Text Mining on Unstructured Data
CAS 2008 Predictive Modeling Seminar

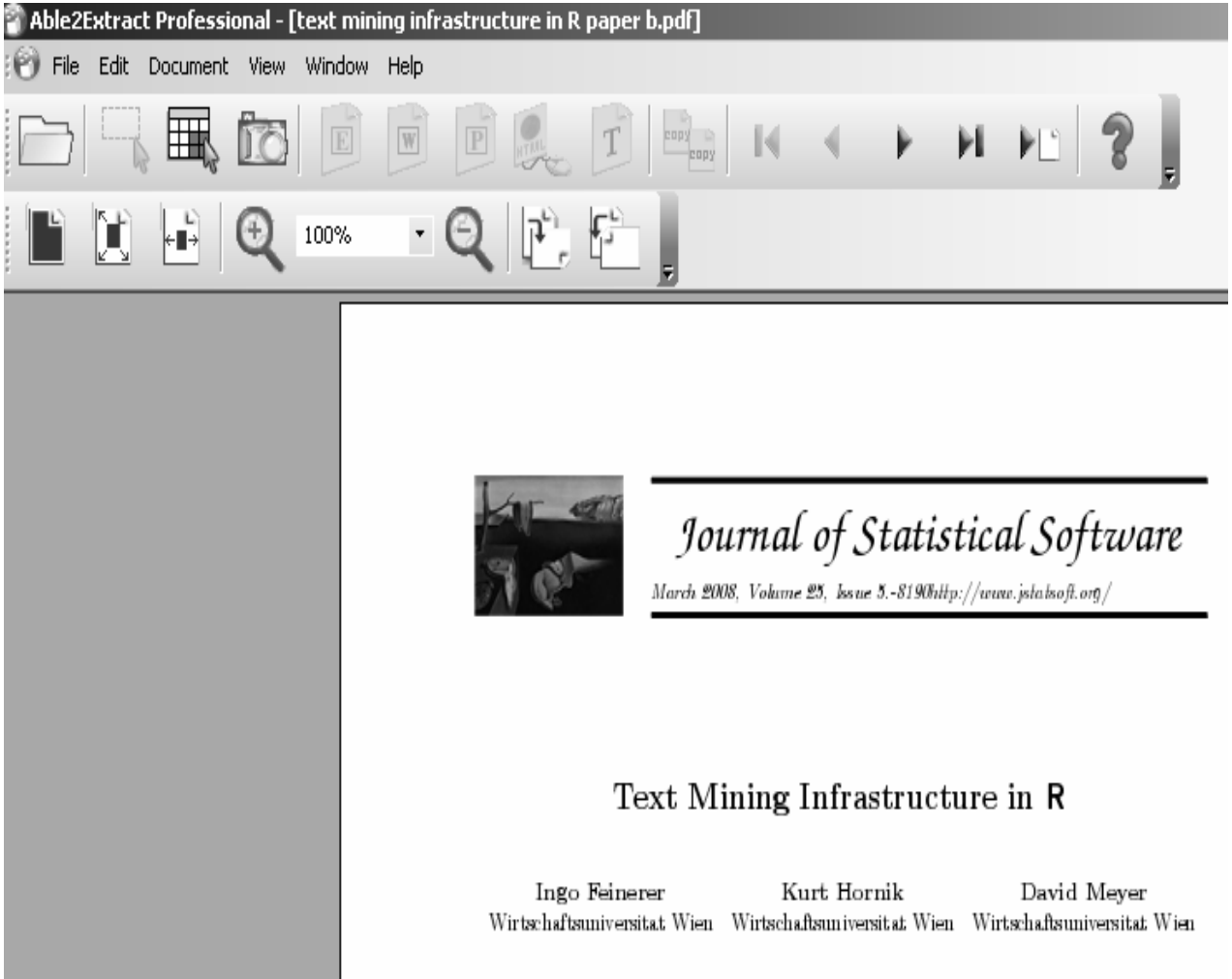
Prepared by
Louise Francis
Francis Analytics and Actuarial Data Mining, Inc.
Oct, 2008
Louise_francis@msn.com
www.data-mines.com

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

Text Mining: Uses Growing in Many Areas

- Optical Character Recognition software used to convert image to document



Major Kinds of Modeling

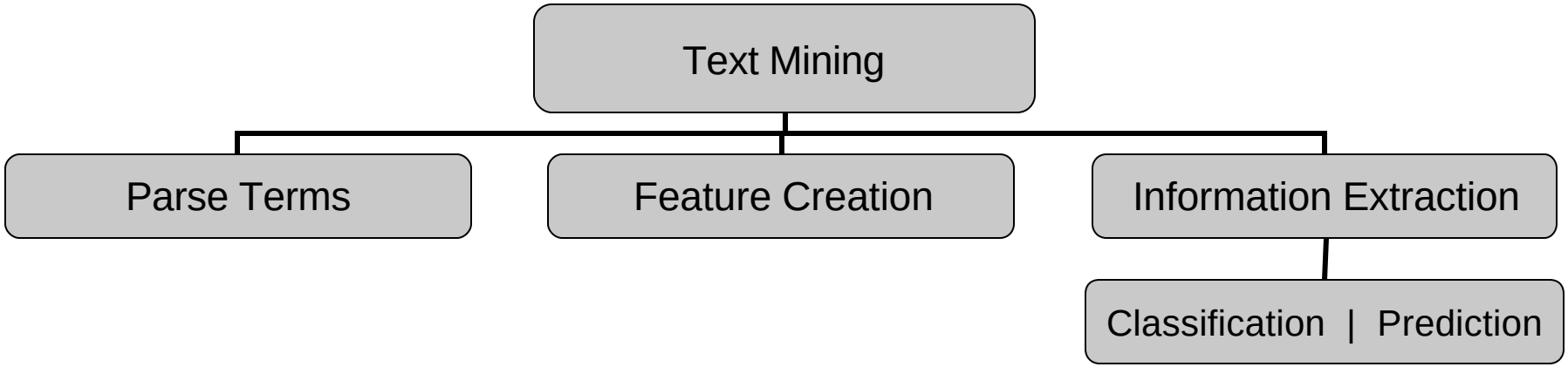
- Supervised learning
  - Most common situation
  - A dependent variable
    - Frequency
    - Loss ratio
    - Fraud/no fraud
  - Some methods
    - Regression
    - CART
    - Some neural networks
- Unsupervised learning
  - No dependent variable
  - Group like records together
    - A group of claims 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

	Analysis Types	***Non-novel information***	***Novel information***	***Comment***
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

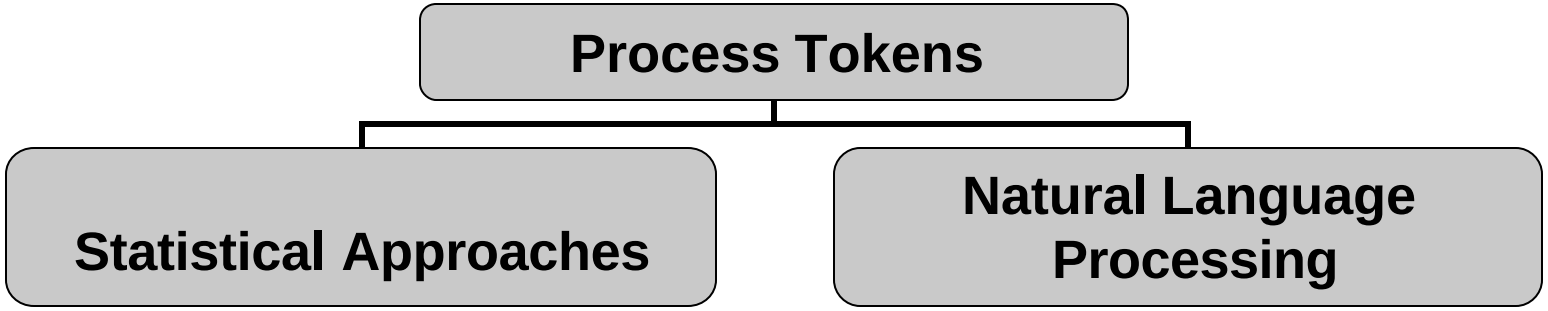
```
# parse out terms
for (i in 1:Linecount)
{
        # first word
        if (Position[i,1]==0) Terms[i,1]<-Description[i]  else if (Position[i,1]>0)
        Terms[i,1]<-substring(Description[i],1,Position[i,j]-1)
        for (j in 1:wordcount)
        {
        if (Position[i,j]>0)
        {
    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

Processing of Tokens

Further Processing



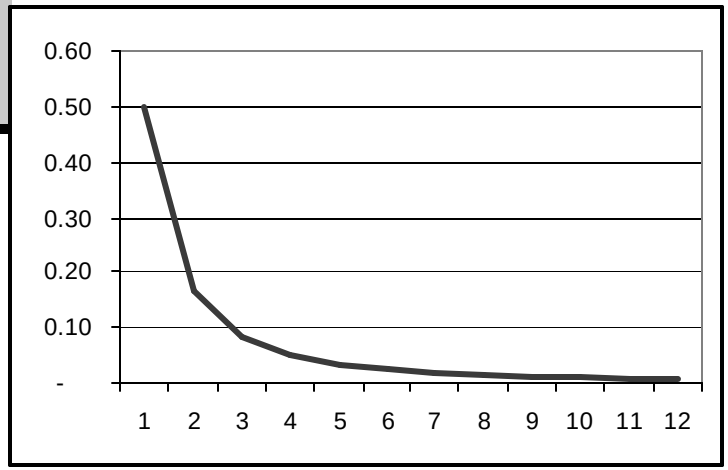
Natural Language Processing

- Draws on many disciplines
  - Artificial Intelligence
  - Linguistics
  - Statistics
  - Speech Recognition
- Its use in text mining is focuses on understanding the structure of language

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

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

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

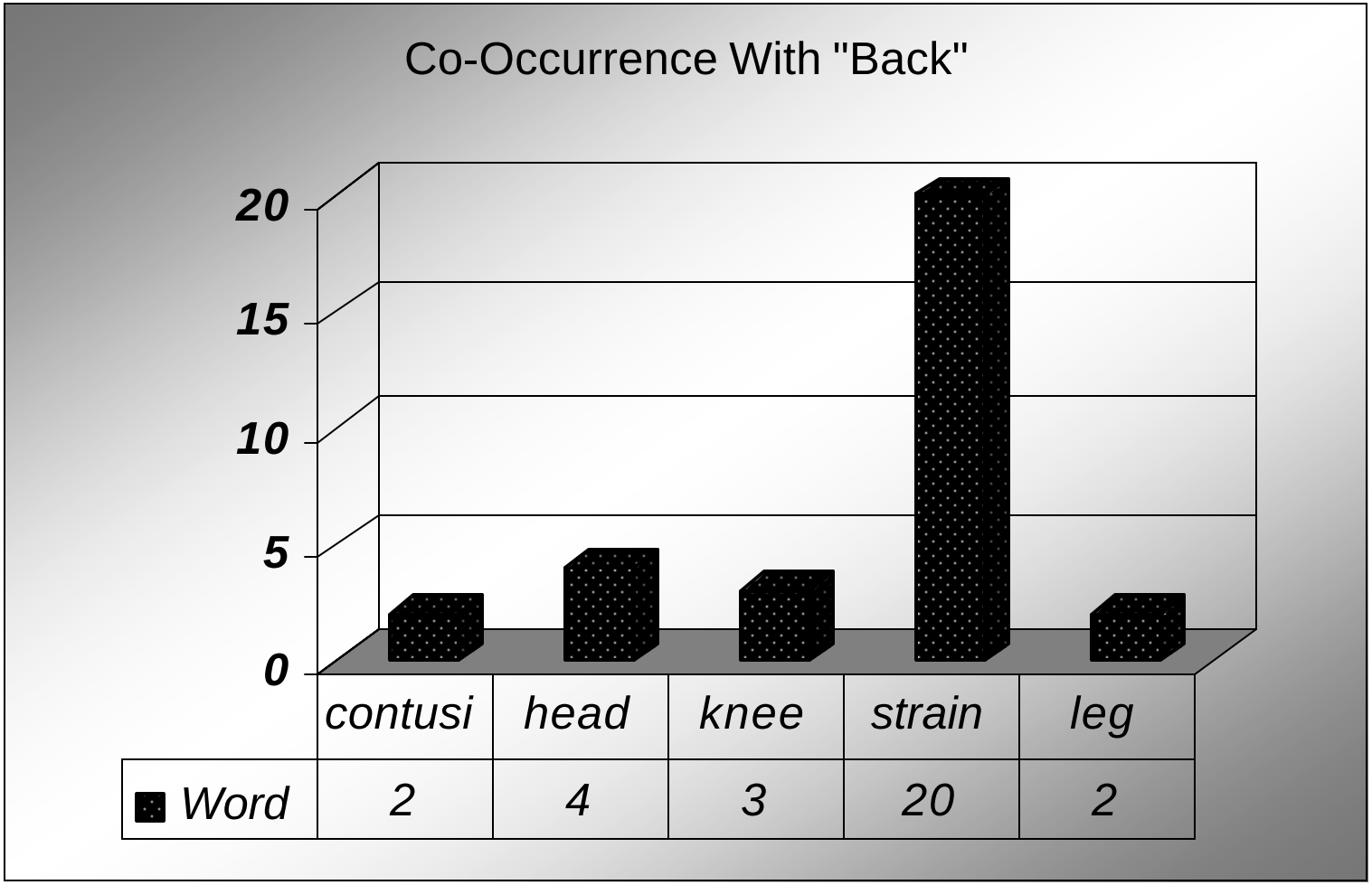
Collocation

- Multiword units, word that go together, phrases with recognized meaning
- Examples from Oct 1 newspaper
  - Philadelphia Inquirer
  - FDIC (Federal Deposit Insurance Corporation)
  - Wall Street
  - New Jersey
  - buffer zone

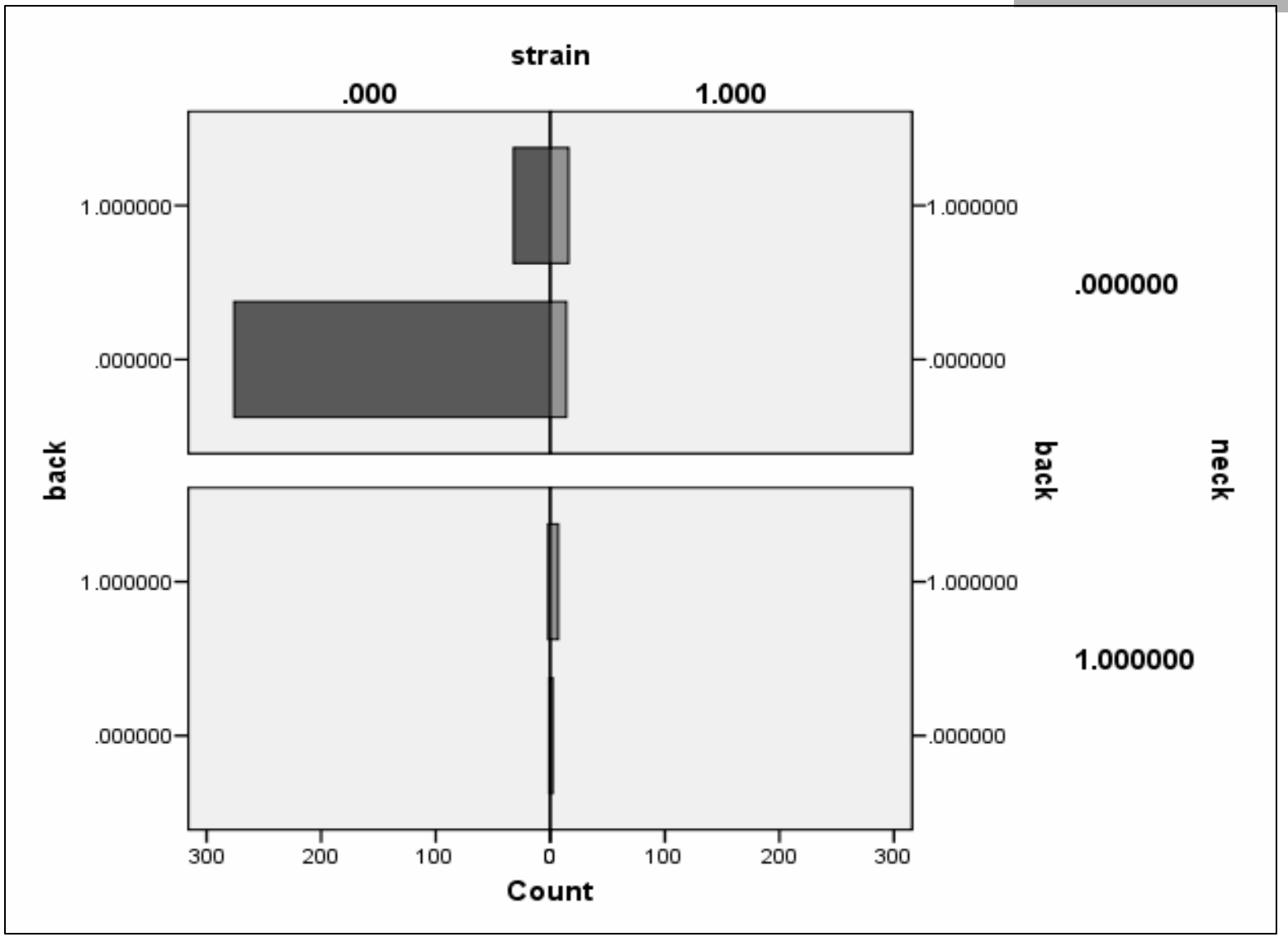
Concordances

- 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

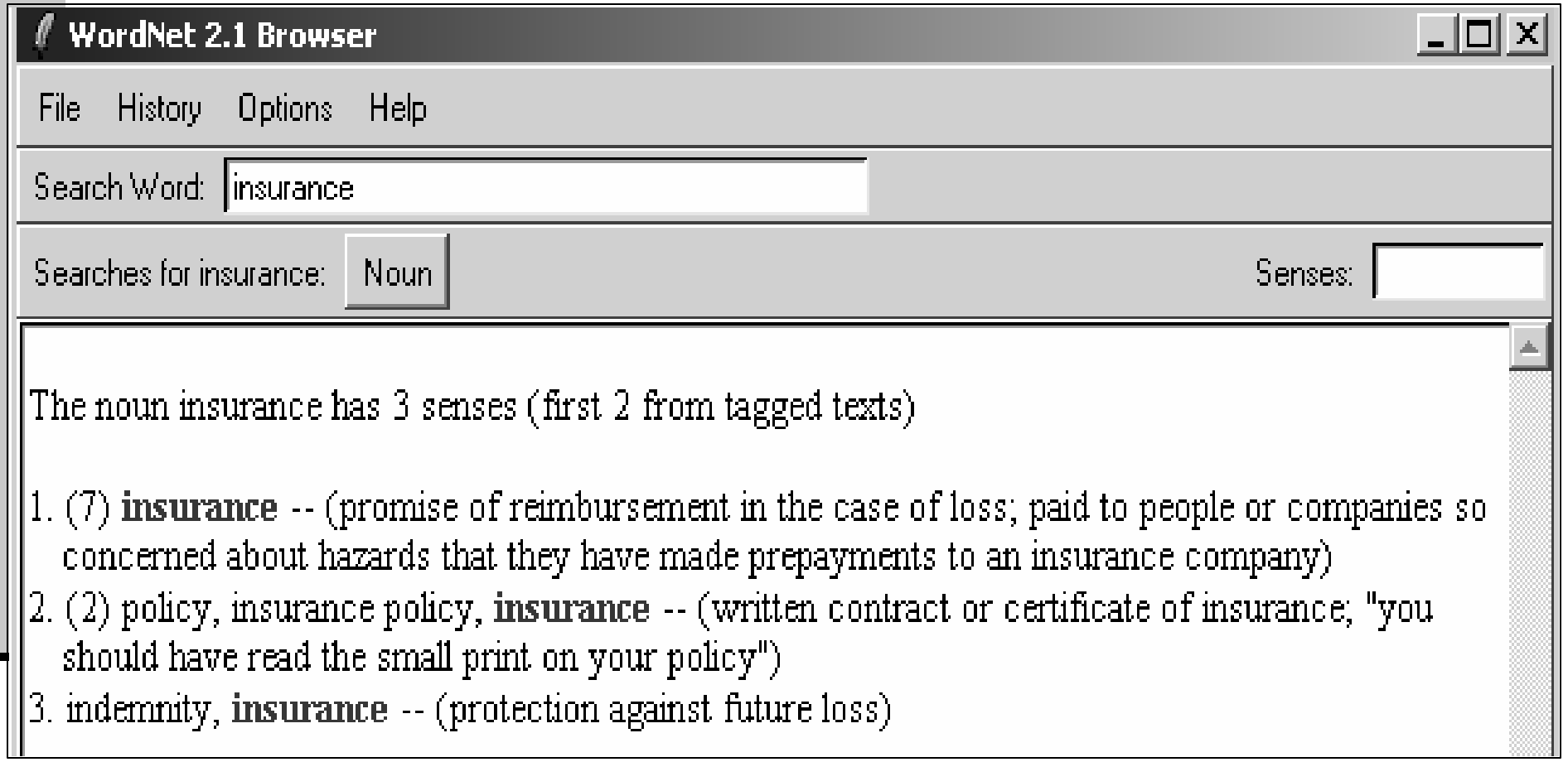
Semantics

- 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 Insurance

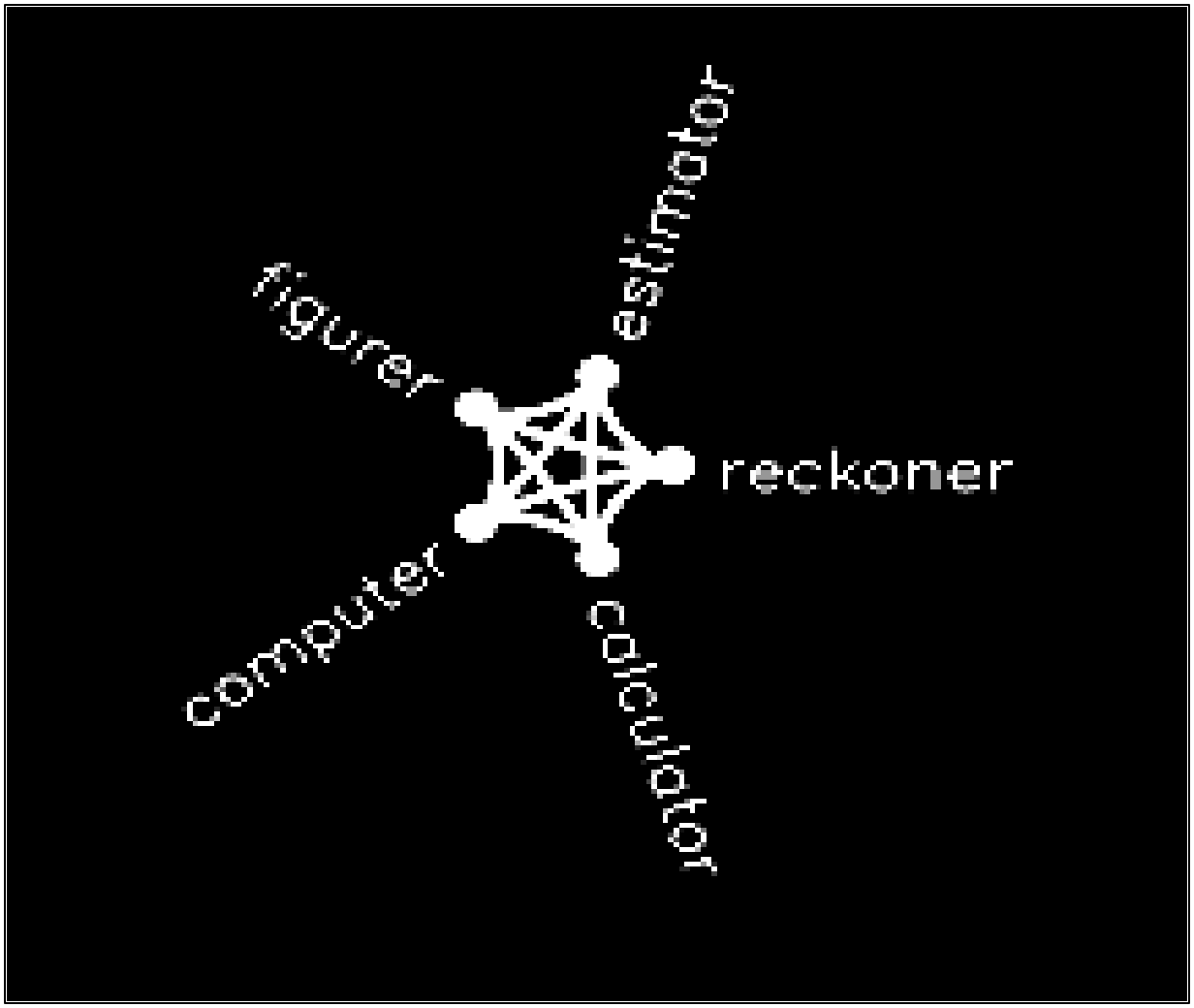


http://wordnet.princeton.edu

Wordnet Visualizations for Underwriter

- http://www.ug.it.usyd.edu.au/~smer3502/assignment3/form.html

Hypernyms of Actuary



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

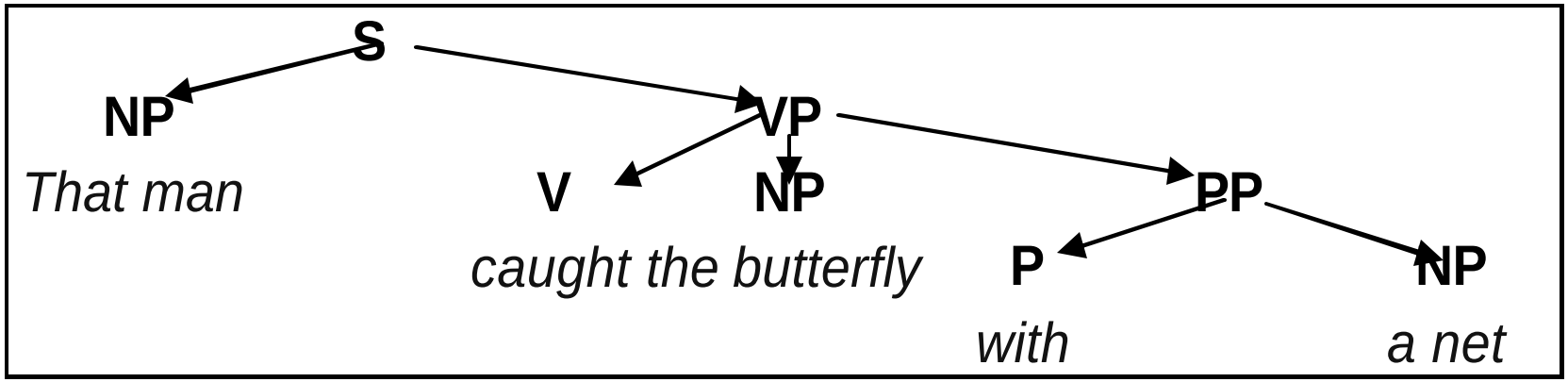
Part of Speech Morphology

- 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

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

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

- Manhattan Distance

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

K-Means Clustering

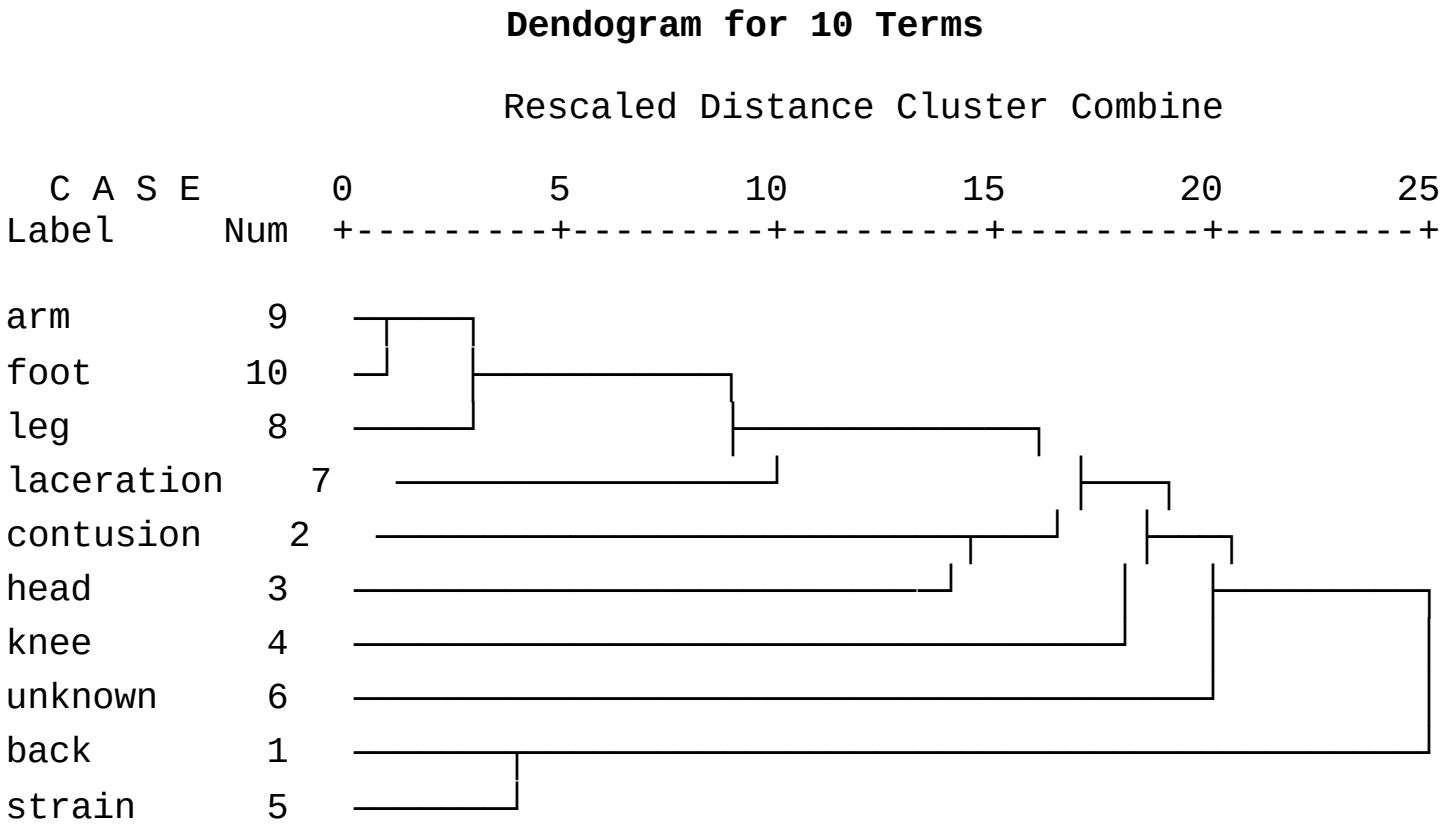
- Determine ahead of time how many clusters or groups you want
- Use dissimilarity measure to assign all records to one of the clusters

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



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

$$Y = B_0 + B_1 Attorney + B_2 Injury\_Group$$

Y = 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

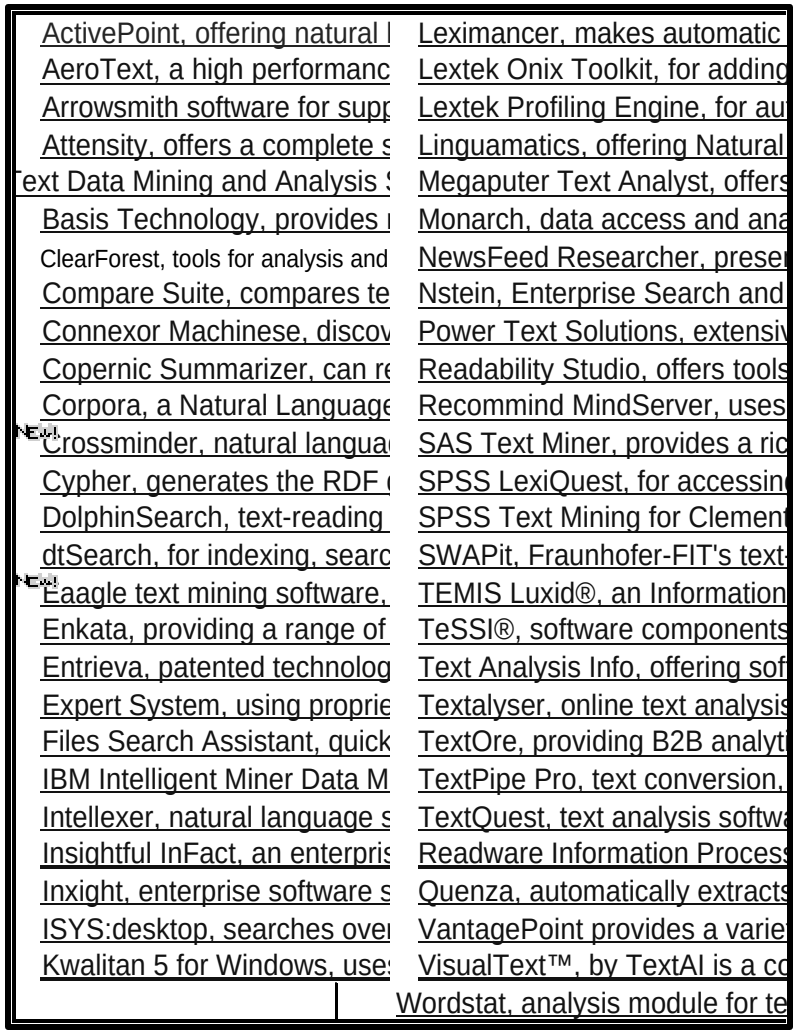
Perl for Text Processing

- Free open source programming language
- www.perl.org
- Used a lot for text processing
- *Perl for Dummies* gives a good introduction

Perl Functions for Parsing

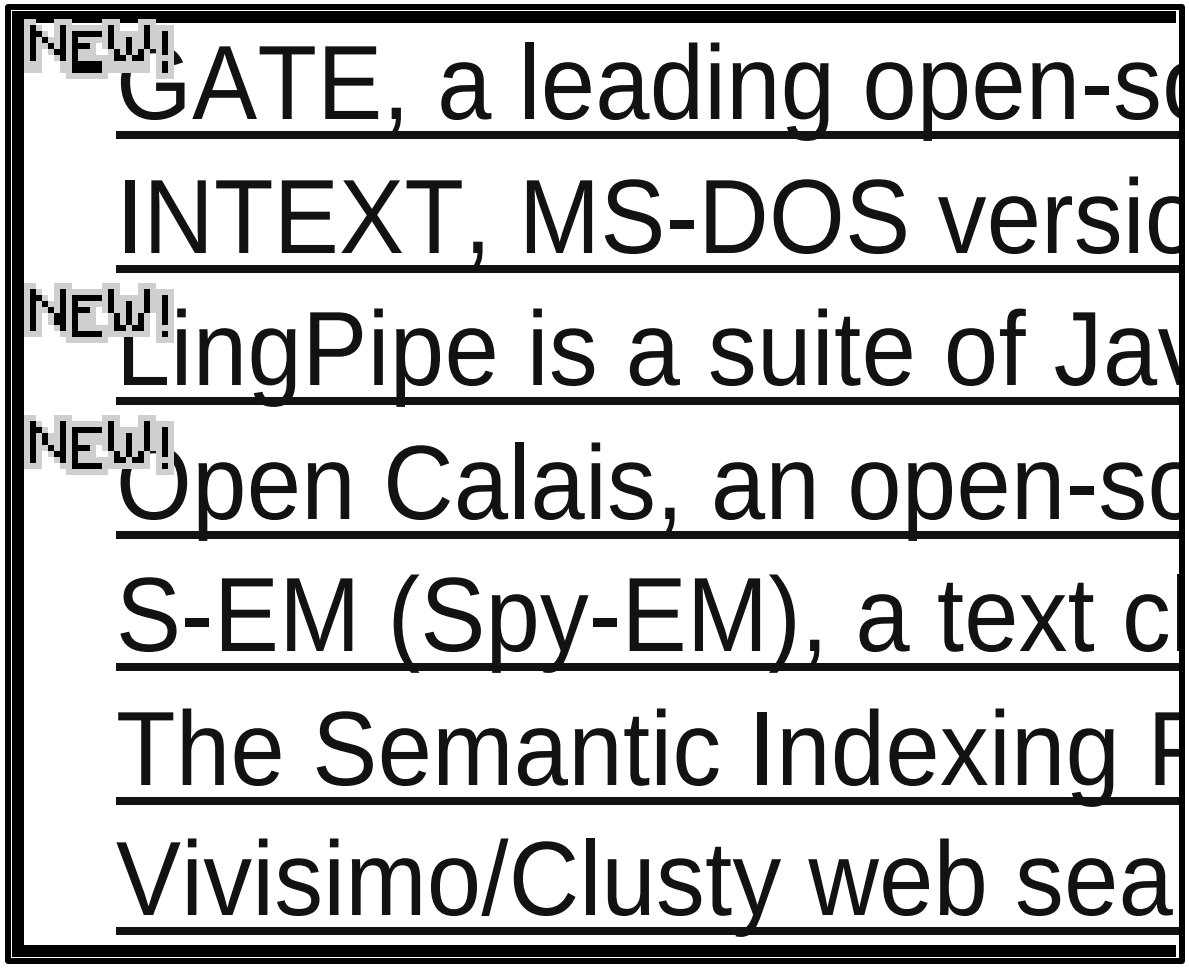
```
$TheFile ="GLClaims.txt";
$Linelength=length($TheFile);
open(INFILE, $TheFile) or die "File not found";
# Initialize variables
$Linecount=0;
@alllines=();
while(<INFILE>){
  $Theline=$_;
  chomp($Theline);
  $Linecount = $Linecount+1;
  $Linelength=length($Theline);
  @Newitems = split(/ /,$Theline);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

Commercial Software for Text Mining



From www.kdnuggets.com

Free Software for Text Mining



From www.kdnuggets.com

References

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

Questions?