
An Introduction to Text Mining

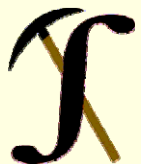
CAS 2009 RPM Seminar

Prepared by
Louise Francis
Francis Analytics and Actuarial Data Mining, Inc.
March 10, 2009
Louise_francis@msn.com
www.data-mines.com



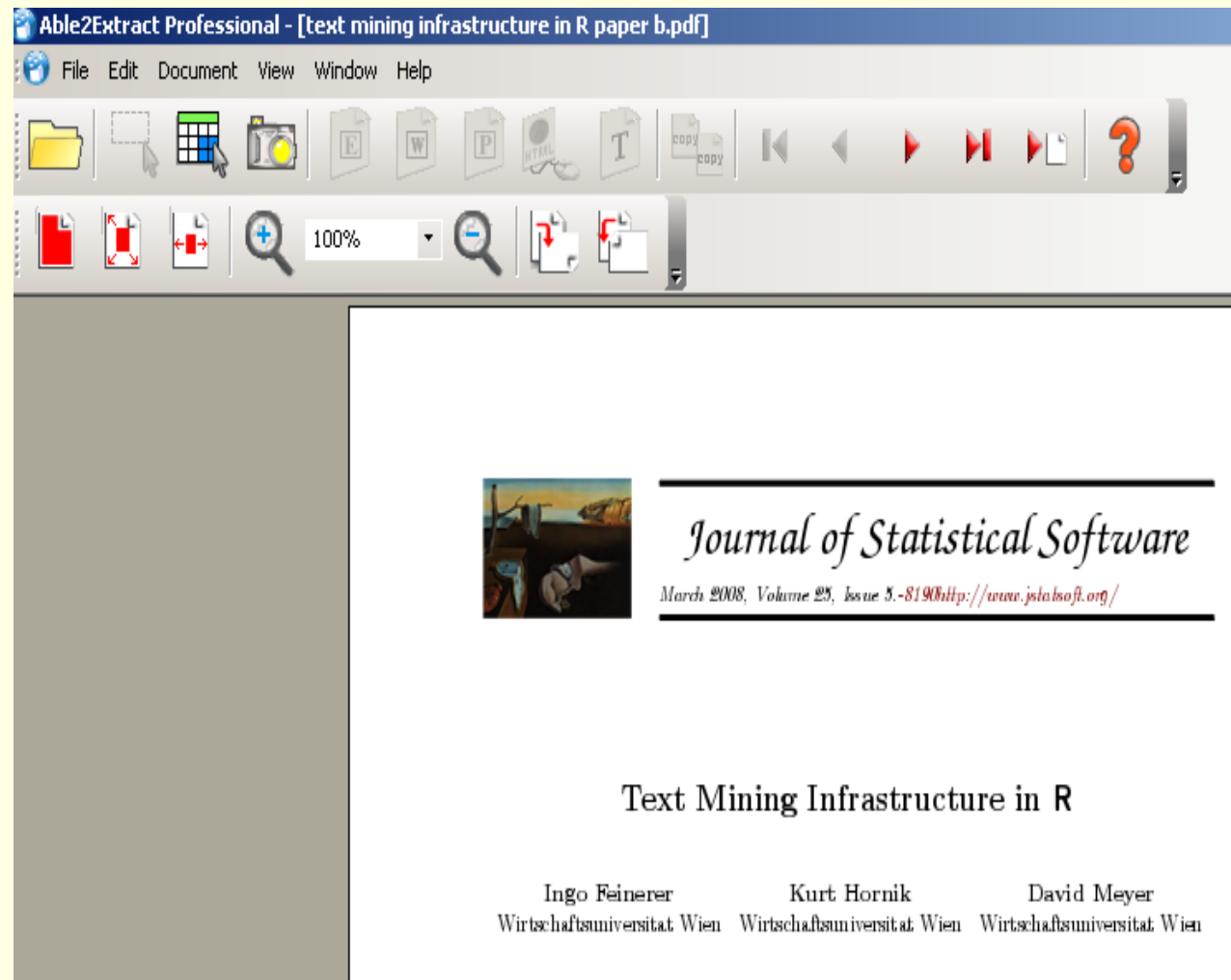
Objectives

- Present a new data mining technology
- Show how the technology uses a combination of
 - String processing functions
 - Natural language processing
 - Common multivariate procedures available in statistical most statistical software
- Discuss practical issues for implementing the methods
- Discuss software for text mining



Analyzing Unstructured Data: Uses Growing in Many Areas

- Optical Character Recognition software used to convert image to document



Major Kinds of Modeling

- Supervised learning
 - Most common situation
 - A dependent variable
 - Frequency
 - Loss ratio
 - Fraud/no fraud
 - Some methods
 - Regression
 - CART
 - Some neural networks
- Unsupervised learning
 - No dependent variable
 - Group like records together
 - A group of claims with similar characteristics might be more likely to be fraudulent
 - Applications:
 - Territory Groups
 - **Text Mining**
 - Some methods
 - Association rules
 - K-means clustering
 - Kohonen neural networks

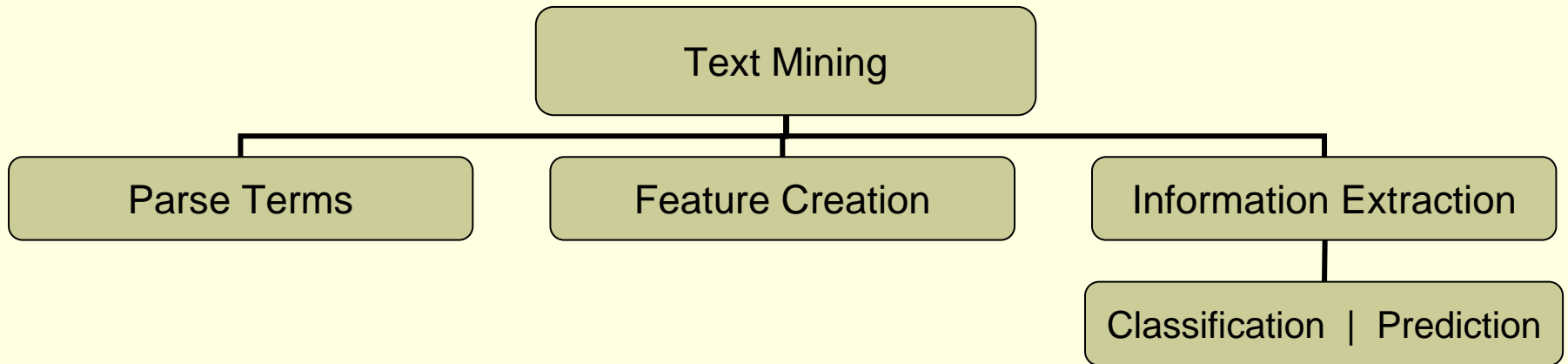


Text Mining vs. Data Mining

	<i>Analysis Types</i>	<i>Non-novel information</i>	<i>Novel information</i>	<i>Comment</i>
Non-text data	standard predictive modeling	database queries	new patterns and relationships	small fraction of data
Text data	computational linguistics/statistical mining of text data	information retrieval	text mining	

modified from Manning/Hearst

Text Mining Process





String Processing

Example: Claim Description Field

INJURY DESCRIPTION
BROKEN ANKLE AND SPRAINED WRIST
FOOT CONTUSION
UNKNOWN
MOUTH AND KNEE
HEAD, ARM LACERATIONS
FOOT PUNCTURE
LOWER BACK AND LEGS
BACK STRAIN
KNEE



Parse Text Into Terms

- Separate free form text into words
- “BROKENANKLE AND SPRAINED WRIST” →
 - BROKEN
 - ANKLE
 - AND
 - SPRAINED
 - WRIST

Parsing Text

- Separate words from spaces and punctuation
- Clean up
- Remove redundant words
- Remove words with no content
- Cleaned up list of Words referred to as tokens

Parsing a Claim Description Field With Microsoft Excel String Functions

Full Description	Total Length	Location of Next Blank	First Word	Remainder Length 1
(1)	(2)	(3)	(4)	(5)
BROKEN ANKLE AND SPRAINED WRIST	31	7	BROKEN	24
Remainder 1		2nd Blank	2nd Word	Remainder Length 2
(6)		(7)	(8)	(9)
ANKLE AND SPRAINED WRIST		6	ANKLE	18
Remainder 2		3rd Blank	3rd Word	Remainder Length 3
(10)		(11)	(12)	(13)
AND SPRAINED WRIST		4	AND	14
Remainder 3		4th Blank	4th Word	Remainder Length 4
(14)		(15)	(16)	(17)
SPRAINED WRIST		9	SPRAINED	5
Remainder 4		5th Blank	5th Word	
(18)		(19)	(20)	
WRIST		0	WRIST	

String Functions

- Use substring function in R/S-PLUS to find spaces

```
# Initialize
charcount<-nchar(Description)
# number of records of text
Linecount<-length(Description)
Num<-Linecount*6
# Array to hold location of spaces
Position<-rep(0,Num)
dim(Position)<-c(Linecount,6)
# Array for Terms
Terms<-rep("",Num)
dim(Terms)<-c(Linecount,6)
wordcount<-rep(0,Linecount)
```

Search for Spaces

```
for (i in 1:Linecount)
{
n<-charcount[i]
k<-1
for (j in 1:n)
{
    Char<-substring(Description[i],j,j)
    if (is.all.white(Char)) { Position[i,k]<-j; k<-k+1 }
    wordcount[i]<-k
}
}
```

Get Words

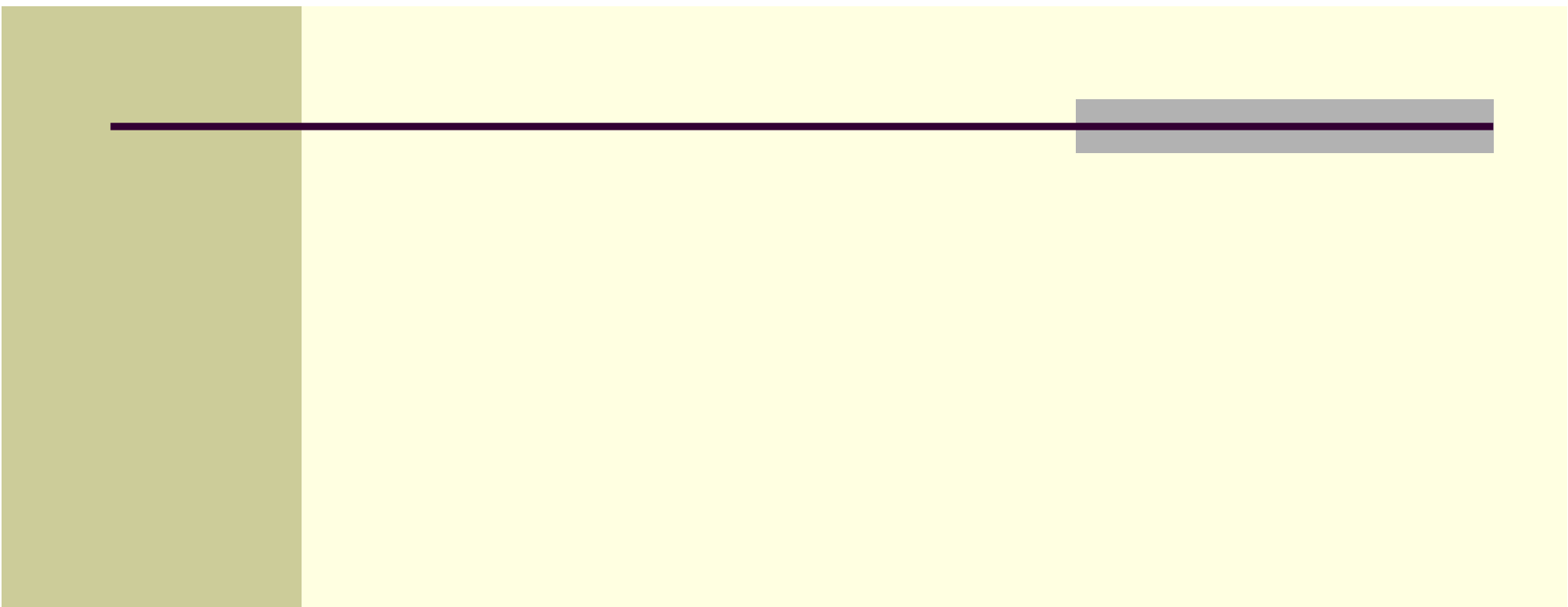
```
# parse out terms
for (i in 1:Linecount)
{
  # first word
  if (Position[i,1]==0) Terms[i,1]<-Description[i] else if (Position[i,1]>0)
  Terms[i,1]<-substring(Description[i],1,Position[i,1]-1)
  for (j in 1:wordcount)
  {
    if (Position[i,j]>0)
    {
      Terms[i,j]<-substring(Description[i],Position[i,j-1]+1,Position[i,j]-1)
    }
  }
}
```

Extraction Creates Binary Indicator Variables

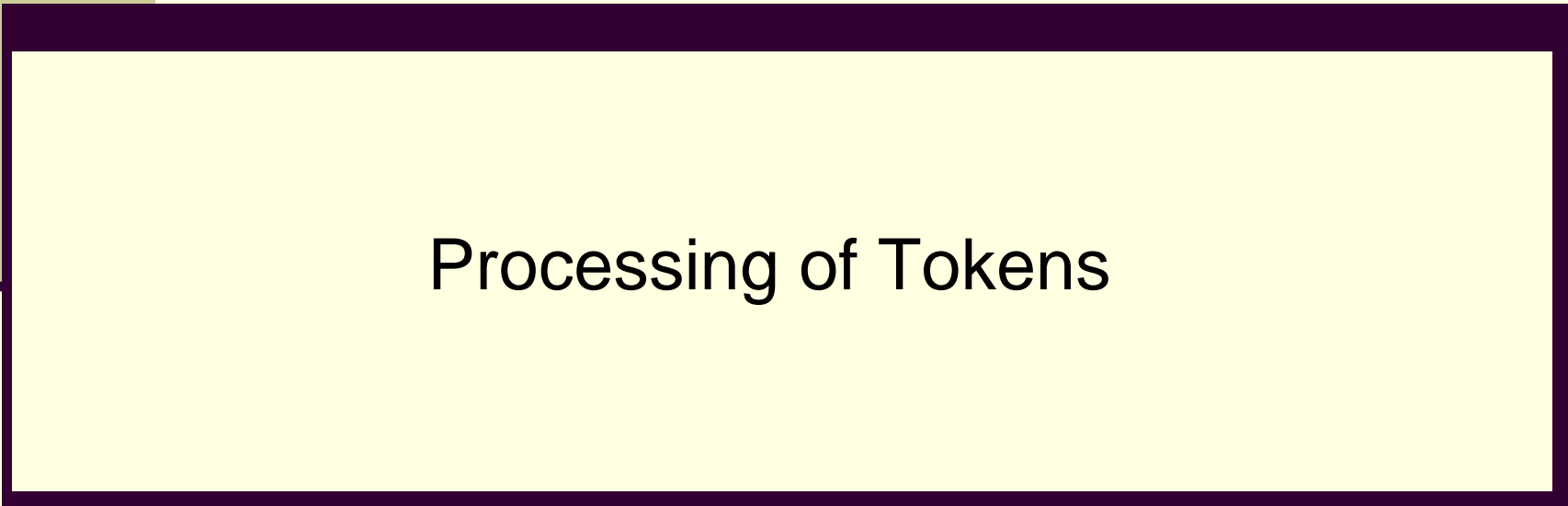
INJURY DESCRIPTION	BROKEN	ANKLE	AND	SPRAINED	W R I S T	F O O T	CONTU - SION	UNKNOWN	N E C K	BACK	STRAIN
BROKEN ANKLE AND SPRAINED WRIST	1	1	1	1	1	0	0	0	0	0	0
FOOT CONTUSION	0	0	0	0	0	1	1	0	0	0	0
UNKNOWN	0	0	0	0	0	0	0	1	0	0	0
NECK AND BACK STRAIN	0	0	1	0	0	0	0	0	1	1	1

Term Document Matrix/Index

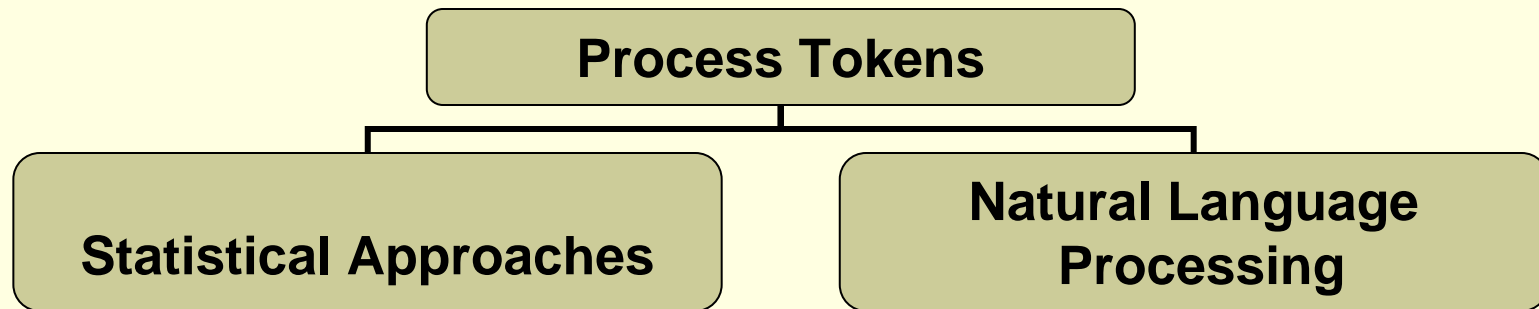
- Uses frequency measure for each word instead of on-off binary indicator
- “The Index representation does not do justice to the complexity of human language but is dictated by the practical difficulty of storing more information objects”
 - Liang et al.



Processing of Tokens



Further Processing



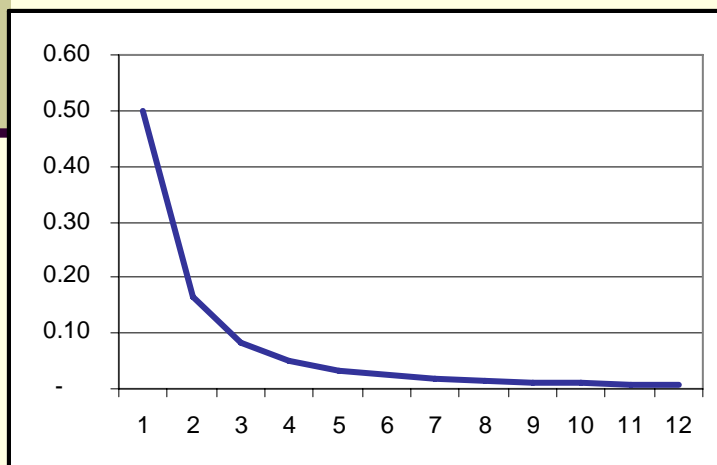
Natural Language Processing

- Draws on many disciplines
 - Artificial Intelligence
 - Linguistics
 - Statistics
 - Speech Recognition
- Includes lexical analysis, multiword phrase groupings, sense disambiguation, part of speech tagging
- Arguments against: it is error-prone and output contains too much detail and noise

Zipff's Law

- Distribution for how often each word occurs in a language
- Inverse relation between rank (r) of word and its frequency (f)

$$f \propto \frac{1}{r}$$



Mandelbrot's Refinement

$$f = p(r + \rho)^{-B}$$

Consequences of Zipf

- There are a few very frequent tokens or words that add little to information
 - Known as stop words
 - Examples: a, the, to, from
- Usually
 - Small number of very common words (i.e., stop words)
 - Medium number of medium frequency words
 - Large number of infrequent words
 - The medium frequency words the most useful

Word Frequency in Tom Sawyer

Word	Frequency (f)	Rank (r)	Word	Frequency (f)	Rank (r)
the	3,332	1	group	13	600
and	2,972	2	lead	11	700
a	1,775	3	friends	10	800
he	877	4	begin	9	900
but	410	5	family	8	1,000
be	294	6	brushed	4	2,000
there	222	7	sins	2	3,000
one	172	8	could	2	4,000
about	158	9	applausive	1	8,000

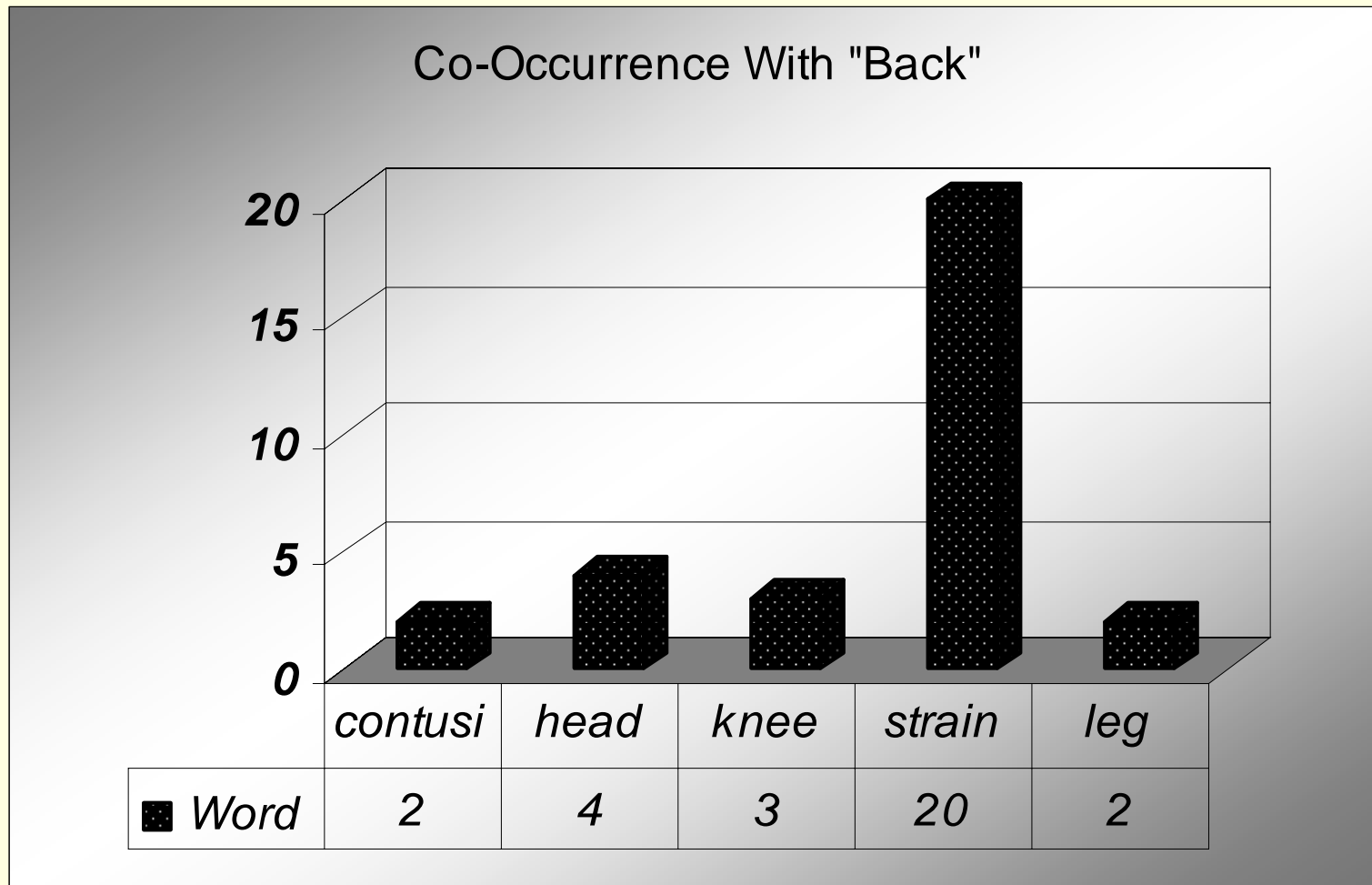
Collocation

- Multiword units, word that go together, phrases with recognized meaning
- Examples from Recent newspaper
 - Philadelphia Inquirer
 - FDIC (Federal Deposit Insurance Corporation)
 - Wall Street
 - Las Vegas

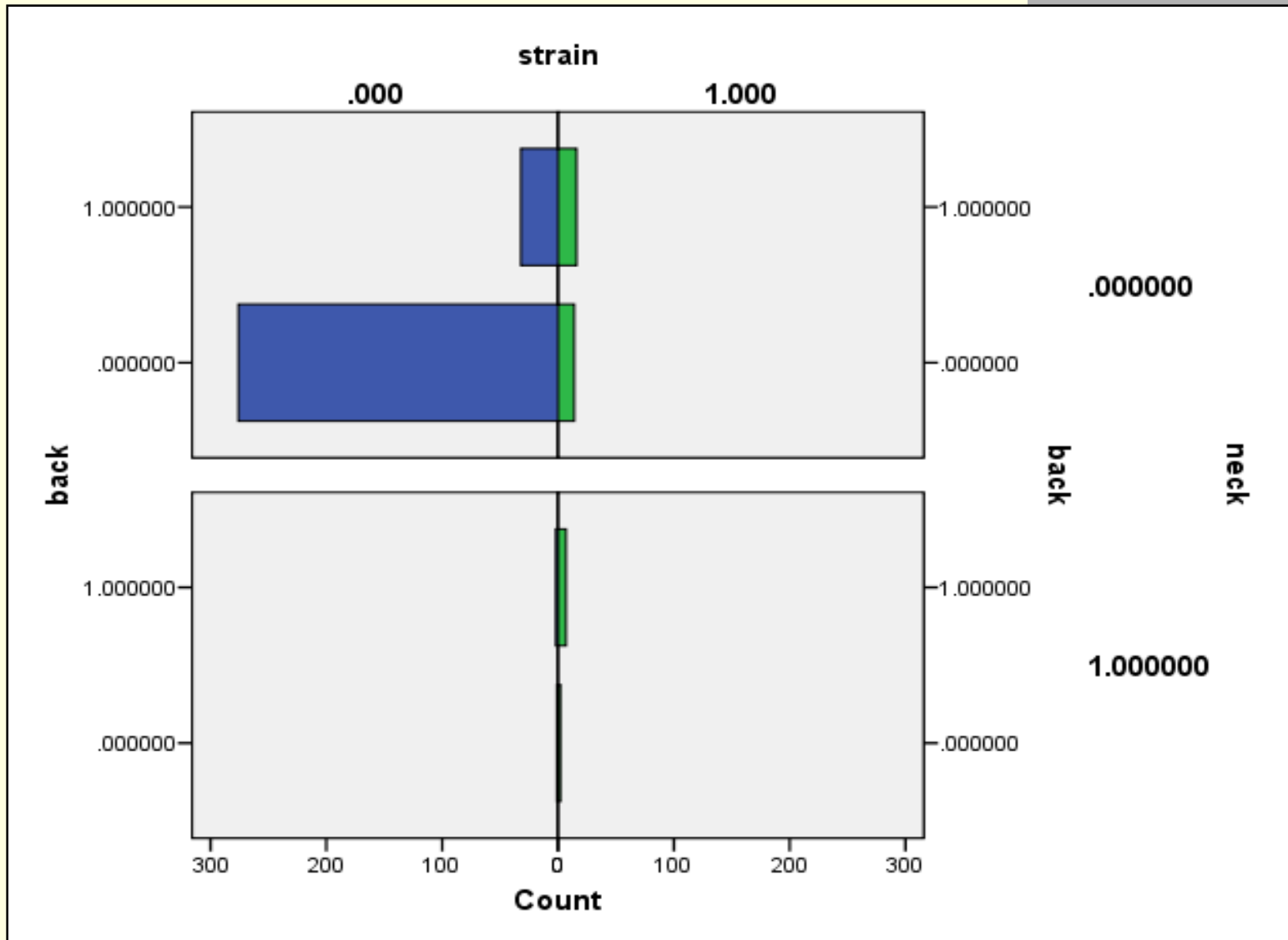
Concordances

- Finding contexts in which verbs appear
- Use key word in context
- Lists all occurrences of the word and the words that occur with it.

The Word “Back” in claim description



Some Co-Occurrences



Identifying Collocations

- Two most frequent patterns
 - Noun- noun
 - Adjective noun
- Analyst will probably want these phrases in a dictionary

Semantics

- Meaning of words, phrases, sentences and other language structures
 - Lexical semantics
 - Meaning of individual words
 - Examples; synonyms, antonyms
 - Meanings of combinations of words

Wordnet

- Semantic lexicon for English language
- Some Features
 - Synonyms
 - Antonyms
 - Hypernyms
 - Hyponyms
- Developed by Princeton University Cognitive Sciences Laboratory

Wordnet Entry for Reserve

Searches for reserve:

Noun

Verb

Senses:

The noun reserve has 7 senses (first 3 from tagged texts)

1. (2) modesty, **reserve** -- (formality and propriety of manner)
2. (1) **reserve**, backlog, stockpile -- (something kept back or saved for future use or a special purpose)
3. (1) substitute, **reserve**, second-stringer -- (an athlete who plays only when a starter on the team is replaced)
4. **reserve** -- ((medicine) potential capacity to respond in order to maintain vital functions)
5. reservation, **reserve** -- (a district that is reserved for particular purpose)
6. military reserve, **reserve** -- (armed forces that are not on active duty but can be called in an emergency)
7. **reserve**, reticence, taciturnity -- (the trait of being uncommunicative; not volunteering anything more than necessary)

The verb reserve has 4 senses (first 3 from tagged texts)

1. (7) **reserve** -- (hold back or set aside, especially for future use or contingency; "they held back their applause in anticipation")
2. (6) allow, appropriate, earmark, set aside, **reserve** -- (give or assign a resource to a particular person or cause; "I will earmark this money for your research"; "She sets aside time for meditation every day")
3. (1) **reserve** -- (obtain or arrange (for oneself) in advance; "We managed to reserve a table at Maxim's")
4. **reserve**, hold, book -- (arrange for and reserve (something for someone else) in advance; "reserve me a seat on a flight"; "The agent booked tickets to the show for the whole family")

Verb Reserve

The verb reserve has 4 senses (first 3 from tagged texts)

1. (7) **reserve** -- (hold back or set aside, especially for future use or contingency; "they held back their applause in anticipation")
2. (6) allow, appropriate, earmark, set aside, **reserve** -- (give or assign a resource to a particular person or cause; "I will earmark this money for your research"; "She sets aside time for meditation every day")
3. (1) **reserve** -- (obtain or arrange (for oneself) in advance; "We managed to reserve a table at Maxim's")
4. **reserve**, hold, book -- (arrange for and reserve (something for someone else) in advance; "reserve me a seat on a flight"; "The agent booked tickets to the show for the whole family"; "please hold a table at Maxim's")

Wordnet Visualizations for Underwriter



Eliminate Stopwords

- Common words with no meaningful content

Stopwords
A
And
Able
About
Above
Across
Aforementioned
After
Again

Stemming: Identify Synonyms and Words with Common Stem

Parsed Words	
HEAD	INJURY
LACERATION	NONE
KNEE	BRUISED
UNKNOWN	TWISTED
L	LOWER
LEG	BROKEN
ARM	FRACTURE
R	FINGER
FOOT	INJURIES
HAND	LIP
ANKLE	RIGHT
HIP	KNEES
SHOULDER	FACE
LEFT	FX
CUT	SIDE
WRIST	PAIN
NECK	INJURED

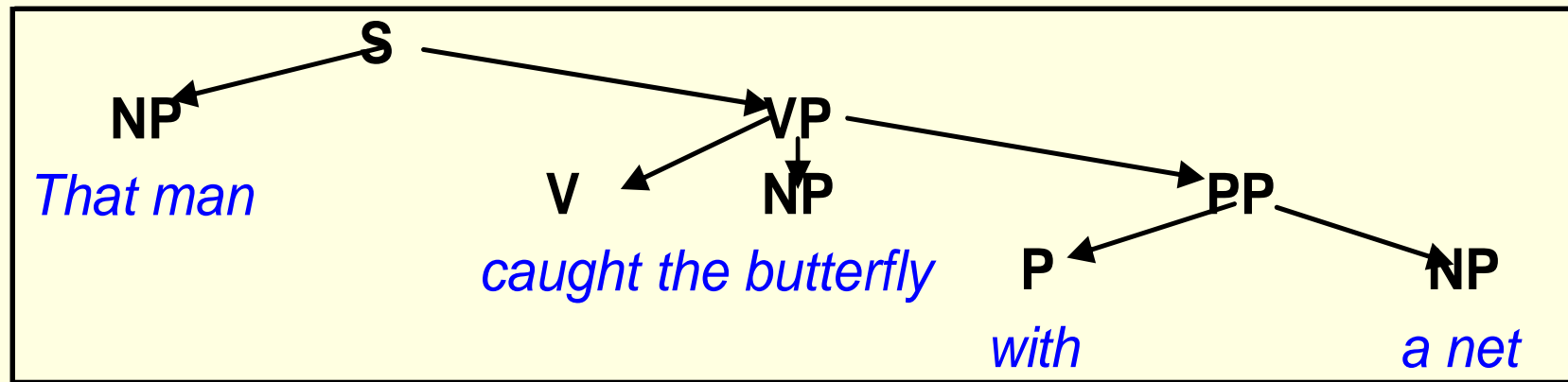
Part of Speech Morphology

- Parts of Speech (POS)
 - Noun
 - Verb
 - Adjective
 - These are open or lexical categories that have large numbers of members and new members frequently added
 - Also prepositions and determiners
 - Of, on, the, a
 - Generally closed categories

Diagrams of Parts of Speech

- Sentence
- Noun Phrase
- Verb Phrase

Diagramming Parts of Speech



Word Sense Disambiguation

- Many words have multiple possible meanings or senses --→ ambiguity about interpretation
- Word can be used as different part of speech
- Disambiguation determines which sense is being used

Disambiguation

- Statistical methods
- NLP based methods

Disambiguation: An Algorithm

- Build list of associated words and weights for ambiguous word
- Read “context” of ambiguous word, save nouns and adjectives in list
- Get list of senses of ambiguous word from dictionary and do for each:
 - Assign initial score to current sense
 - Scan list of context words
 - For each check if it is associated word, then increment or decrement score
- Sort scores in descending order and list top scoring senses

From Konchady, Text Mining Application Programming



Statistical Approaches

Objective

- Create a new variable from free form text
- Use words in injury description to create an injury code
- New injury code can be used in a predictive model or in other analysis

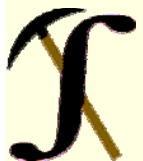


Dimension Reduction

	CLAIM NUMBER	DATE OF LOSS	STATUS	INCURRED LOSS
				VARIABLES
RECORDS	1998001	09/15/97	C	407.61
	1998002	09/25/97	C	0.00
	1998003	09/26/97	C	0.00
	1998004	09/29/97	C	8,247.16
	1998005	09/29/97	C	0.00
	1998006	10/02/97	C	0.00
	1998007	10/10/97	C	0.00
	1998008	10/24/97	G	0.00
	1998009	10/29/97	C	21,211.66
	1998010	10/29/97	C	0.00
	1998011	11/03/97	G	0.00
	1998012	11/03/97	C	0.00
	1998013	11/04/97	C	451.66
	1998014	11/04/97	C	0.00
	1998015	11/04/97	C	0.00
	1998016	11/06/97	C	15,903.66
	1998017	11/11/97	C	465.10

The Two Major Categories of Dimension Reduction

- Variable reduction
 - Factor Analysis
 - Principal Components Analysis
- Record reduction
 - Clustering
- Other methods tend to be developments on these



Clustering

- Common Method: k-means and hierarchical clustering
- No dependent variable – records are grouped into classes with similar values on the variable
- Start with a measure of similarity or dissimilarity
- Maximize dissimilarity between members of different clusters



Dissimilarity (Distance) Measure – Continuous Variables

■ Euclidian Distance

$$d_{ij} = \left(\sum_{k=1}^m (x_{ik} - x_{jk})^2 \right)^{1/2} \quad i, j = \text{records} \quad k = \text{variable}$$

■ Manhattan Distance

$$d_{ij} = \left(\sum_{k=1}^m |x_{ik} - x_{jk}| \right)$$



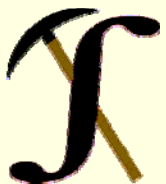
K-Means Clustering

- Determine ahead of time how many clusters or groups you want
- Use dissimilarity measure to assign all records to one of the clusters

Cluster Number	back	contusion	head	knee	strain	unknown	laceration
1	0.00	0.15	0.12	0.13	0.05	0.13	0.17
2	1.00	0.04	0.11	0.05	0.40	0.00	0.00

Hierarchical Clustering

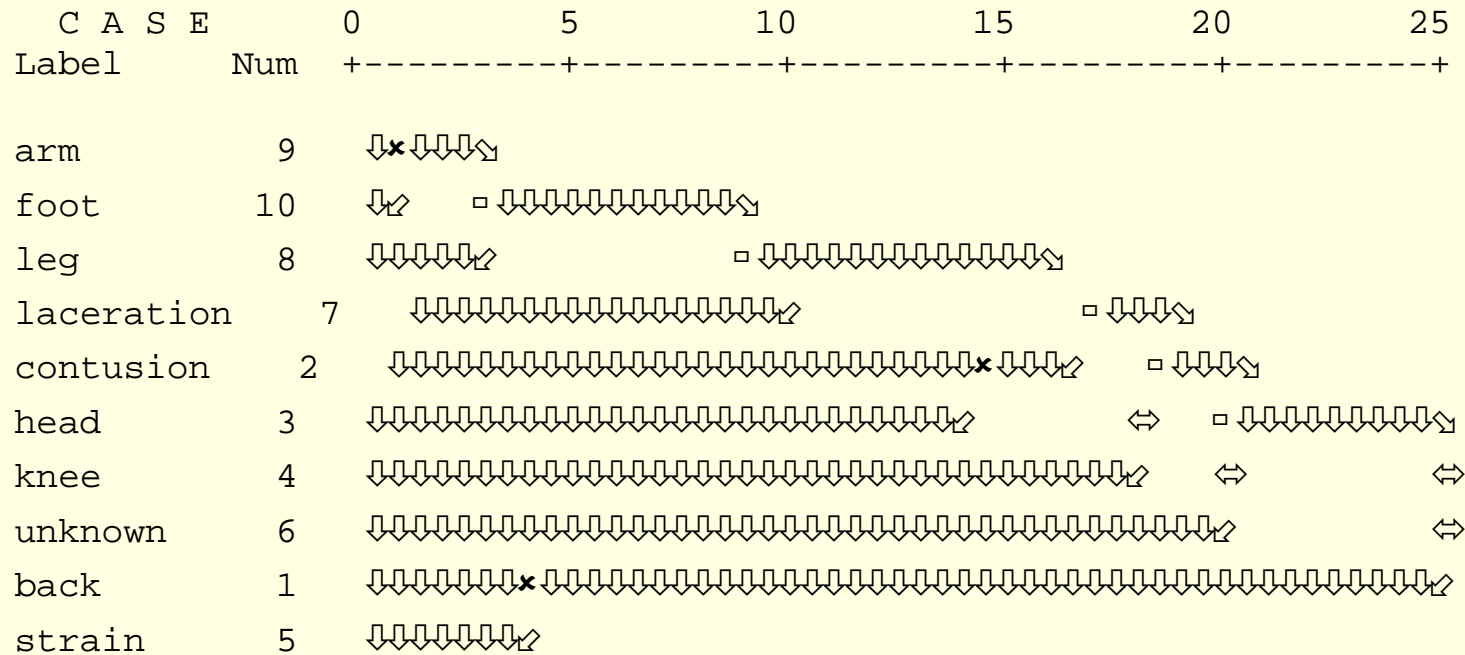
- A stepwise procedure
- At beginning, each records is its own cluster
- Combine the most similar records into a single cluster
- Repeat process until there is only one cluster with every record in it



Hierarchical Clustering Example

Dendrogram for 10 Terms

Rescaled Distance Cluster Combine



Final Cluster Selection

Cluster	Back	Contusion	head	knee	strain	unknown	laceration	Leg
1	0.000	0.000	0.000	0.095	0.000	0.277	0.000	0.000
2	0.022	1.000	0.261	0.239	0.000	0.000	0.022	0.087
3	0.000	0.000	0.162	0.054	0.000	0.000	1.000	0.135
4	1.000	0.000	0.000	0.043	1.000	0.000	0.000	0.000
5	0.000	0.000	0.065	0.258	0.065	0.000	0.000	0.032
6	0.681	0.021	0.447	0.043	0.000	0.000	0.000	0.000
7	0.034	0.000	0.034	0.103	0.483	0.000	0.000	0.655
Weighted Average	0.163	0.134	0.120	0.114	0.114	0.108	0.109	0.083

Use New Injury Code in a Logistic Regression to Predict Serious Claims

$$Y = B_0 + B_1 \text{Attorney} + B_2 \text{Injury_Group}$$

$$Y = \text{Claim Severity} > \$10,000$$

Mean Probability of Serious Claim vs. Actual Value

	Actual Value	
	1	0
Avg Prob	0.31	0.01

Software for Text Mining- Commercial Software

- Most major software companies, as well as some specialists sell text mining software
 - These products tend to be for large complicated applications, such as classifying academic papers
 - They also tend to be expensive
- One inexpensive product reviewed by *American Statistician* had disappointing performance



Perl for Text Processing

- Free open source programming language
- www.perl.org
- Used a lot for text processing
- *Perl for Dummies* gives a good introduction
- *Practical Text Mining With Perl*, Roger Bilisoly

Perl Functions for Parsing

- `$TheFile = "GLClaims.txt";`
- `$Linelength=length($TheFile);`
- `open(INFILE, $TheFile) or die "File not found";`
- `# Initialize variables`
- `$Linecount=0;`
- `@alllines=();`
- `while(<INFILE>){`
- `$Theline=$_;`
- `chomp($Theline);`
- `$Linecount = $Linecount+1;`
- `$Linelength=length($Theline);`
- `#Use space for splitting`
- `@Newitems = split(/ /,$Theline);`
- `print "@Newitems \n";`
- `push(@alllines, [@Newitems]);`
- `} # end while`

What if more than one space

- Regular Expressions

```
$Test = "This is a list  of words";  
@words =split (/ \s+/, $Test);  
print "@words\n";
```

Commercial Software for Text Mining

ActivePoint , offering natural language processing	Leximancer , makes automatic text analysis
AeroText , a high performance text processing engine	Lextek Onix Toolkit , for adding text mining capabilities
Arrowsmith software for support of text mining	Lextek Profiling Engine , for automatic text analysis
Attensity , offers a complete solution for text mining	Linguamatics , offering Natural Language Processing
Text Data Mining and Analysis (TDM)	Megaputer Text Analyst , offers text mining capabilities
Basis Technology , provides text mining solutions	Monarch , data access and analysis
ClearForest , tools for analysis and text mining	NewsFeed Researcher , presents text mining capabilities
Compare Suite , compares text documents	Nstein , Enterprise Search and text mining
Connexor MachineSense , discovers patterns in text	Power Text Solutions , extensive text mining capabilities
Copernic Summarizer , can reduce text to its essence	Readability Studio , offers tools for text analysis
Corpora , a Natural Language Processing tool	Recommind MindServer , uses text mining for recommendations
NEW! Crossminder , natural language processing	SAS Text Miner , provides a rich set of text mining capabilities
Cypher , generates the RDF graph from text	SPSS LexiQuest , for accessing text mining capabilities
DolphinSearch , text-reading and search engine	SPSS Text Mining for Clementine , text mining capabilities
dtSearch , for indexing, searching and text mining	SWAPit , Fraunhofer-FIT's text mining capabilities
NEW! Eagle text mining software, for text mining	TEMIS Luxid® , an Information Management tool
Enkata , providing a range of text mining capabilities	TeSSI® , software components for text mining
Entrieva , patented technology for text mining	Text Analysis Info , offering software for text mining
Expert System , using proprietary text mining capabilities	Textalyser , online text analysis tool
Files Search Assistant , quick search and text mining	TextOre , providing B2B analytical capabilities
IBM Intelligent Miner Data Mining , text mining capabilities	TextPipe Pro , text conversion and analysis
Intellexer , natural language processing and text mining	TextQuest , text analysis software
Insightful InFact , an enterprise text mining solution	Readware Information Process , text mining capabilities
Inxight , enterprise software for text mining	Quenza , automatically extracts text mining capabilities
ISYS:desktop , searches over text documents	VantagePoint provides a variety of text mining capabilities
Kwalitan 5 for Windows , uses text mining capabilities	VisualText™ , by TextAI is a commercial text mining tool
	Wordstat , analysis module for text mining

Free Software for Text Mining

NEW!

[GATE](#), a leading open-so

[INTEXT](#), MS-DOS versio

NEW!

[LingPipe](#) is a suite of Jav

NEW!

[Open Calais](#), an open-so

[S-EM \(Spy-EM\)](#), a text cl

[The Semantic Indexing F](#)

[Vivisimo/Clusty](#) web sea

References

- Hoffman, P, *Perl for Dummies*, Wiley, 2003
- Francis, L., “Taming Text”, 2006 CAS Winter Forum
- Weiss, Shalom, Indurkha, Nitin, Zhang, Tong and Damerau, Fred, *Text Mining*, Springer, 2005
- Konchady, Manu, *Text Mining Application Programming*, Thompson, 2006
- Liang et. al., “Extracting Statistical Data Frames From Text”, Insightful Corporation
- Manning and Schultze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999



Questions?
