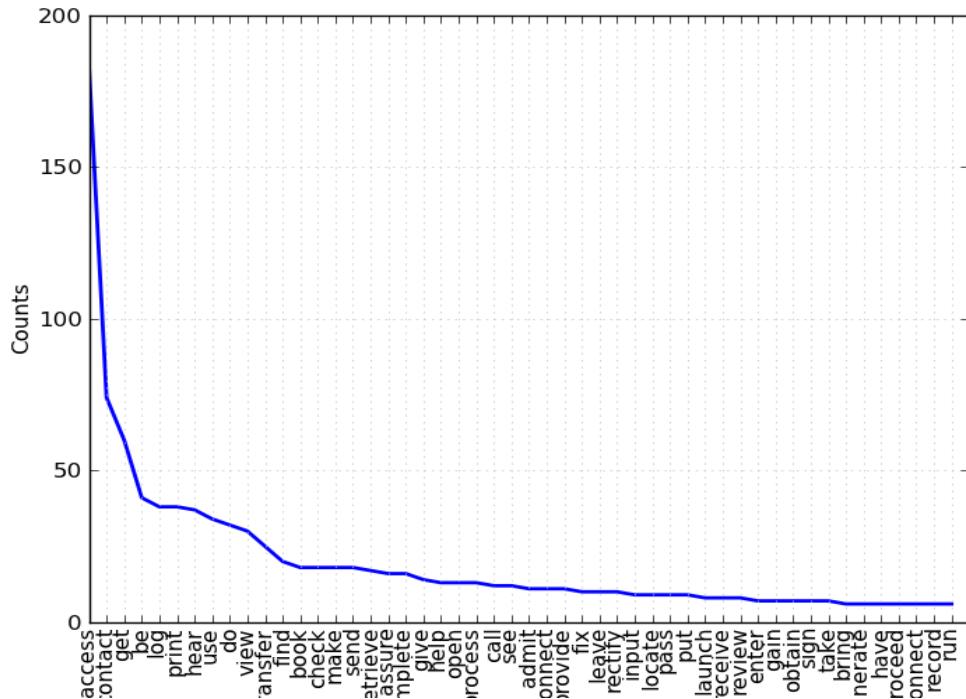


Figure 7.2: Frequency distribution plot of the 50 most common words to follow the words “unable to” in the text.

Solr search results (996 documents) - Mozilla Firefox

Solr search results (996 documents)

file:///home/sam/ongoingproblems.xml

Diigo - Sign in

Documents Tools Validation Styles Transformations Views Stats S.O.A. Authoring Help

**Solr 3.6 search results for incident description field (IN07)**

**Search term(s): ongoing OR recurrent OR recurring OR already OR again**

**996 results**

IN07
Zebra path label printer works intermittently - already escalated .
Record of EPR on / off Line . 05 / 05 / 06 Went off at 09:45hrs . Record of time back on not recorded .
Serial port 299 in Theatre 3 unable to link Faxitron to PACS system .. Theatre 3 Serial Port 299 Serial Port 300 - Not working . Unable to transfer operative image to PACS system .
Paed bleeped twice from ex 89829 to attend bradycardia and instrumental delivery in room 5 no answer to bleep request , therefore bleeped twice from ex 85140 still no reply . NICU extension rang 85644 this number engaged , extension no 85041 no reply extension 85644 tried again ANNP contacted . ANNP stated bleep had not gone off .
UNABLE TO LOG ONTO PACS TO CHECK A PATIENTS RESULTS AS THE MAXIMUM NUMBER OF USERS HAS BEEN REACHED . THIS A RECURRING THEME ! If this is a capacity issue then I recommend that system owner is made aware and makes a bid for the additional funds to increase the capacity , if this is indeed a viable option . It would appear that if this is a common occurrence , then such representations would have already been made and if rejected than this risk was found to be manageable .
unable to log onto the computer . recurring problem .
Record of EPR going off Line . 12 / 10 / 06 Off line at 09:15hrs . Never worked all day . 13 / 10 / 06 Off line at 09:25hrs . Back on at 10:30hrs . 16 / 10 / 06 Off line at 15:00hrs . Back on at 18:00hrs . 02 / 11 / 06 Off line at 09:50hrs . Back on at 13:00hrs .
Computers on Ward 6 and 8 still have no access to PACS , PINs and ICM , despite continual reporting to IT . This is an ongoing problem , already reported as an incident . Patient discharges are being held up by this problem . Repeatedly reported to IT .
GR system did not record my dictation . A recurrent problem with GR . ( Entered from gold copy ) .
Is entered onto the system one terminal under 2 prison no however one of the no only gives access to his records if you advance the search for deleted patients . These notes contain medication issue , reception details etc and have not been merged into LIDS updated recently . Concern that there will be an error if ( photocopier cut off bottom of sheet so can read no further ) .
A urology acute patient , was admitted under the general surgical team , this seems a recurring problem in A&E and never happens the other way round ..
Attempted to get patient X-ray on medient . The right now came up but for a different patient . This is a recurring medient problem ..
P1 call picked from top of faq , attempted to call patient , answer machine said person was already on telephone and line was busy , repeatedly attempted to call patient without success , I placed the call in my advice line callback queue . I then received a call from a male health adviser to say caller had rung back and was on the line , he attempted to transfer the call to me but said the patient had hung up . I immediately rang the patient straight back and again the telephone message was still saying the line was busy , I repeatedly tried to call the patient without success . I spoke to the ct at this site ( SW ) who said to continue trying for the next few minutes , I placed the call back in my advice line callback queue and within minutes was contacted again by another health adviser ( BT ) who had got the caller on the line again , I asked her to transfer the caller straight to me which she did , this time we were connected and I apologised for the communication problems , whilst introducing myself and the service , I attempted to open the call from my advice line callback queue but BT had already taken it from my queue with a higher priority and then closed the call , I was unable to access any information on the patient and therefore was unable to safely assess him . The patients wife was very upset by this time and the

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 accurately assess their clinical stability and plan treatment accordingly . 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.<sup>6</sup>

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

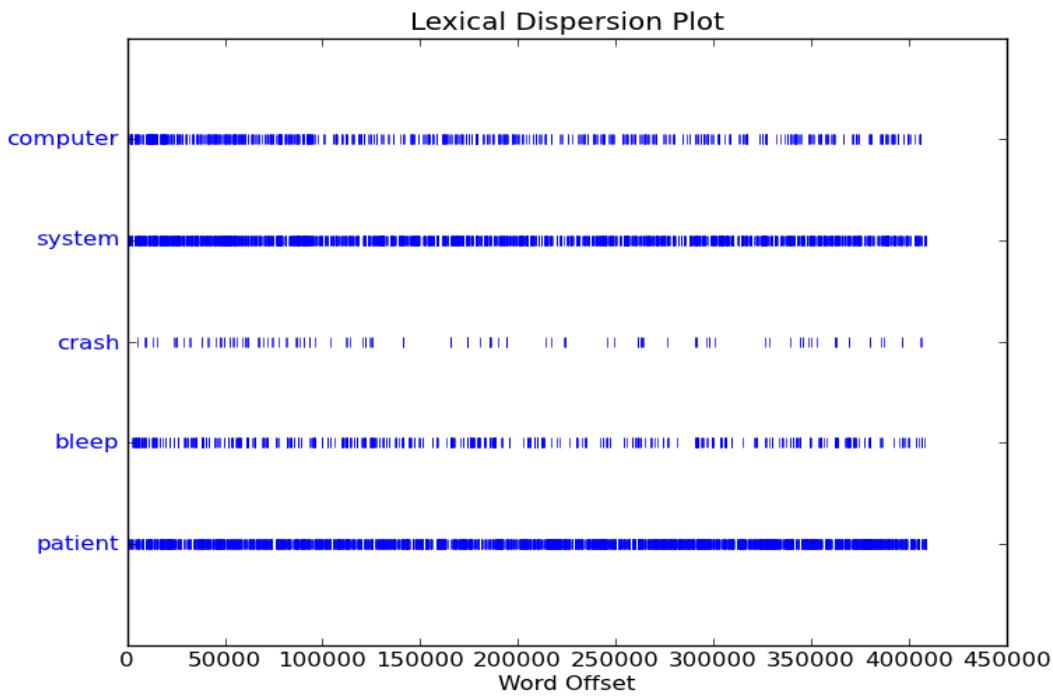


Figure 7.4: Each stripe represents an instance of a word and each row represents the entire text. If present, trends in the frequency of the use of words over time can be visualized in this way (the incident descriptions forming the text have been time ordered).

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

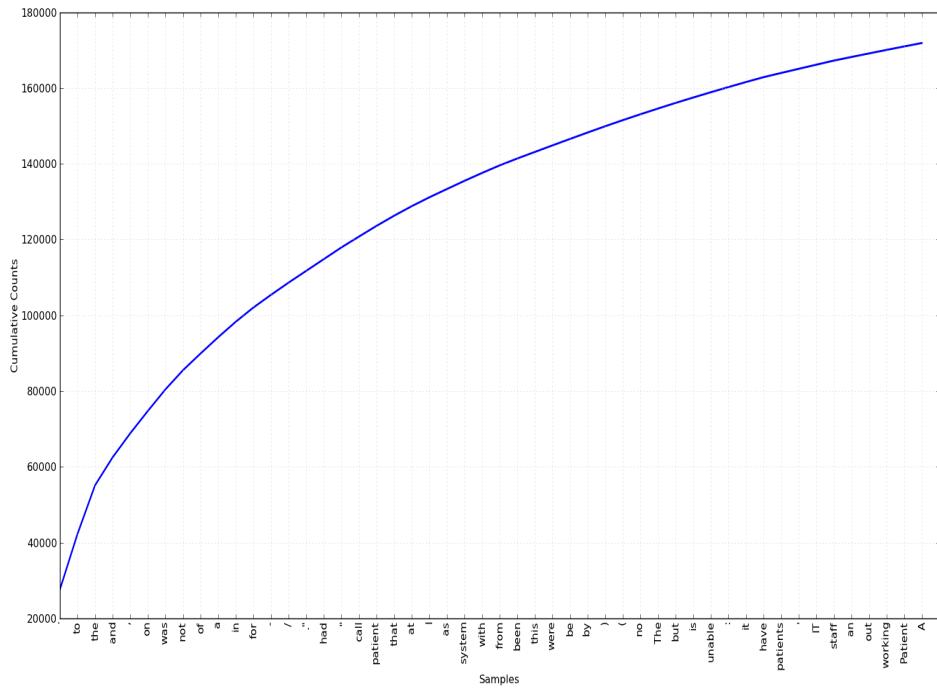


Figure 7.5: Cumulative frequency plot for the 50 most frequently used words in incident reports about computer problems.

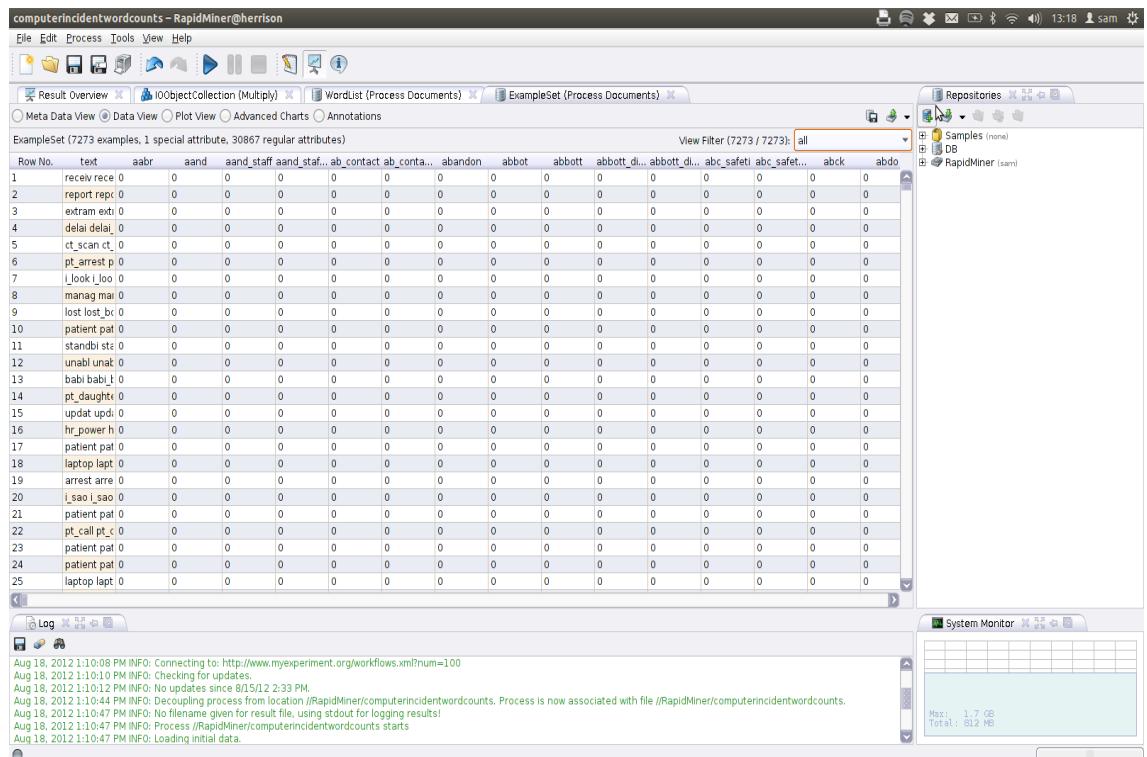


Figure 7.6: Each row represents a single incident report. Each column heading is a single term present in the collection of incidents.

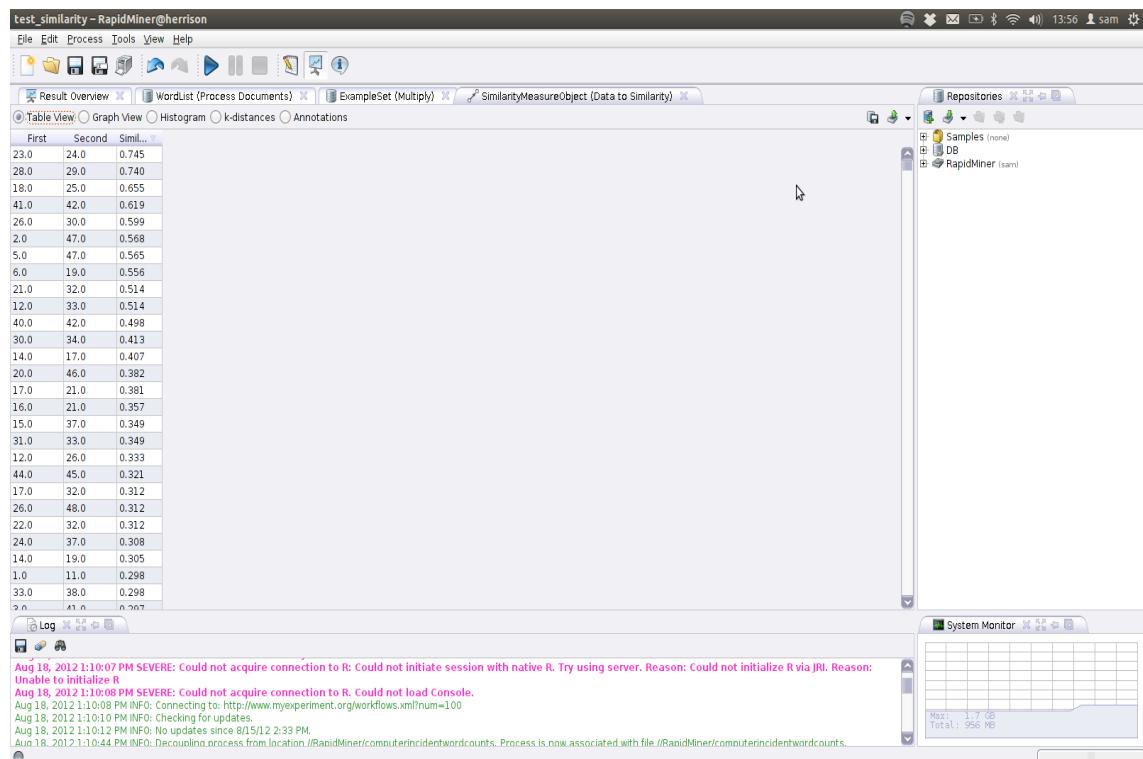


Figure 7.7: Cosine similarity for sample of 50 incidents.

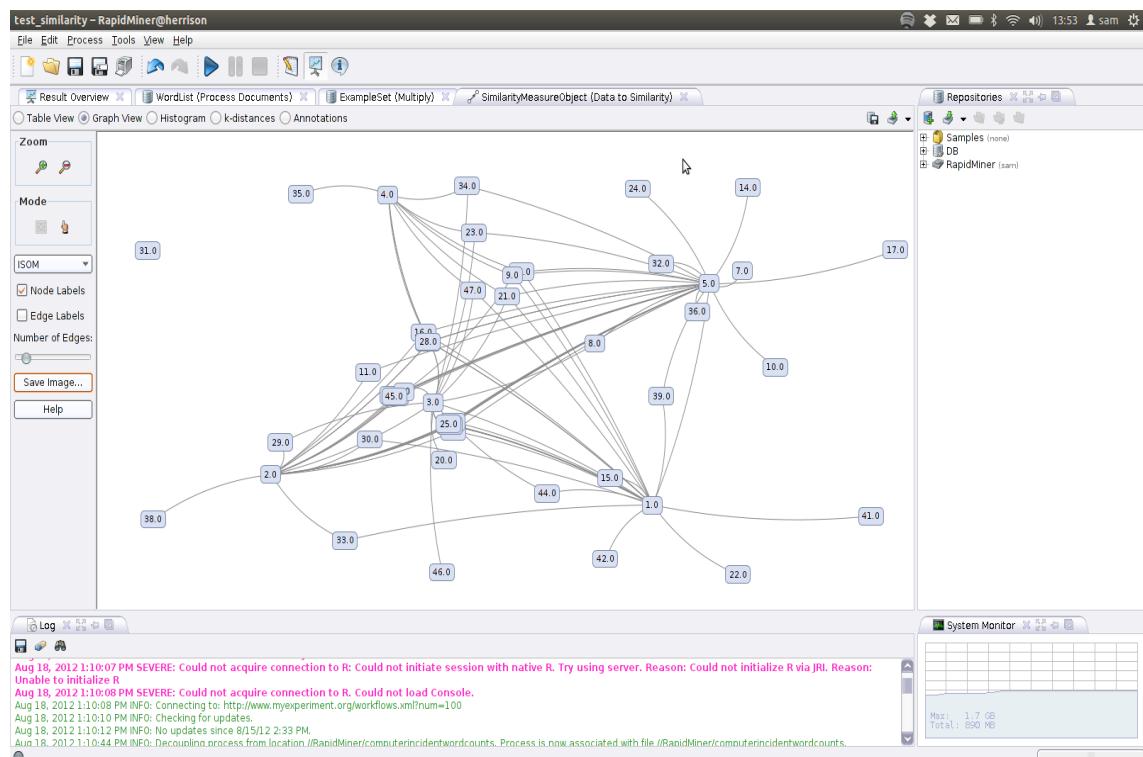


Figure 7.8: Cosine similarity for sample of 50 incidents, ISOM visualisation

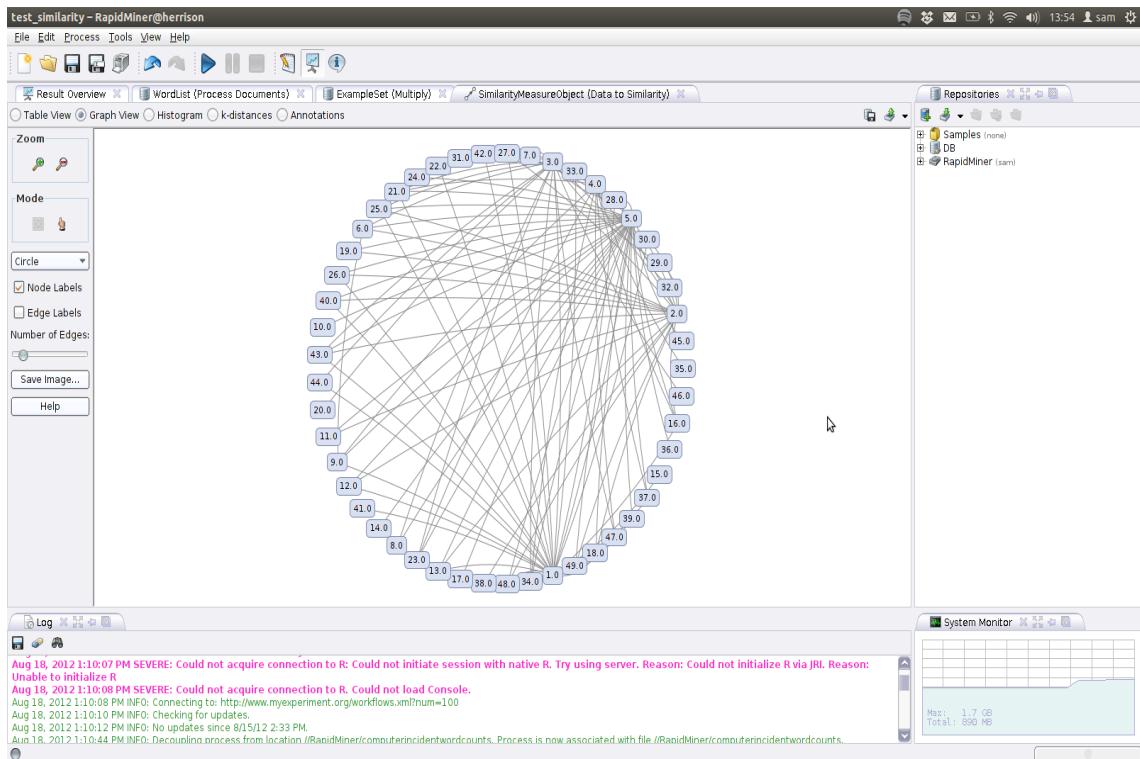


Figure 7.9: Cosine similarity for sample of 50 incidents, circle visualisation

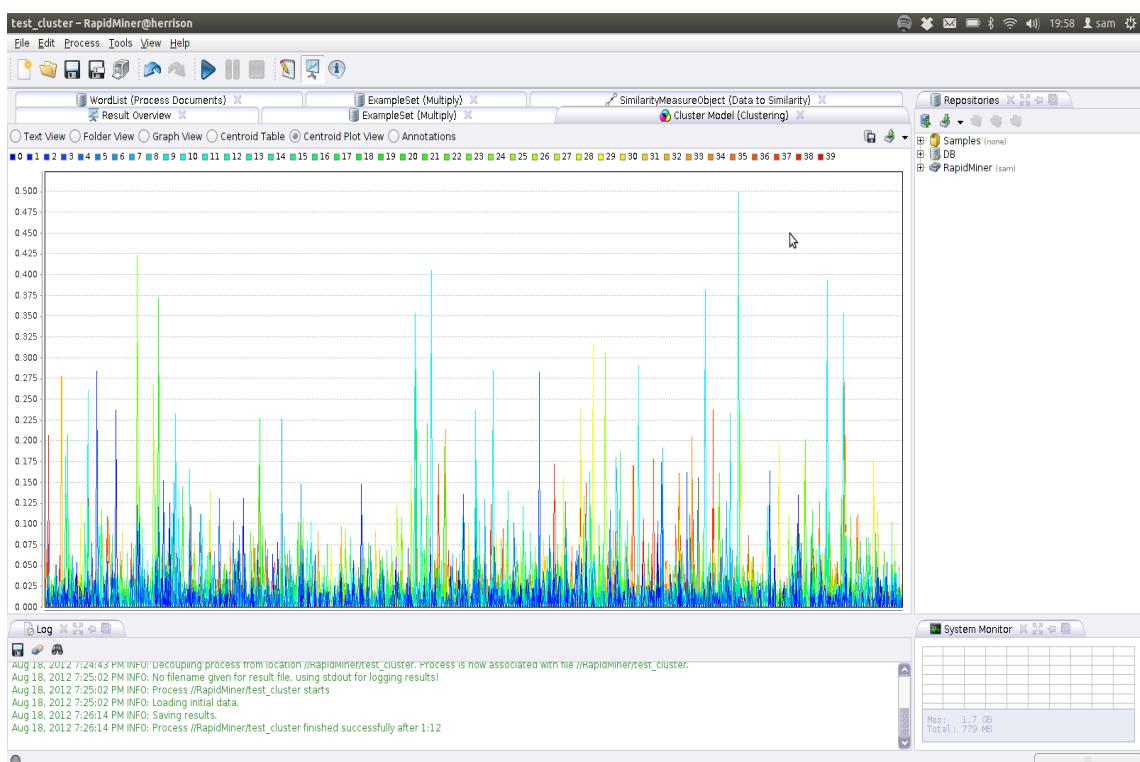


Figure 7.10: Centroid plot view of K means clustering of a sample of 500 incident reports.

Computers Unable		System Patient		Inform Patients		System Staff		Called to Ask		Phone System		Called Switchboard		Unable to Transfer Patient		Bleep Tested		Informed via Sleep System		Bleeps for Results		Bleeped Twice		New Sleep	
		Patient Called	Patient Called	Unable to Access	System Working	System Failure	System Failure	Patients Unable to Contact	Patients Unable to Contact	Answer Calls from Patient	Answer Calls from Patient	Doctors Bleep	Doctors Bleep	Unable to make or receive Calls	Unable to make or receive Calls	Bleeped by Switchboard	Bleeped by Switchboard	Pediatric Sleep	Pediatric Sleep	Reg. Bleep	Reg. Bleep	Given on Bleep	Given on Bleep	State Sleep	State Sleep
Patients Unable	Patients and Staff	Bleep Problem	Calls being Received	System Called	Staff Called	Called to Advise	Called to Advise	Contacted Bleep	Contacted Bleep	Attempted to call Patient	Attempted to call Patient	Doctors Bleep	Doctors Bleep	Unable to Access Patient	Unable to Access Patient	Bleeped Staff	Bleeped Staff	Normal Bleep	Normal Bleep	Added to Bed List	Added to Bed List	Carrying Bleep	Carrying Bleep	Anesthetic Bleep	Anesthetic Bleep
System Unable						Results System	Results System	Bleeped Worked	Bleeped Worked	Bleeped System	Bleeped System	System Used	System Used	Ward System	Ward System	3 Times	3 Times	Fast Sleep	Fast Sleep	Attempted to call Patient	Attempted to call Patient	Normal Bleep	Normal Bleep	Bleep Message	Bleep Message
Computer System Having	Information System	Caller Called	Access System	Working Unable	Number of Calls	Record System	Record System	Bleep Received	Bleep Received	Bleeped to Attend	Bleeped to Attend	Emergency Bleep	Emergency Bleep	Bleep from Switch	Bleep from Switch	Unable to Admin Patient	Unable to Admin Patient	Unable to Connect the Patient	Unable to Connect the Patient	Unable to Admin Patient	Unable to Admin Patient	Given on Bleep	Given on Bleep	New Sleep	New Sleep
Staff Unable	Reporting System		Tried Calling	Calls from Wards	Phone Unable	Team Called	Team Called	Asked to Bleep	Asked to Bleep	Answered Bleep	Answered Bleep	Phone and Bleep	Phone and Bleep	Bleep from SHO	Bleep from SHO	Unable to Admin Patient	Unable to Admin Patient	Unable to Connect the Patient	Unable to Connect the Patient	Unable to Admin Patient	Unable to Admin Patient	Normal Bleep	Normal Bleep	Bleep Failed	Bleep Failed
Time the System	Unable to Contact	Called the Service	Clinical System	System Due	Unable to make Clinical	System with Number	System with Number	Crash Bleep	Crash Bleep	Bleeped using the Theatre	Bleeped using the Theatre	Registrar Bleep	Registrar Bleep	Arrest Bleep	Arrest Bleep	Unable to Book Patient	Unable to Book Patient	Unable to Book Patient	Unable to Book Patient	Unable to Book Patient	Unable to Book Patient	Normal Bleep	Normal Bleep	Bleep Failed	Bleep Failed

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

Computers Unable		System Patient		Inform Patients		System Staff		Called to Ask		Phone System		Called Switchboard		Unable to Transfer Patient		Bleep Tested		Informed via Sleep System		Bleeps for Results		Bleeped Twice		New Sleep	
		Patient Called	Patient Called	Unable to Access	System Working	System Failure	System Failure	Patients Unable to Contact	Patients Unable to Contact	Answer Calls from Patient	Answer Calls from Patient	Doctors Bleep	Doctors Bleep	Unable to make or receive Calls	Unable to make or receive Calls	Bleeped by Switchboard	Bleeped by Switchboard	Pediatric Sleep	Pediatric Sleep	Reg. Bleep	Reg. Bleep	Given on Bleep	Given on Bleep	State Sleep	State Sleep
Patients Unable	Patients and Staff	Bleep Problem	Calls being Received	System Called	Staff Called	Called to Advise	Called to Advise	Contacted Bleep	Contacted Bleep	Attempted to call Patient	Attempted to call Patient	Doctors Bleep	Doctors Bleep	Unable to Access Patient	Unable to Access Patient	Bleeped Staff	Bleeped Staff	Normal Bleep	Normal Bleep	Added to Bed List	Added to Bed List	Carrying Bleep	Carrying Bleep	Anesthetic Bleep	Anesthetic Bleep
System Unable						Results System	Results System	Bleeped Worked	Bleeped Worked	Bleeped System	Bleeped System	System Used	System Used	Ward System	Ward System	3 Times	3 Times	Fast Sleep	Fast Sleep	Attempted to call Patient	Attempted to call Patient	Normal Bleep	Normal Bleep	Bleep Message	Bleep Message
Computer System Having	Information System	Caller Called	Access System	Working Unable	Number of Calls	Record System	Record System	Bleep Received	Bleep Received	Bleeped to Attend	Bleeped to Attend	Emergency Bleep	Emergency Bleep	Bleep from Switch	Bleep from Switch	Unable to Admin Patient	Unable to Admin Patient	Unable to Connect the Patient	Unable to Connect the Patient	Unable to Admin Patient	Unable to Admin Patient	Given on Bleep	Given on Bleep	New Sleep	New Sleep
Staff Unable	Reporting System		Tried Calling	Calls from Wards	Phone Unable	Team Called	Team Called	Asked to Bleep	Asked to Bleep	Answered Bleep	Answered Bleep	Phone and Bleep	Phone and Bleep	Bleep from SHO	Bleep from SHO	Unable to Admin Patient	Unable to Admin Patient	Unable to Connect the Patient	Unable to Connect the Patient	Unable to Admin Patient	Unable to Admin Patient	Normal Bleep	Normal Bleep	Bleep Failed	Bleep Failed
Time the System	Unable to Contact	Called the Service	Clinical System	System Due	Unable to make Clinical	System with Number	System with Number	Crash Bleep	Crash Bleep	Bleeped using the Theatre	Bleeped using the Theatre	Registrar Bleep	Registrar Bleep	Arrest Bleep	Arrest Bleep	Unable to Book Patient	Unable to Book Patient	Unable to Book Patient	Unable to Book Patient	Unable to Book Patient	Unable to Book Patient	Normal Bleep	Normal Bleep	Bleep Failed	Bleep Failed

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

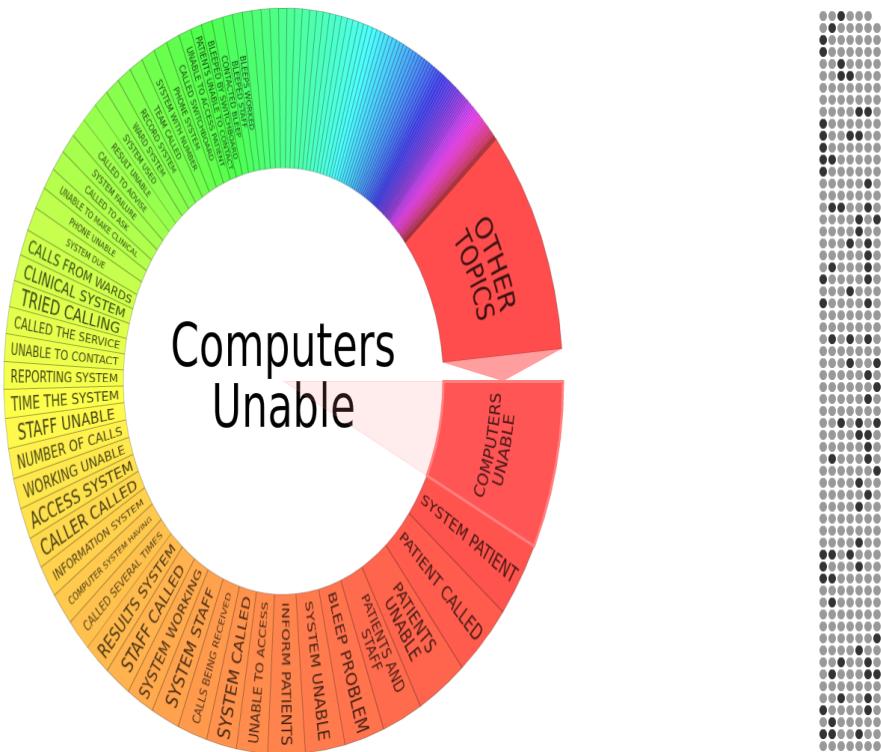


Figure 7.13: Each row represents a single incident report. Each column heading is a single term present in the collection of incidents.

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 system 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

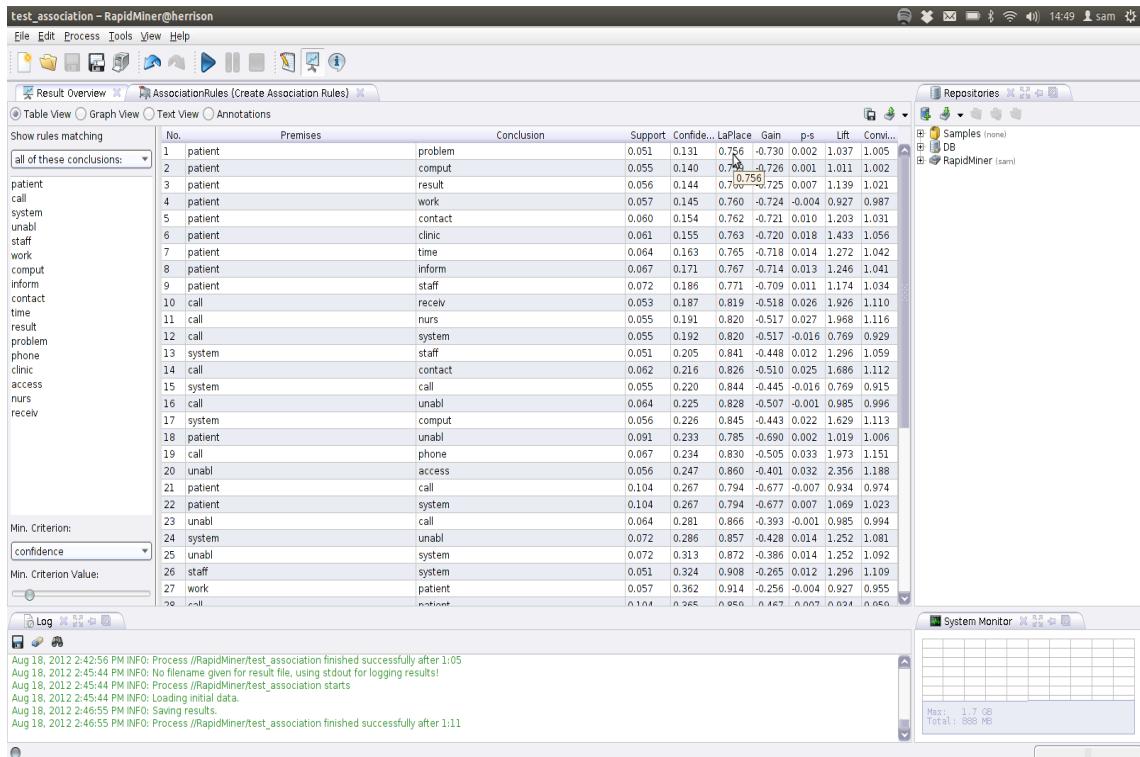


Figure 7.14: Unsupervised association rule learning results for all incidents.

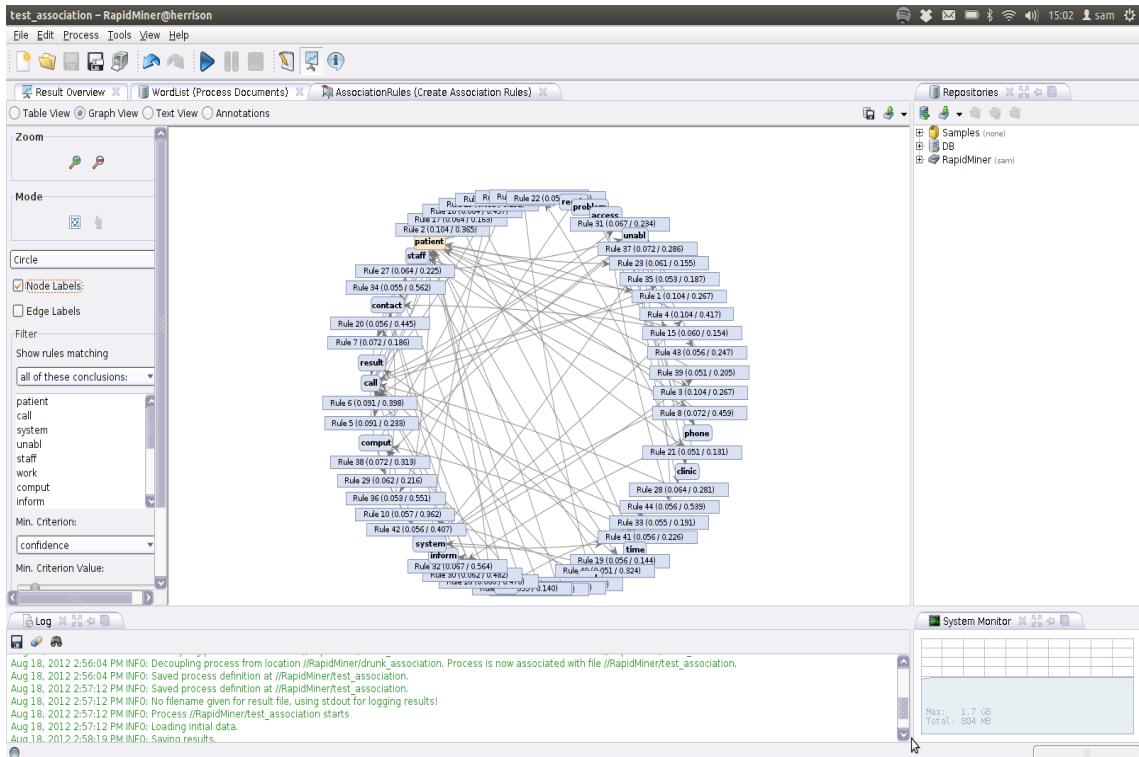


Figure 7.15: Unsupervised association rule learning results for all incidents, graph visualisation.

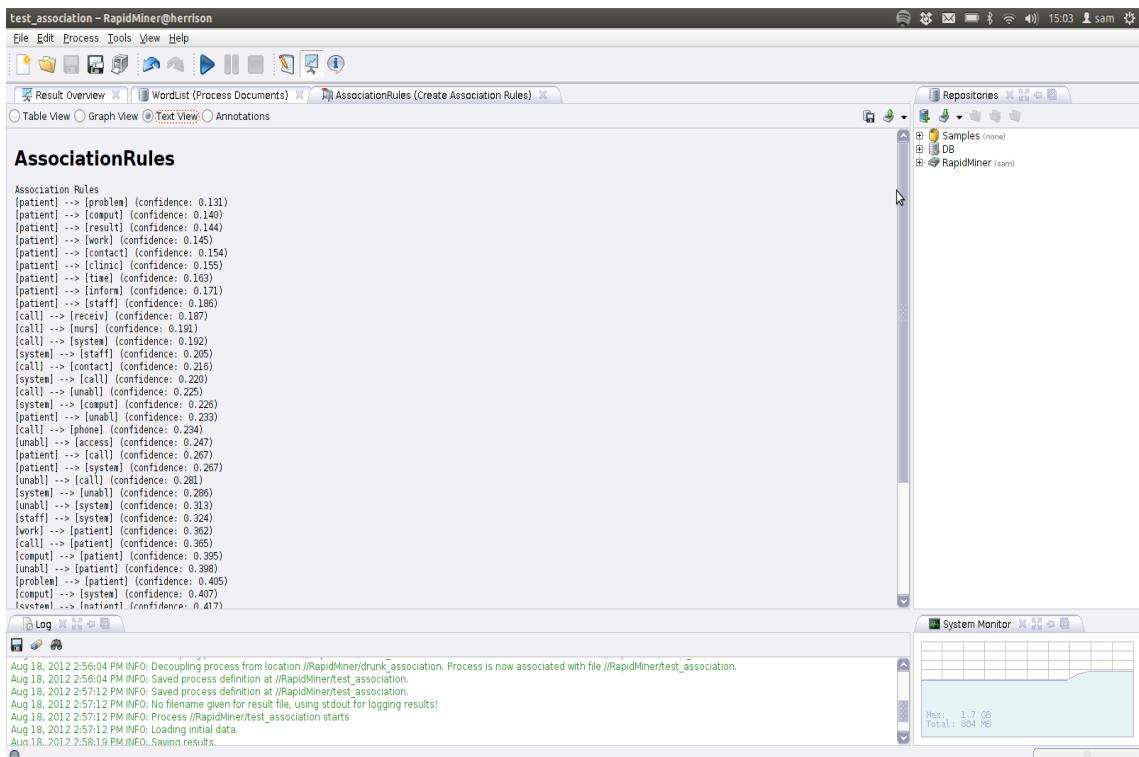


Figure 7.16: Unsupervised association rule learning results for all incidents, text summary.

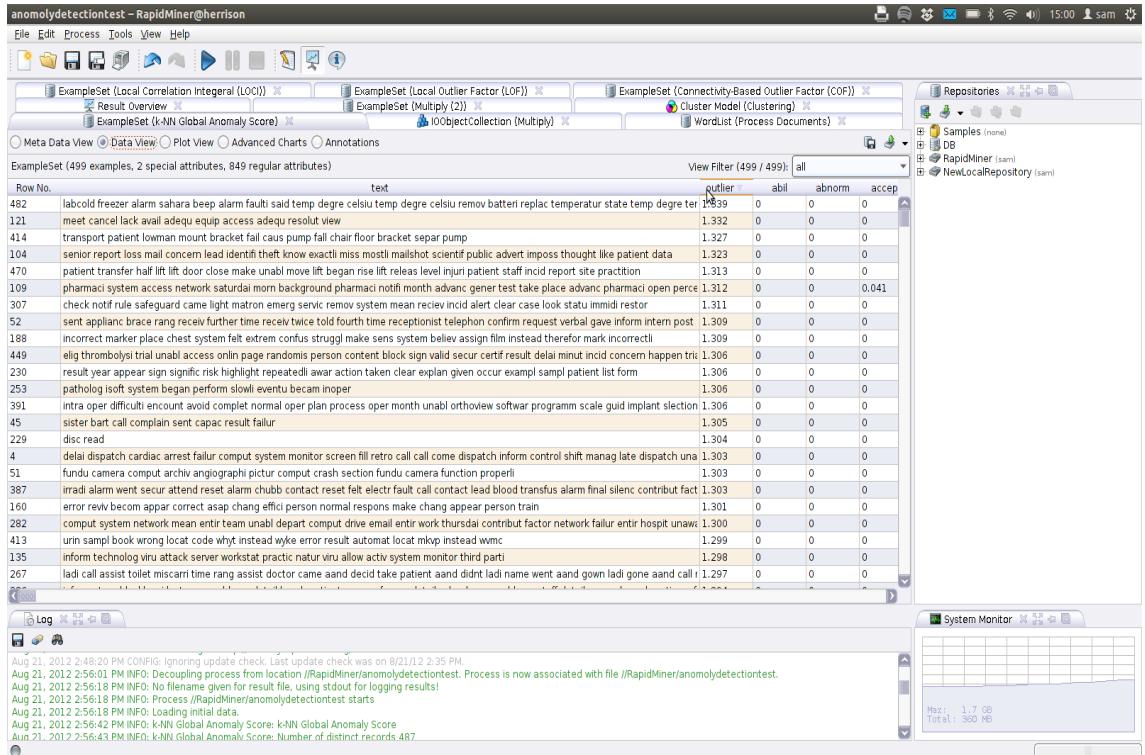


Figure 7.17: Centroid plot view of K means clustering of a sample of 500 incident reports.

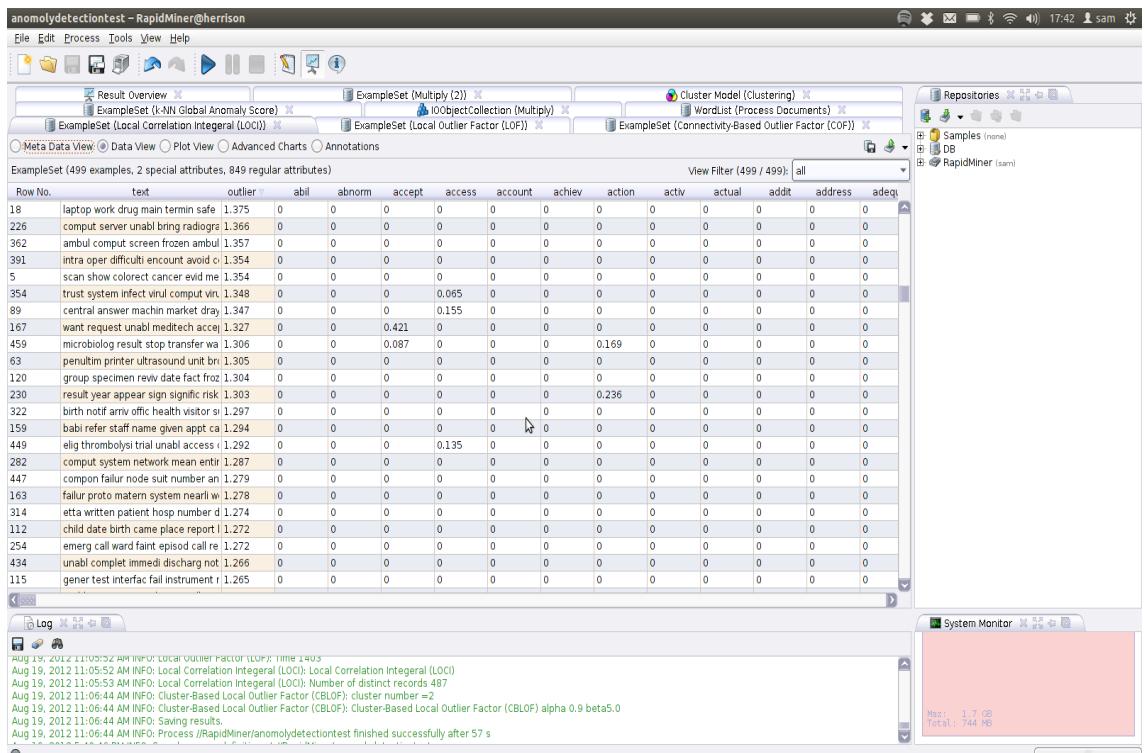


Figure 7.18: Centroid plot view of K means clustering of a sample of 500 incident reports.

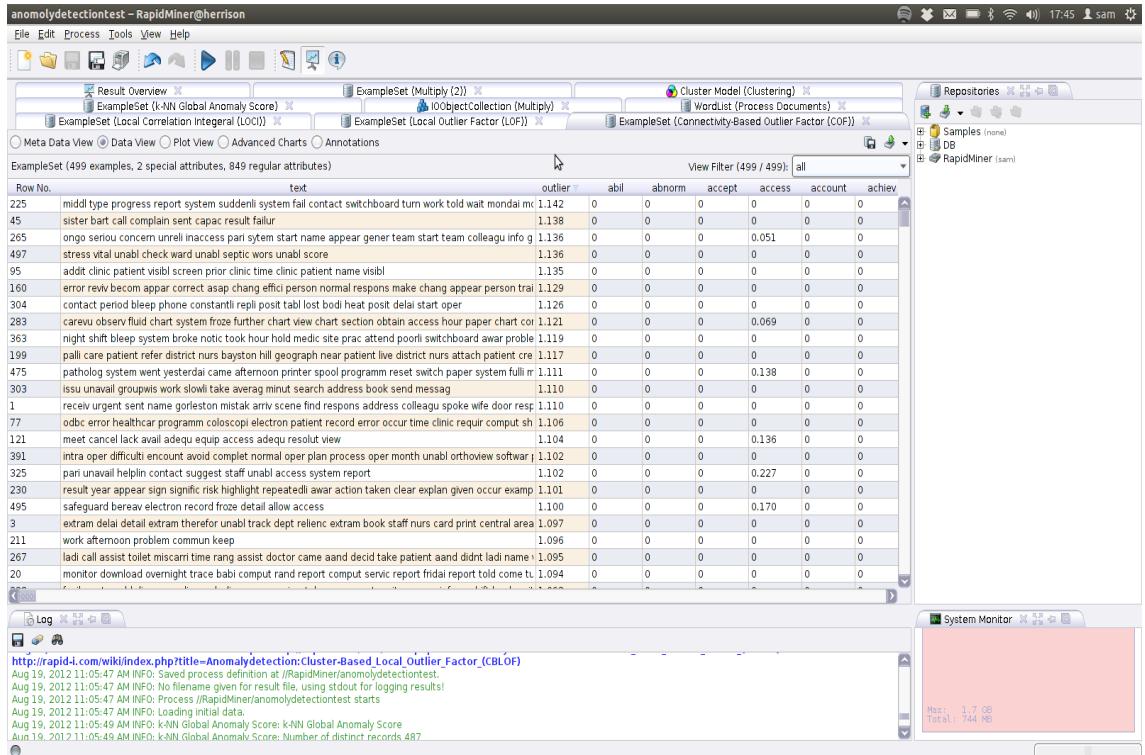


Figure 7.19: Centroid plot view of K means clustering of a sample of 500 incident reports.

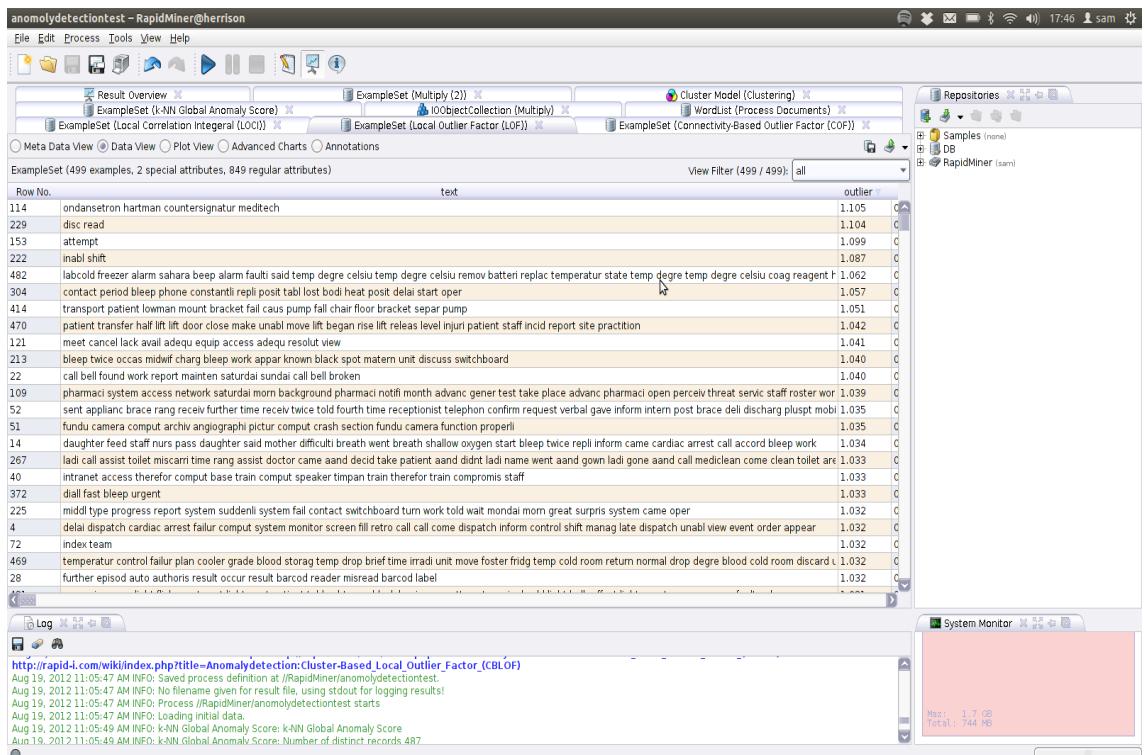


Figure 7.20: Centroid plot view of K means clustering of a sample of 500 incident reports.

## 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

The screenshot shows the RapidMiner software interface. The main window displays a 'Data View' of a dataset named 'ExampleSet'. The table has 273 rows and 20 columns. The columns are labeled: Row No., text, aabr, aand, aand\_staff, aand\_staf..., ab\_contact, ab\_abandon, abbot, abbott, abbott\_dil, abbott\_dil..., abc\_safeti, abc\_safet..., abck, abdo. The first few rows of data are:

Row No.	text	aabr	aand	aand_staff	aand_staf...	ab_contact	ab_abandon	abbot	abbott	abbott_dil	abbott_dil...	abc_safeti	abc_safet...	abck	abdo
1	receiv rece 0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	report rept 0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	extram ext 0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	delai delai_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	ct_scan ct_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	pt_arrest p_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	i_look i_loo_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	manag man_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	lost lost_b_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	patient pat_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	standbi st_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	unabl unat_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	babi babi_1_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	pt_daughtt_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	updat upd_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	hr_power h_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	patient pat_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	laptop lapt_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	arrest arre_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	i_sao i_sao_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	patient pat_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	pt_call pt_c_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	patient pat_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	patient pat_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	laptop lapt_0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

The bottom right corner shows a 'System Monitor' window displaying log messages and system statistics.

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

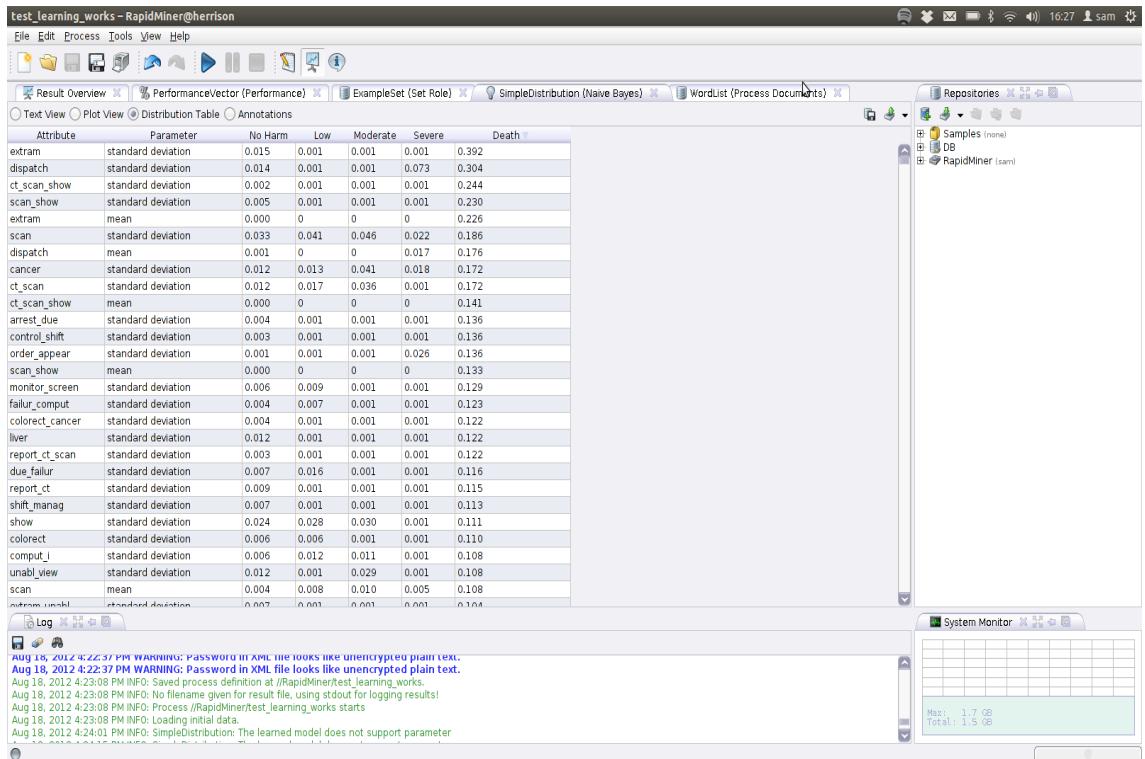


Figure 7.22: Top terms identified as predicting death by naive bayes classifier.

### 7.7.3 K nearest neighbour classifier

### 7.7.4 Support vector modelling

#### LibSVM

### 7.7.5 Neural networks

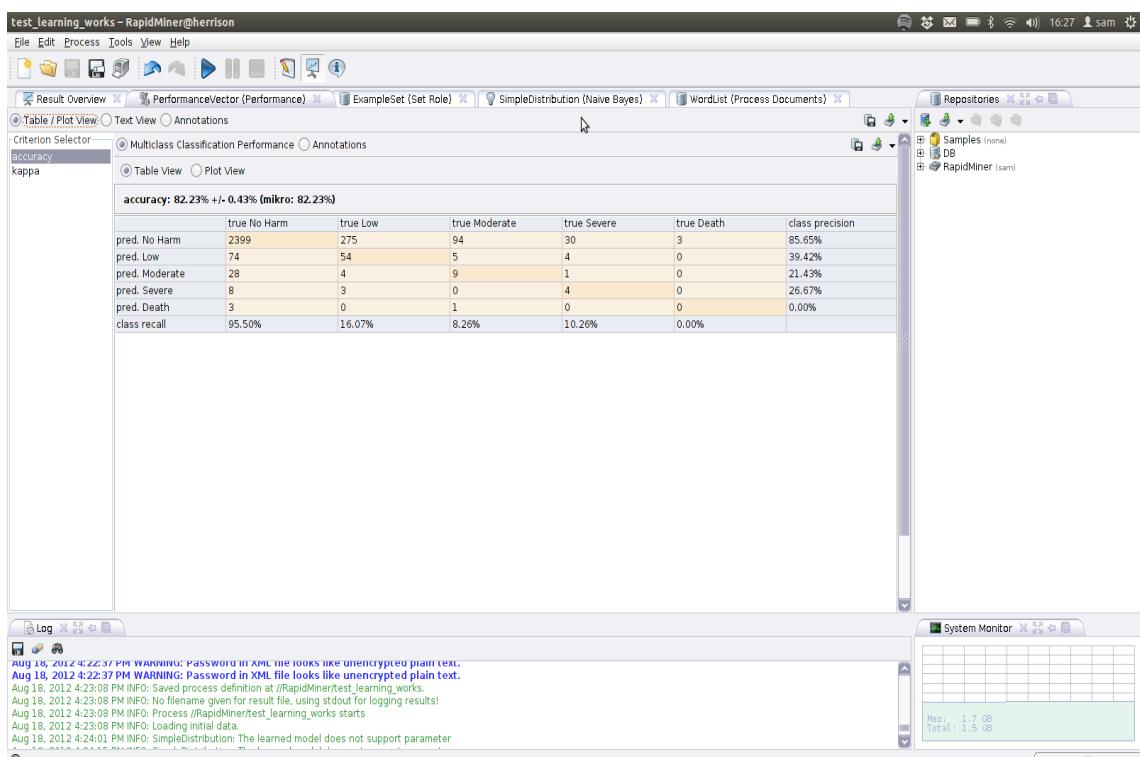


Figure 7.23: Performance of the naive bayes classifier with laplace correction for classifying degree of harm from incident description for a sample of 3000 reports.

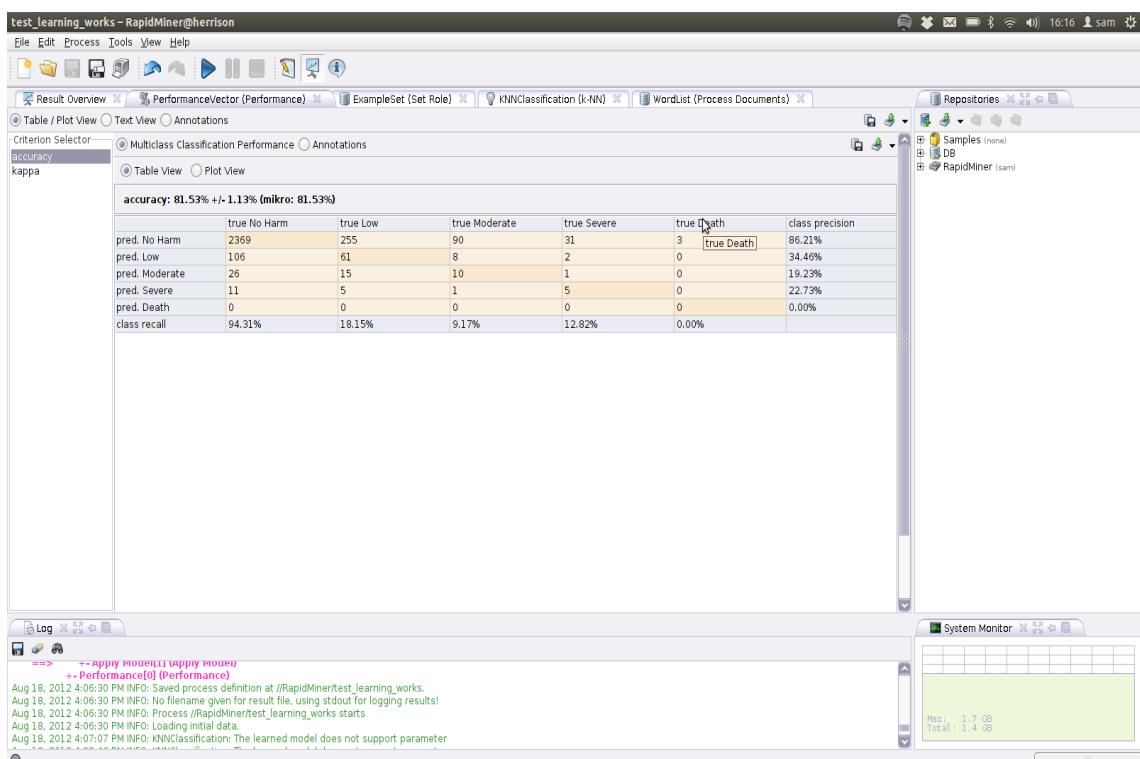


Figure 7.24: Performance of the K nearest neighbour classifier for classifying degree of harm from incident description for a sample of 3000 reports. K=3, numerical measures with cosine similarity are used.

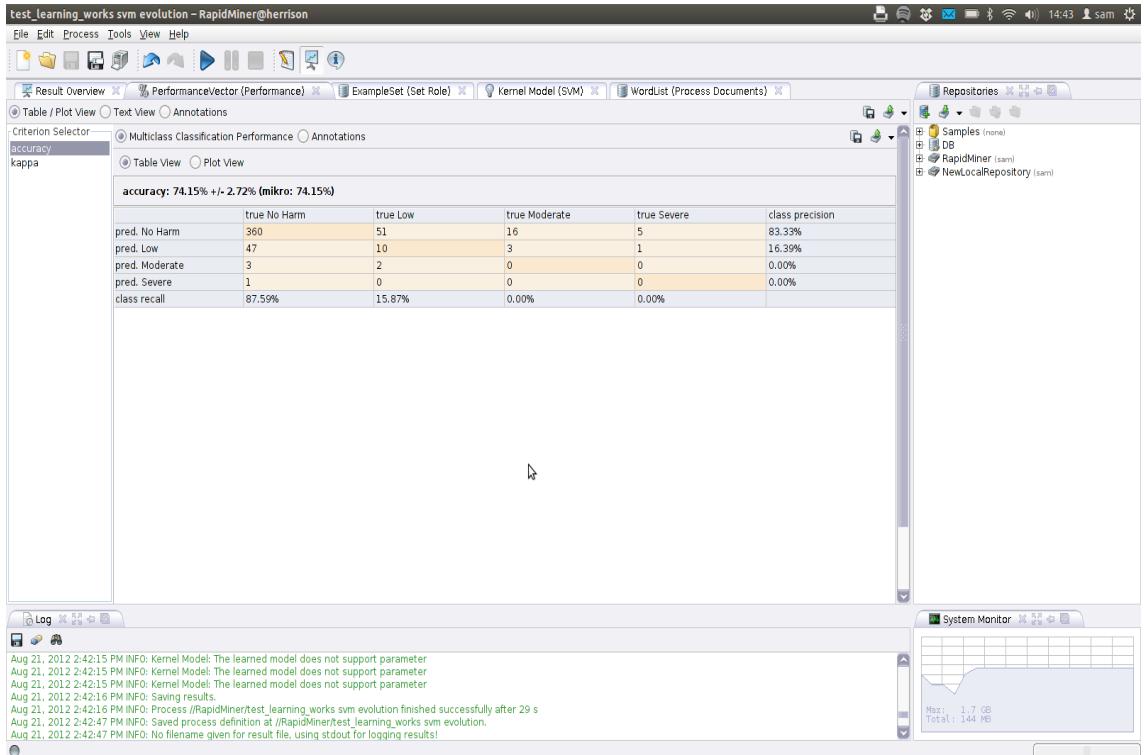


Figure 7.25: Support vector machine learner for classifying degree of harm from incident description for a sample of 500 reports. C=1000.

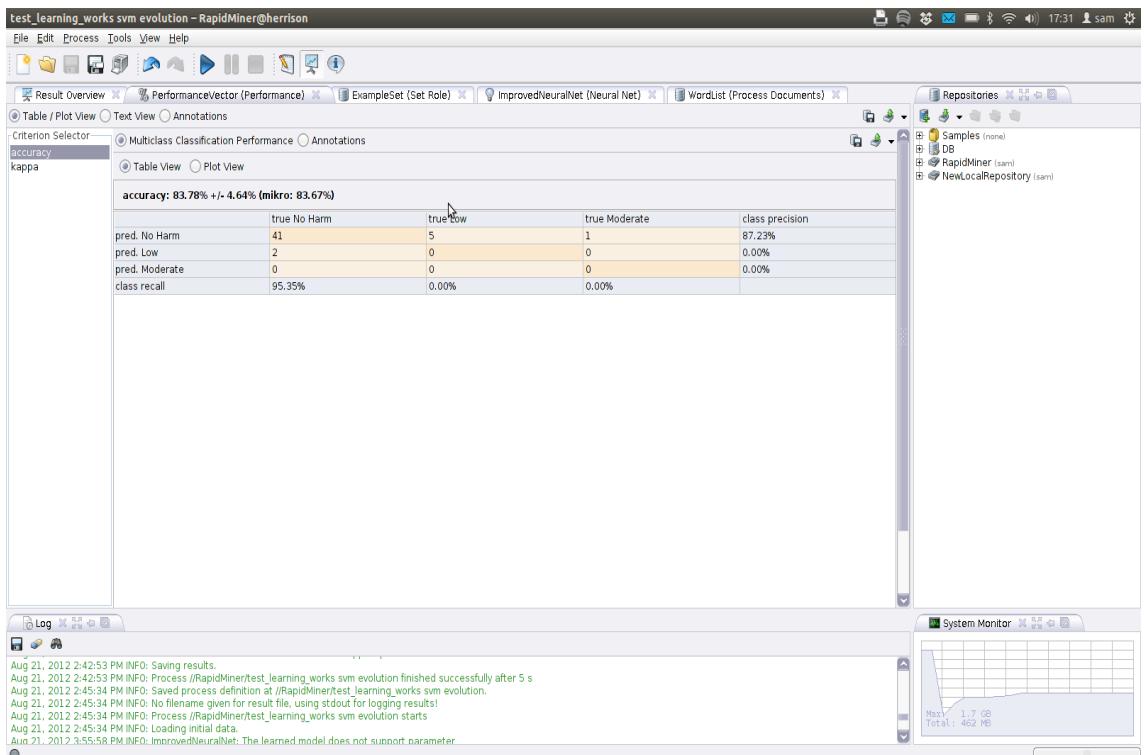


Figure 7.26: Neural network machine learner for classifying degree of harm from incident description for a sample of 50 reports (there were major performance issues).

# **Chapter 8**

## **Discussion**

### **8.1 Results**

### **8.2 Limitations**

### **8.3 Evaluation of data mining**

### **8.4 Future applications**

possible 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**

## C.2. SCREENSHOTS. RAPID MINER FILES AND EXAMPLE SCREEN SHOTS

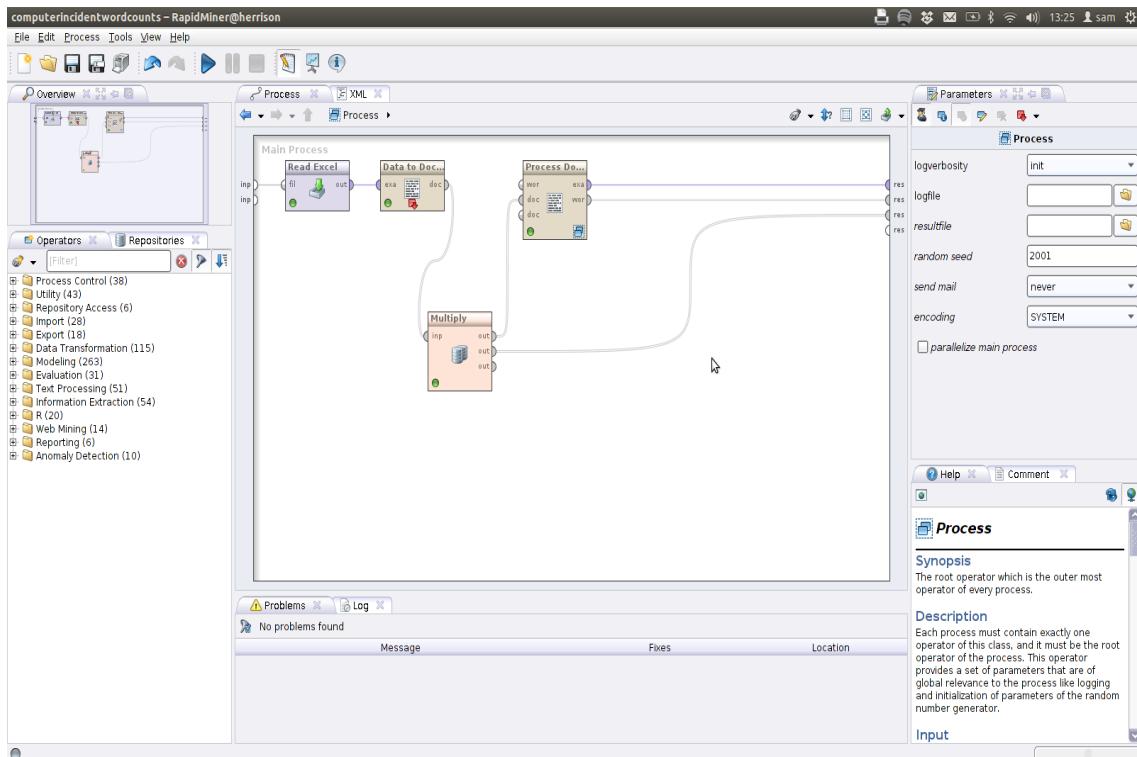


Figure C.1: Overview of Rapid Miner process to create term-frequency inverse document frequency table.

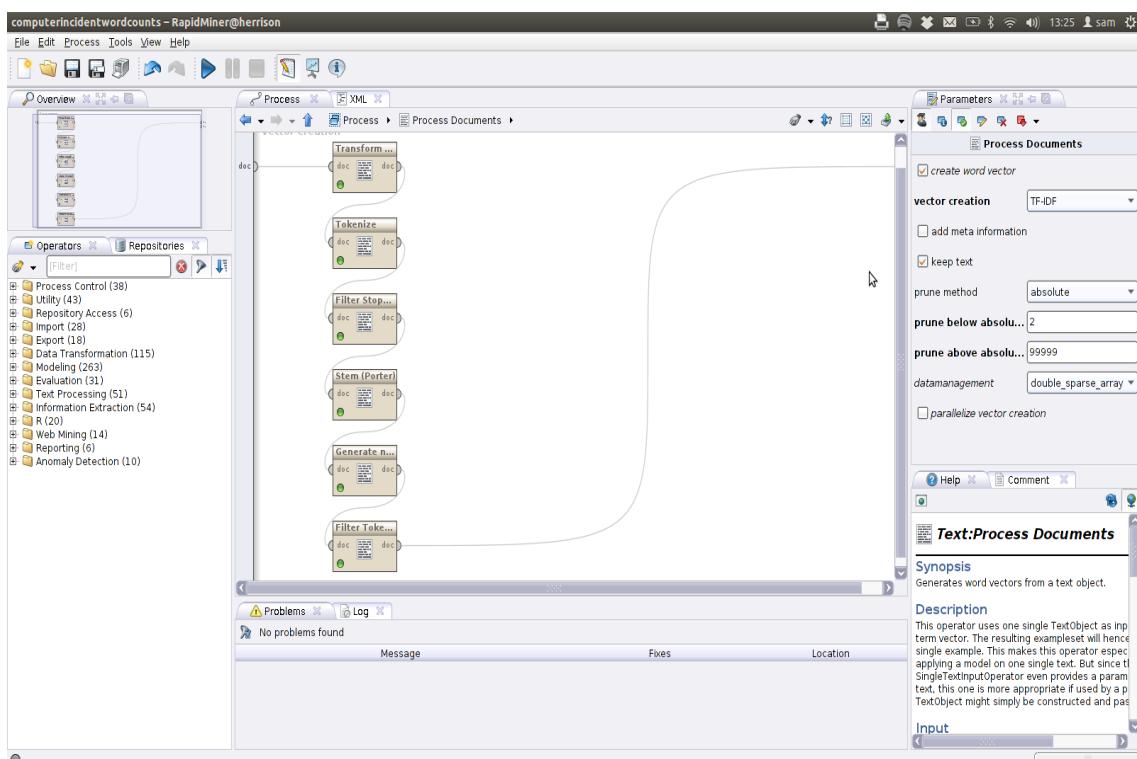


Figure C.2: Detail of Rapid Miner Document processing stages.

## C.2. APPENDIXES. RAPID MINER FILES AND EXAMPLE SCREEN SHOTS

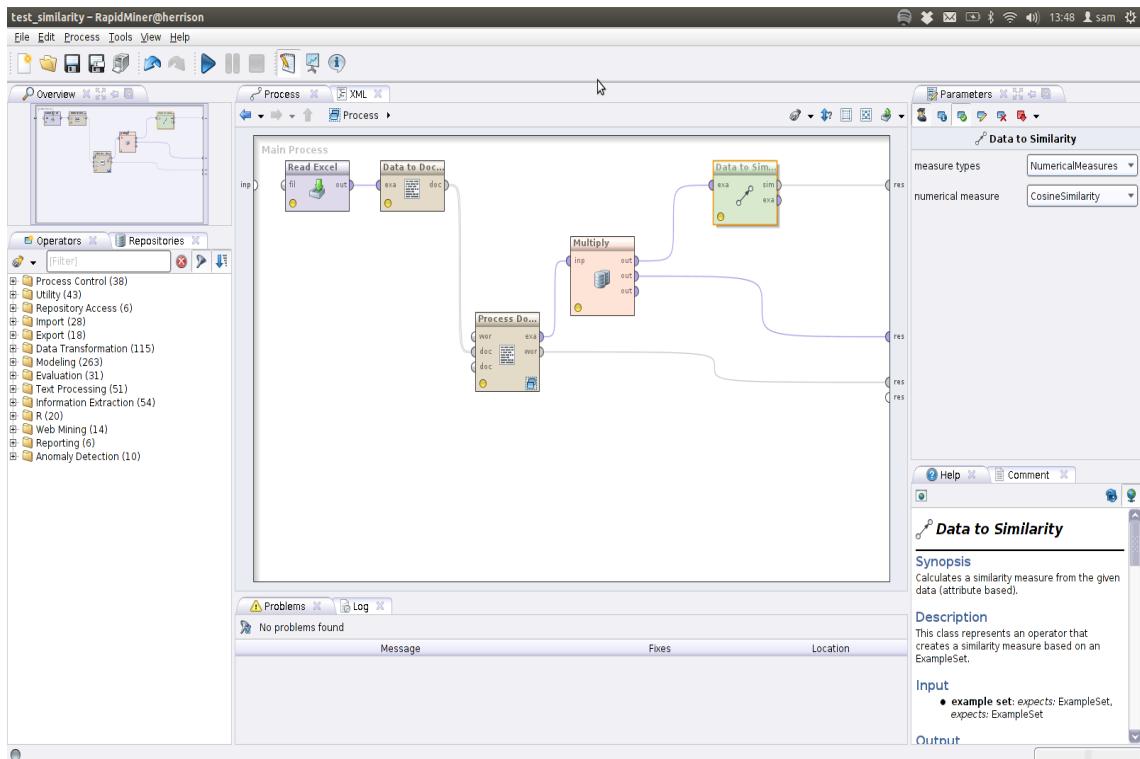


Figure C.3: Cosine similarity for sample of 50 incidents, process diagram.

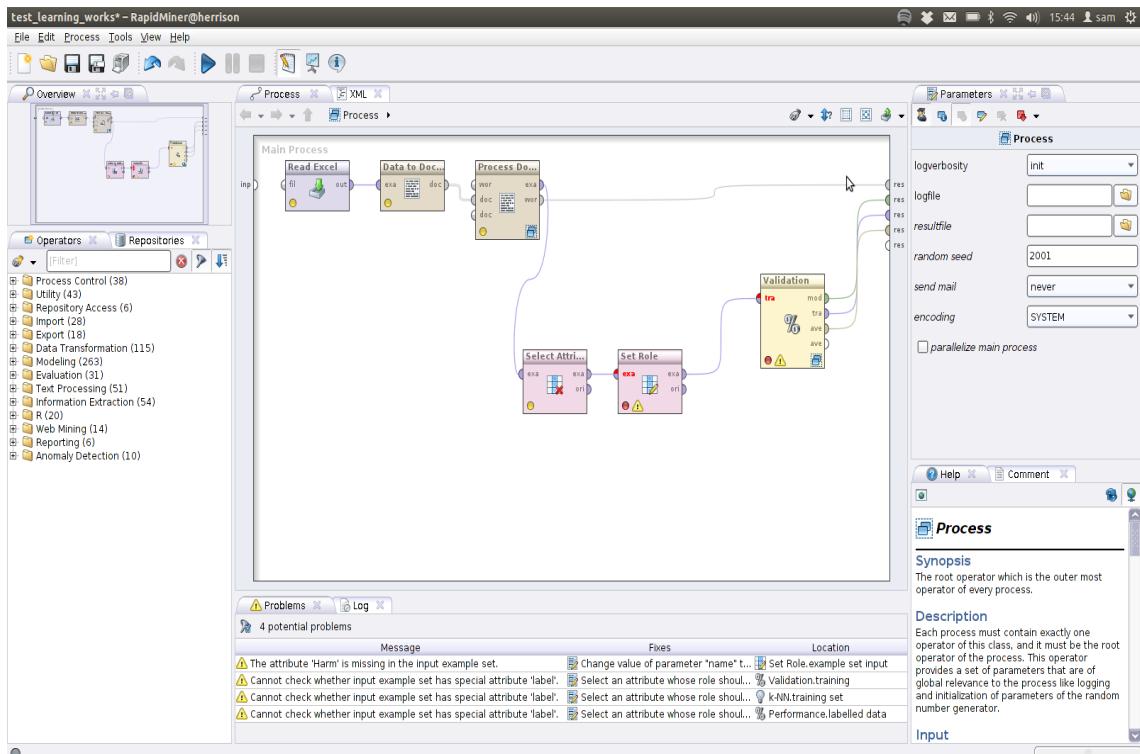


Figure C.4: Overview of Rapid Miner process for rule learning

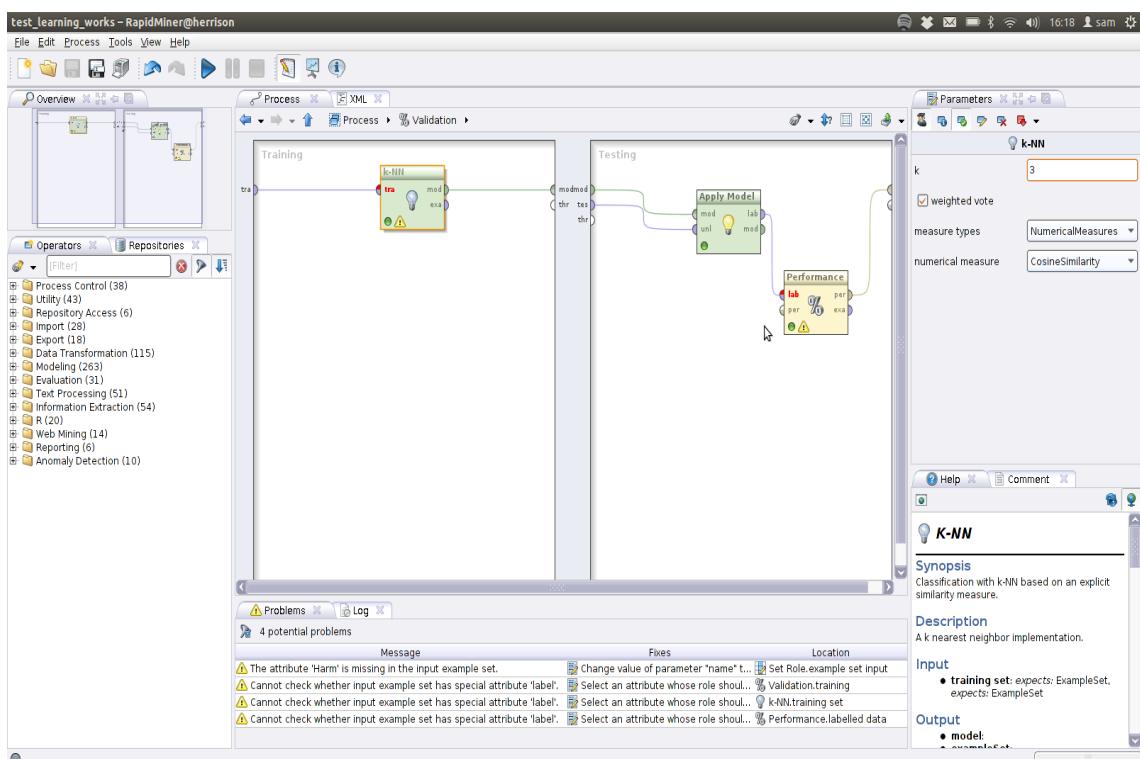


Figure C.5: Details of the Rapid Miner process for rule learning

# **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](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](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.