

---

# An Introduction to Text Mining

## CAS 2008 Predictive Modeling Seminar

Prepared by  
Louise Francis  
Francis Analytics and Actuarial Data Mining, Inc.  
Oct, 2008  
[Louise\\_francis@msn.com](mailto:Louise_francis@msn.com)  
[www.data-mines.com](http://www.data-mines.com)



# Objectives

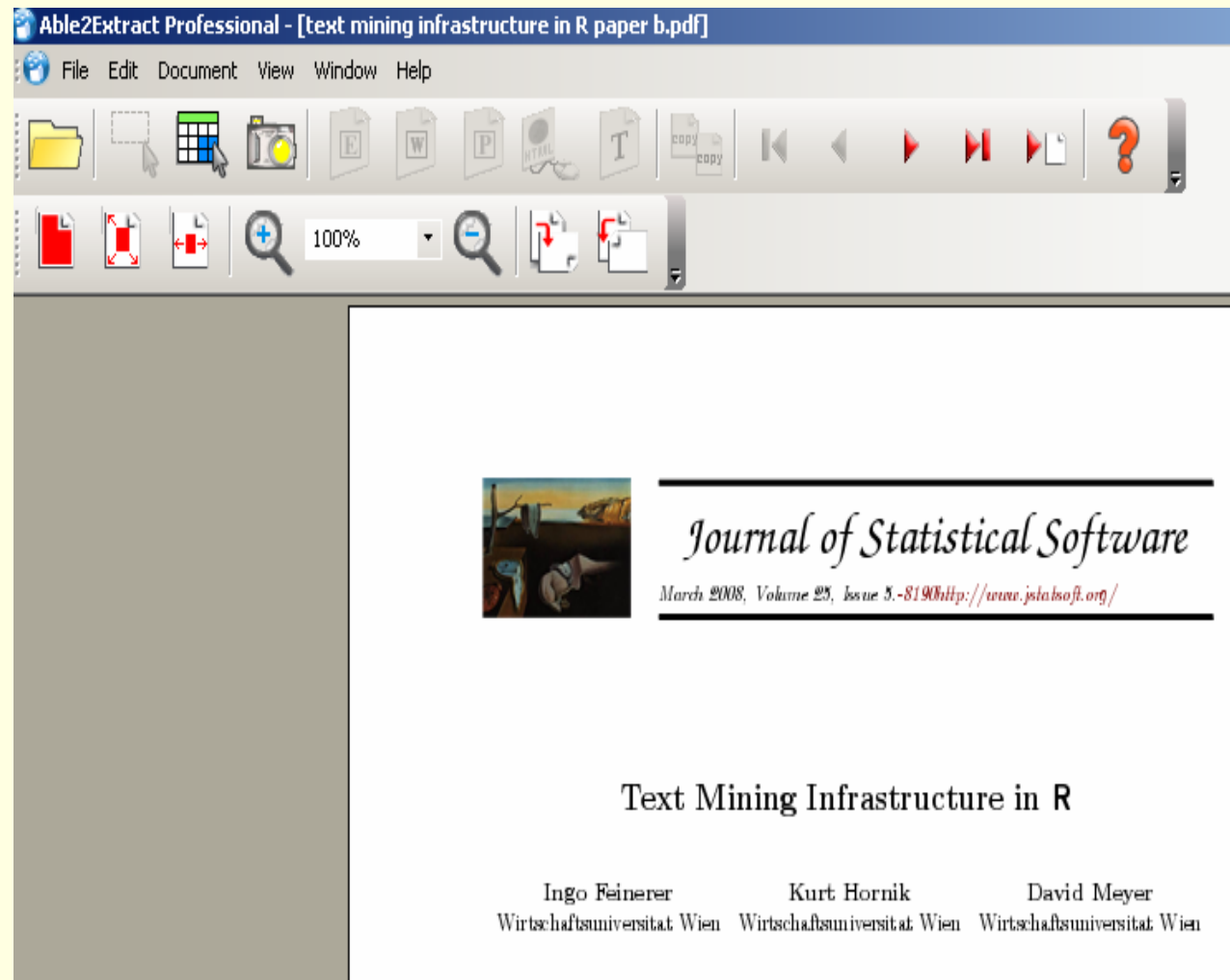
---

- 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



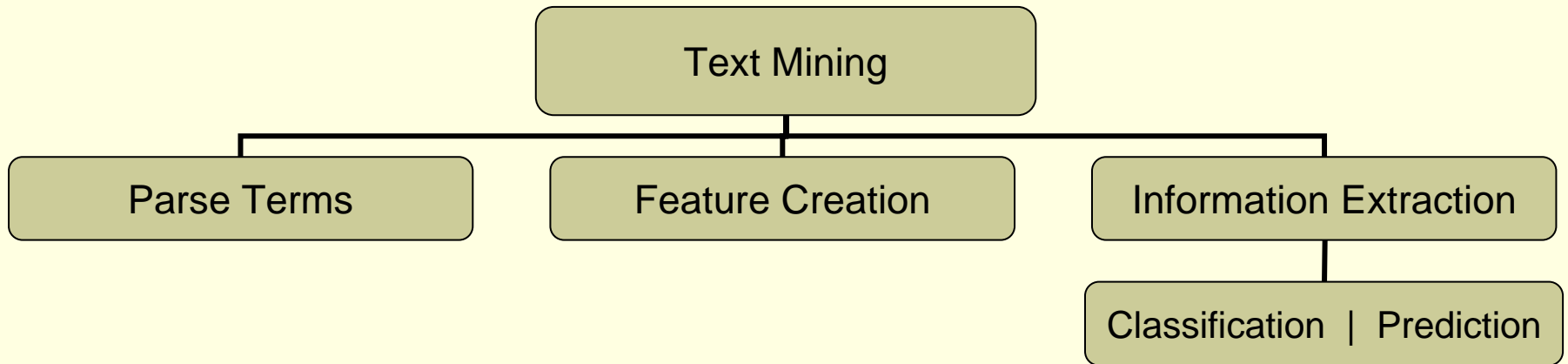
# Text Mining vs. Data Mining

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

modified from Manning/Hearst

# Text Mining Process

---





# String Processing

# Example: Claim Description Field

<b>INJURY DESCRIPTION</b>
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)
<b>BROKEN ANKLE AND SPRAINED WRIST</b>	<b>31</b>	7	BROKEN	24
<b>Remainder 1</b>		<b>2<sup>nd</sup> Blank</b>	<b>2<sup>nd</sup> Word</b>	<b>Remainder Length 2</b>
(6)		(7)	(8)	(9)
ANKLE AND SPRAINED WRIST		6	ANKLE	18
<b>Remainder 2</b>		<b>3<sup>rd</sup> Blank</b>	<b>3<sup>rd</sup> Word</b>	<b>Remainder Length 3</b>
(10)		(11)	(12)	(13)
AND SPRAINED WRIST		4	AND	14
<b>Remainder 3</b>		<b>4<sup>th</sup> Blank</b>	<b>4<sup>th</sup> Word</b>	<b>Remainder Length 4</b>
(14)		(15)	(16)	(17)
SPRAINED WRIST		9	SPRAINED	5
<b>Remainder 4</b>		<b>5<sup>th</sup> Blank</b>	<b>5<sup>th</sup> Word</b>	
(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
}
}
```

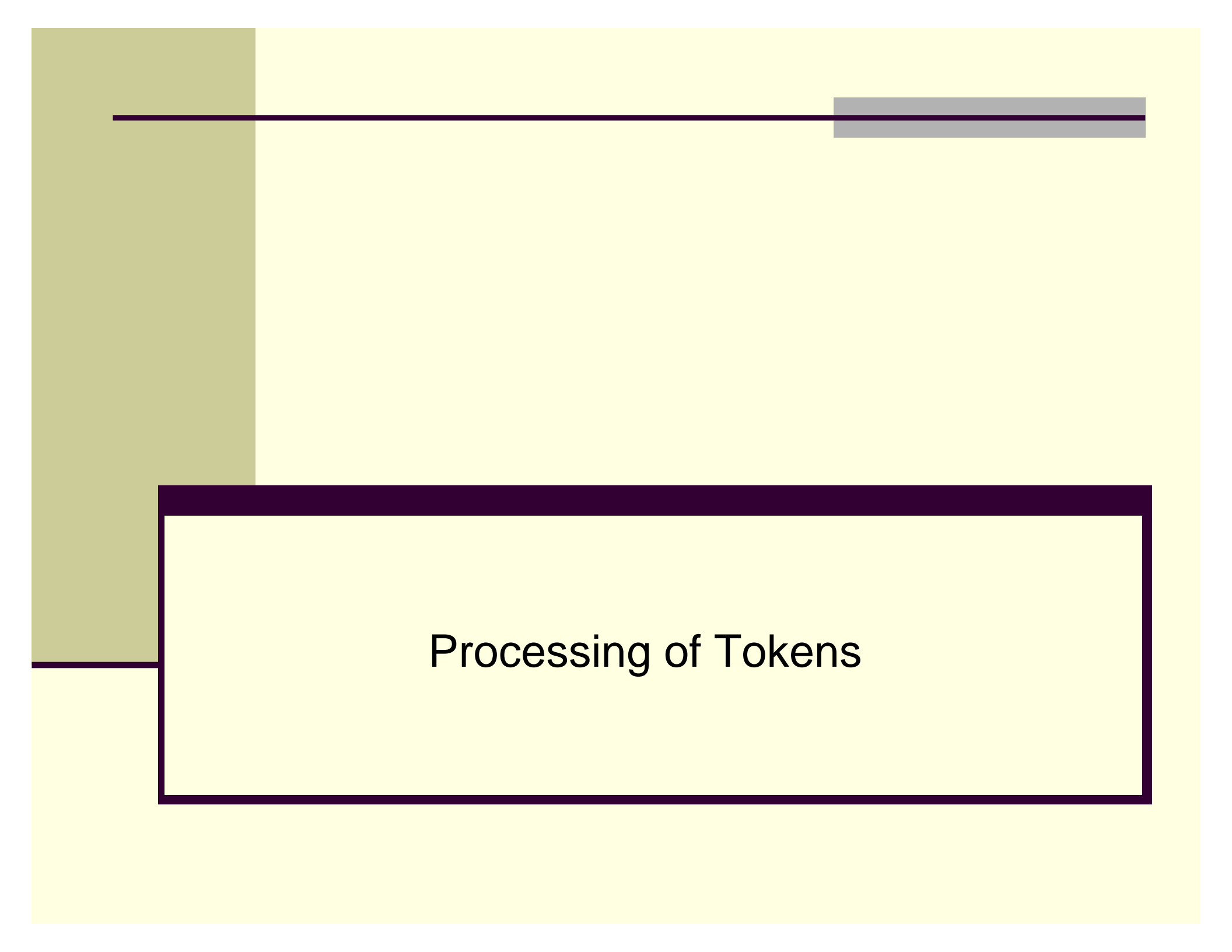
# Get Words

---

```
# 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

<b>INJURY DESCRIPTION</b>	<b>BROKEN</b>	<b>ANKLE</b>	<b>AND</b>	<b>SPRAINED</b>	<b>W R I S T</b>	<b>F O O T</b>	<b>CONTU - SION</b>	<b>UNKNOWN</b>	<b>N E C K</b>	<b>BACK</b>	<b>STRAIN</b>
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

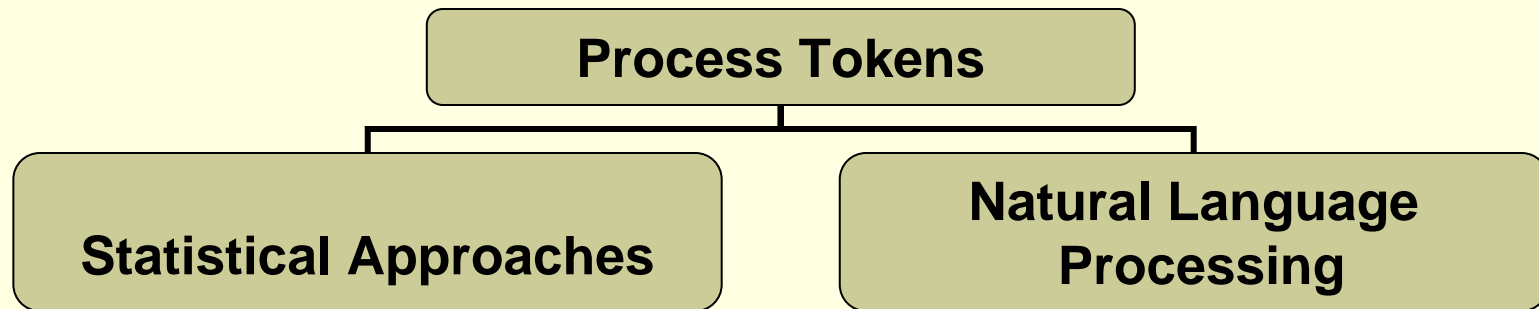


# Processing of Tokens



# Further Processing

---



# Natural Language Processing

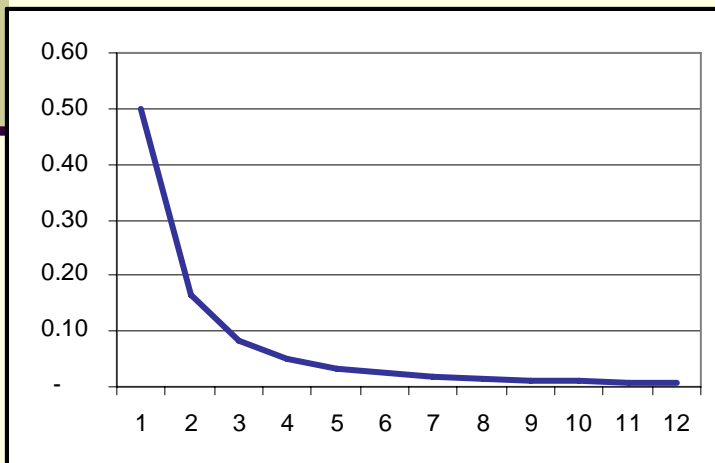
---

- 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

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

# 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

---

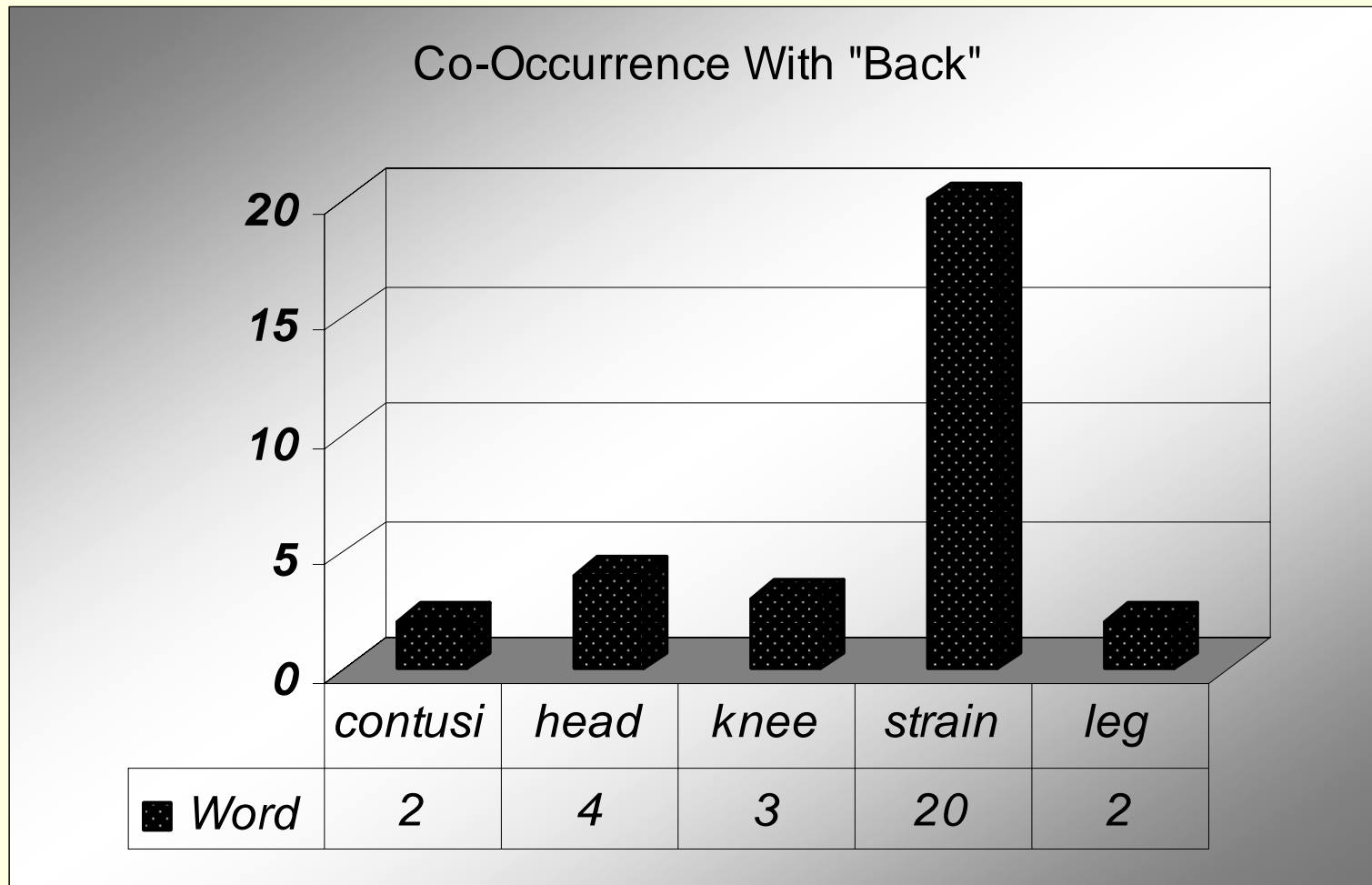
- Multiword units, word that go together, phrases with recognized meaning
- Examples from Oct 1 newspaper
  - Philadelphia Inquirer
  - FDIC (Federal Deposit Insurance Corporation)
  - Wall Street
  - New Jersey
  - buffer zone

# Concordances

---

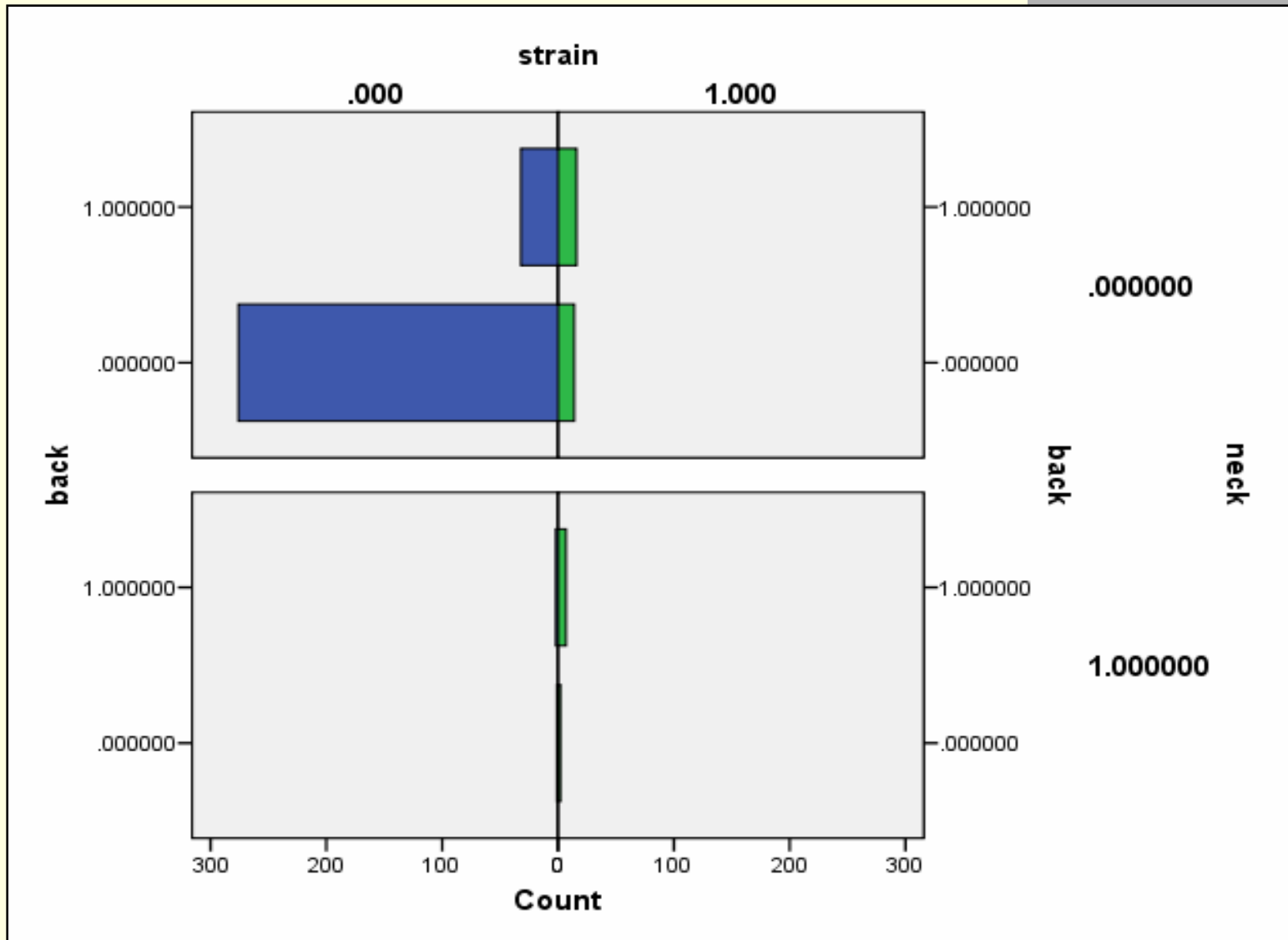
- Finding contexts in which verbs appear
- Use key word in context
- Lists all occurrences of the word and the words that occur with it.

# 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

---

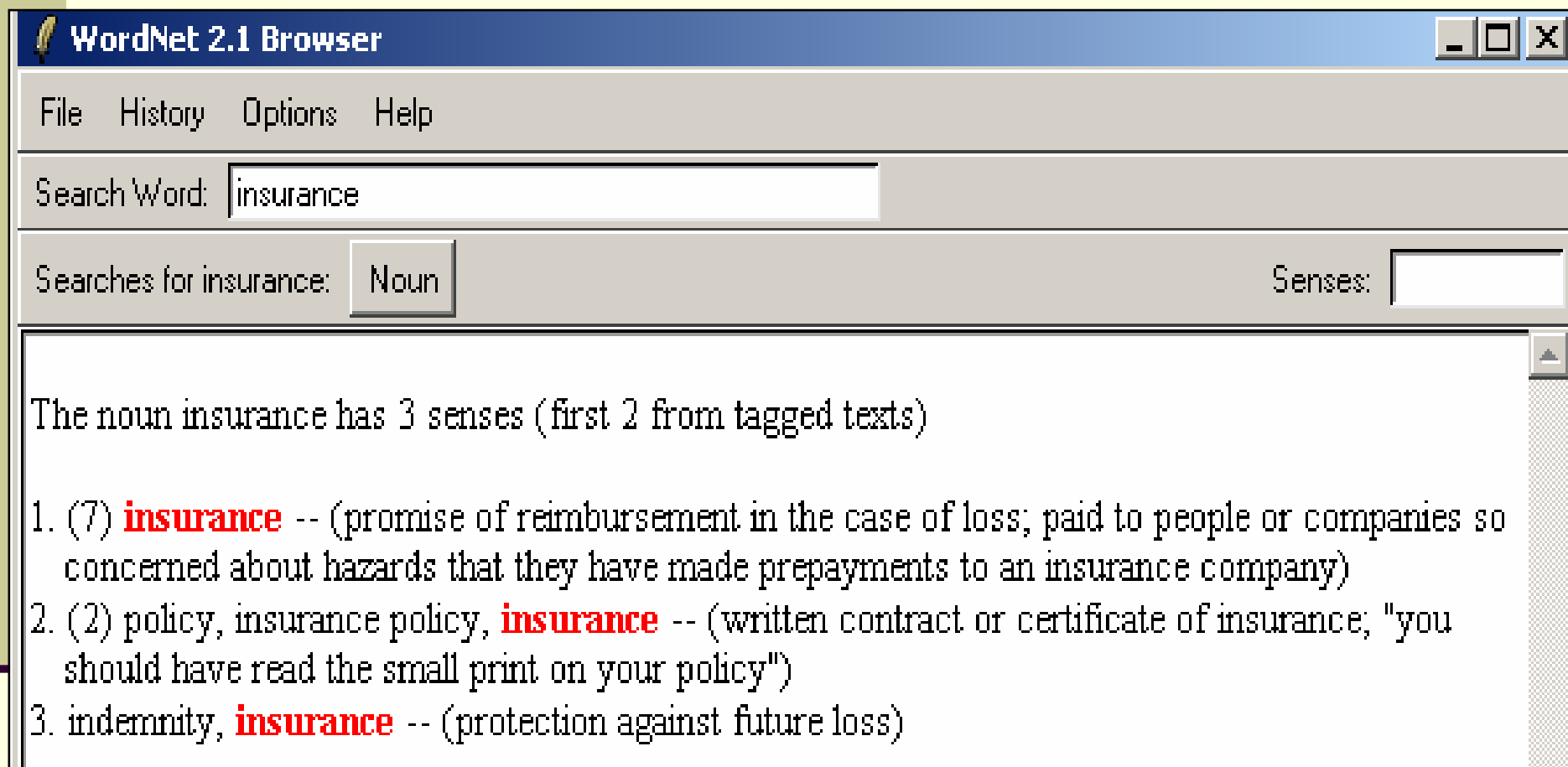
- Meaning of words, phrases, sentences and other language structures
  - Lexical semantics
    - Meaning of individual words
    - Examples; synonyms, antonyms
  - Meanings of combinations of words

# Wordnet

---

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

# Wordnet Entry for Insurance



The screenshot shows the WordNet 2.1 Browser interface. The title bar reads "WordNet 2.1 Browser". The menu bar includes "File", "History", "Options", and "Help". The search bar contains the text "insurance". Below the search bar, it shows "Searches for insurance: Noun" and "Senses: [ ]". The main content area displays the following text:

The noun insurance has 3 senses (first 2 from tagged texts)

1. (7) **insurance** -- (promise of reimbursement in the case of loss; paid to people or companies so concerned about hazards that they have made prepayments to an insurance company)
2. (2) policy, insurance policy, **insurance** -- (written contract or certificate of insurance; "you should have read the small print on your policy")
3. indemnity, **insurance** -- (protection against future loss)

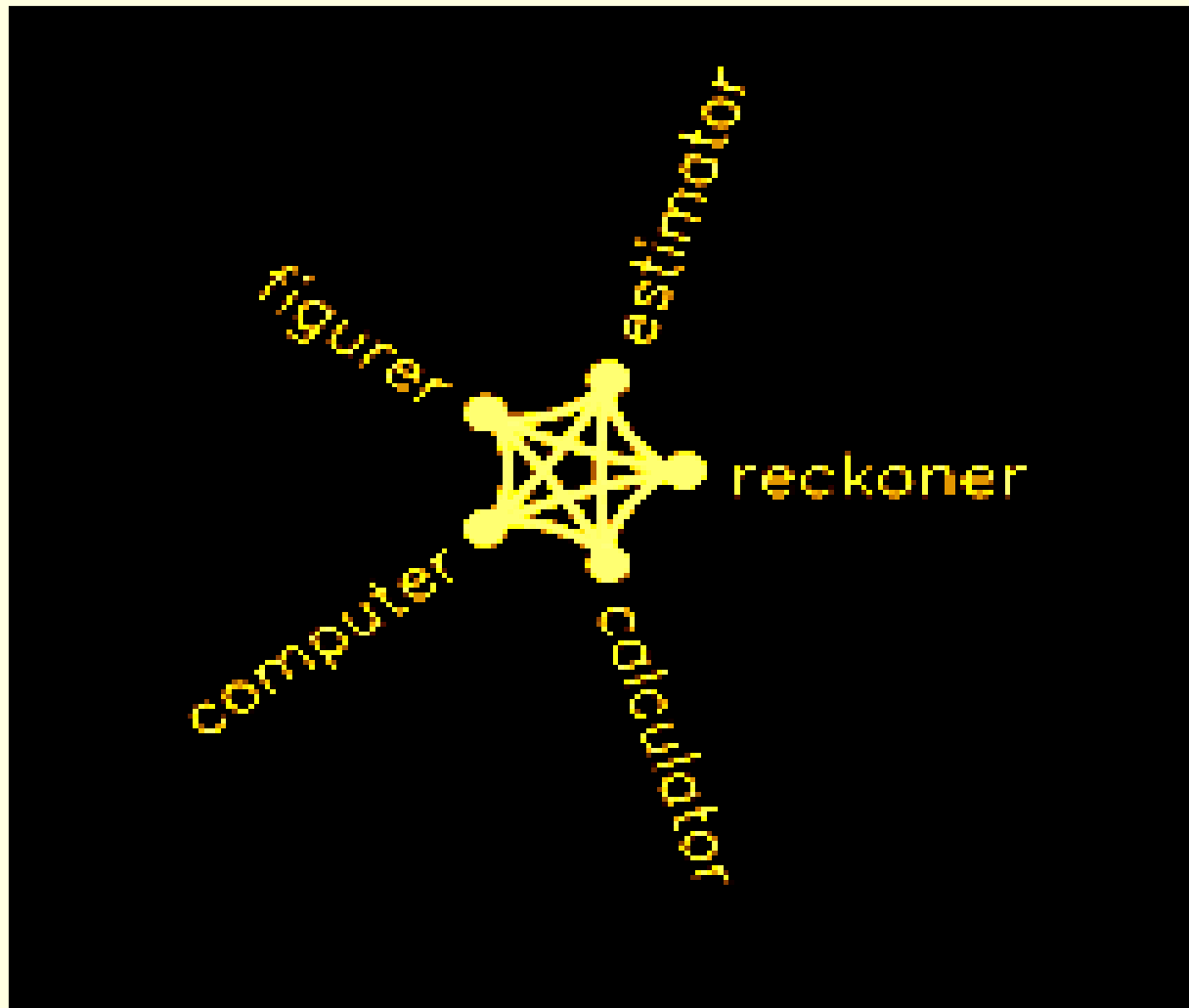
<http://wordnet.princeton.edu>

# Wordnet Visualizations for Underwriter



# 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

---

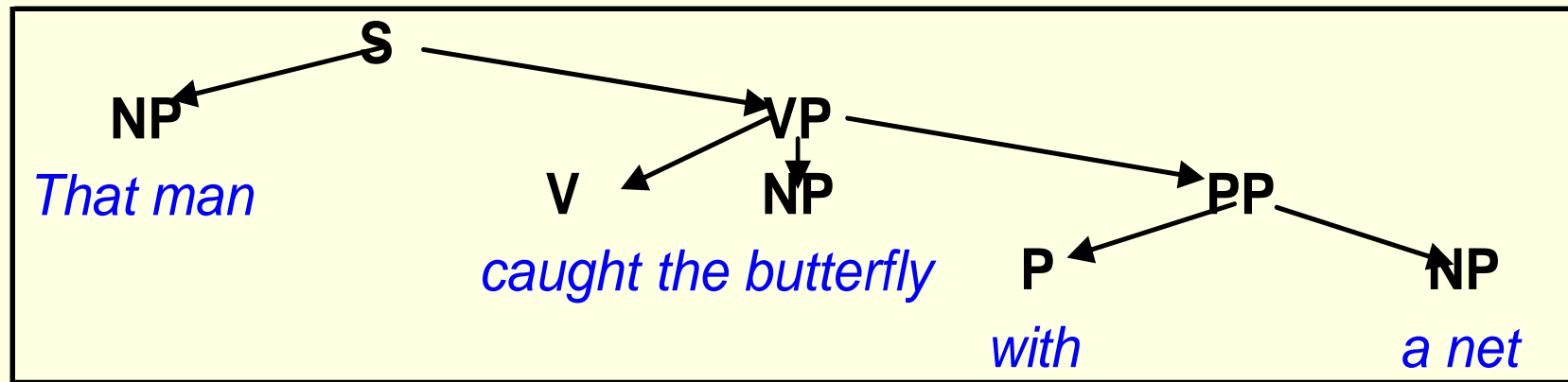
- Parts of Speech (POS)
  - Noun
  - Verb
  - Adjective
    - These are open or lexical categories that have large numbers of members and new members frequently added
  - Also prepositions and determiners
    - Of, on, the, a
    - Generally closed categories

# 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

	CLAIM NUMBER	DATE OF LOSS	STATUS	INCURRED LOSS
				VARIABLES
RECORDS	1998001	09/15/97	C	407.81
	1998002	09/25/97	C	0.00
	1998003	09/28/97	C	0.00
	1998004	09/29/97	C	8,247.18
	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	C	0.00
	1998009	10/29/97	C	21,211.88
	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	451.88
	1998014	11/04/97	C	0.00
	1998015	11/04/97	C	0.00
	1998016	11/08/97	C	15,903.88
	1998017	11/11/97	C	485.10

# 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

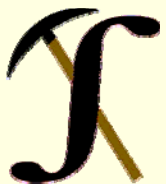
- Determine ahead of time how many clusters or groups you want
- Use dissimilarity measure to assign all records to one of the clusters

<b>Cluster Number</b>	<b>back</b>	<b>contusion</b>	<b>head</b>	<b>knee</b>	<b>strain</b>	<b>unknown</b>	<b>laceration</b>
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

---

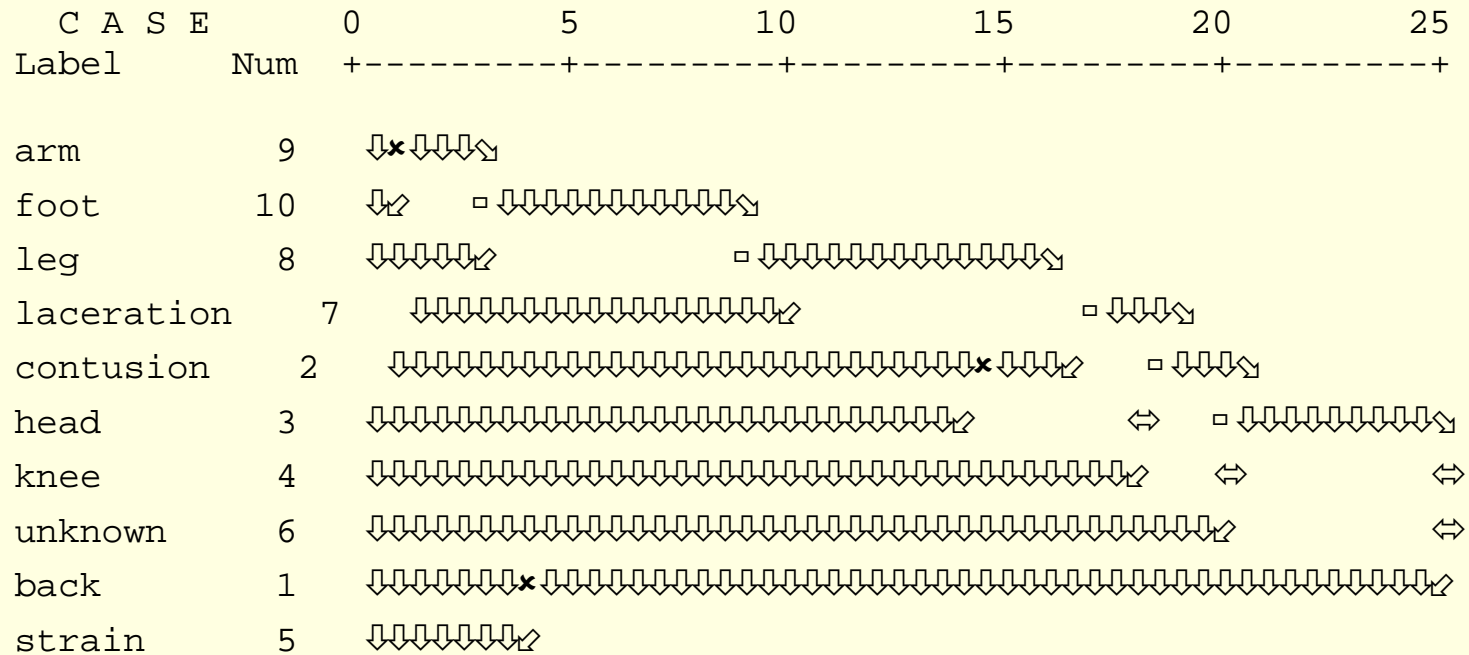
- A stepwise procedure
- At beginning, each records is its own cluster
- Combine the most similar records into a single cluster
- Repeat process until there is only one cluster with every record in it



# 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

---

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

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

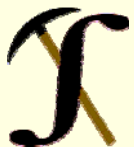
**Mean Probability of Serious Claim vs. Actual Value**

	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](http://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

<a href="#">ActivePoint</a> , offering natural language processing	<a href="#">Leximancer</a> , makes automatic text analysis
<a href="#">AeroText</a> , a high performance text processing engine	<a href="#">Lextek Onix Toolkit</a> , for adding text mining capabilities
<a href="#">Arrowsmith</a> software for support of text mining	<a href="#">Lextek Profiling Engine</a> , for automatic text analysis
<a href="#">Attensity</a> , offers a complete suite of text mining tools	<a href="#">Linguamatics</a> , offering Natural Language Processing
<a href="#">Text Data Mining and Analysis</a> (TDM)	<a href="#">Megaputer Text Analyst</a> , offers text mining capabilities
<a href="#">Basis Technology</a> , provides text mining solutions	<a href="#">Monarch</a> , data access and analysis
<a href="#">ClearForest</a> , tools for analysis and text mining	<a href="#">NewsFeed Researcher</a> , presents text mining capabilities
<a href="#">Compare Suite</a> , compares text documents	<a href="#">Nstein</a> , Enterprise Search and text mining
<a href="#">Connexor MachineSense</a> , discovers patterns in text	<a href="#">Power Text Solutions</a> , extensive text mining capabilities
<a href="#">Copernic Summarizer</a> , can reduce text to its essence	<a href="#">Readability Studio</a> , offers tools for text analysis
<a href="#">Corpora</a> , a Natural Language Processing tool	<a href="#">Recommind MindServer</a> , uses text mining for recommendations
<a href="#">Crossminder</a> , natural language processing tool	<a href="#">SAS Text Miner</a> , provides a rich set of text mining capabilities
<a href="#">Cypher</a> , generates the RDF graph from text	<a href="#">SPSS LexiQuest</a> , for accessing text mining capabilities
<a href="#">DolphinSearch</a> , text-reading and search engine	<a href="#">SPSS Text Mining for Clementine</a> , text mining capabilities
<a href="#">dtSearch</a> , for indexing, searching and text mining	<a href="#">SWAPit</a> , Fraunhofer-FIT's text mining tool
<a href="#">Eagle</a> text mining software, for text analysis	<a href="#">TEMIS Luxid®</a> , an Information Management tool
<a href="#">Enkata</a> , providing a range of text mining capabilities	<a href="#">TeSSI®</a> , software components for text mining
<a href="#">Entrieva</a> , patented technology for text mining	<a href="#">Text Analysis Info</a> , offering software for text analysis
<a href="#">Expert System</a> , using proprietary text mining capabilities	<a href="#">Textalyser</a> , online text analysis tool
<a href="#">Files Search Assistant</a> , quick search and text mining	<a href="#">TextOre</a> , providing B2B analytical capabilities
<a href="#">IBM Intelligent Miner Data Mining</a> , text mining capabilities	<a href="#">TextPipe Pro</a> , text conversion and analysis
<a href="#">Intellexer</a> , natural language processing tool	<a href="#">TextQuest</a> , text analysis software
<a href="#">Insightful InFact</a> , an enterprise text mining solution	<a href="#">Readware Information Process</a> , text mining capabilities
<a href="#">Inxight</a> , enterprise software for text mining	<a href="#">Quenza</a> , automatically extracts text mining capabilities
<a href="#">ISYS:desktop</a> , searches over text documents	<a href="#">VantagePoint</a> provides a variety of text mining capabilities
<a href="#">Kwalitan 5 for Windows</a> , uses text mining capabilities	<a href="#">VisualText™</a> , by TextAI is a commercial text mining tool
	<a href="#">Wordstat</a> , 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
- Manning and Schultze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999





# Questions?

---