Taking Customers at their Word

Natural Language Processing
And the Analysis of Text Data

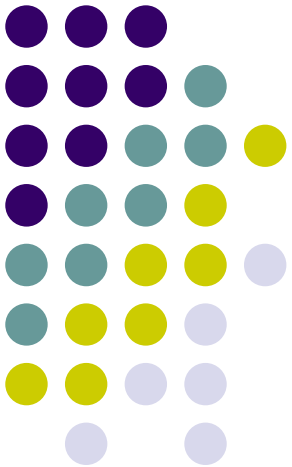
Lynd Bacon & Nick Haddock
2004 AMA ART Forum

www.lba.com

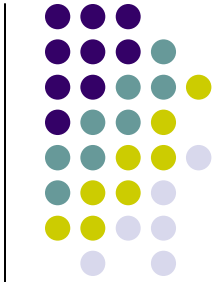
Solutions for Business Growth
lbacon@lba.com

www.atomicintelligence.com

Analysis of Unstructured Content
nick@atomicintelligence.com

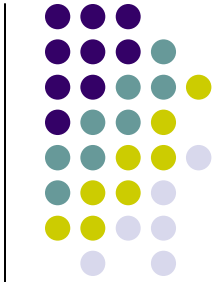


Agenda



- Nature of language
- State of the art in NLP
- Text data in marketing research
- Text analysis methods
- Latent variable modeling example using text and quantitative data
- Take-aways

The crux of the problem

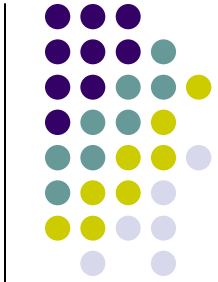


Time flies like arrow.

Fruit flies like a banana.

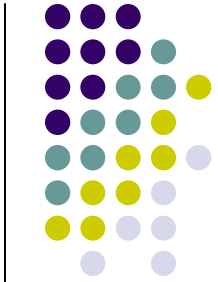
Groucho Marx

What is NLP



- Long-term goals
  - Text understanding
  - Text generation
  - Spoken dialogue
  - Machine translation
- Focused, practical goals
  - Fact extraction from news articles
  - Analyzing authorship
  - Question-answering from the web
- Sub-problems
  - Morphological analysis
  - Part of speech tagging
  - Grammar and parsing
  - Word sense disambiguation
  - Semantic interpretation
  - Discourse and reference
- Approaches
  - Rule-based
    - Linguistic and world knowledge encoded by intuited rules and representations
  - Probabilistic models
    - Models built directly from language data

Why NLP is hard: Vast ambiguity



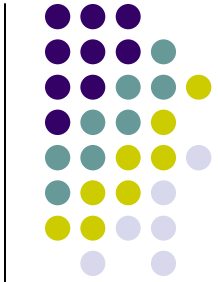
Headline	Ambiguities
Ban on Nude Dancing on Governor's desk	Syntactic parsing, sem. role: on Governor's desk
Iraqi Head Seeks Arms	Word-sense: head, arms
Stolen Painting Found by Tree	Semantic role: by tree
Red Tape Holds Up New Bridges	Semantic interpretation: red tape, holds up
Hospitals are sued by 7 Foot Doctors	Part-of-speech, syntactic parsing: 7 foot doctors
Kids Make Nutritious Snacks	Word-sense: make

Adapted from Stanford University online course material, Spring 2004

Why NLP is hard

Ambiguous One expression can mean multiple things - Part of speech - Syntactic structure - Word-sense - Semantic interpretation - Referential ambiguity	***"a noisy fan"*** ***"carrying people"*** ***"…and the car was a headache after it ran out"*** ***"I had no input"*** ***"today" ("The problem that I had today" vs. "on the market today")***
Unrestricted and irregular One intention can be expressed in expressed in multiple (infinite) ways	***"… seats that support while driving"*** ***"… the feelings you get in your legs and buttocks area after a long trip"*** ***" … if I can drive for 6 hours and I don't hurt"***
Context-dependent - Discourse knowledge - Task knowledge - World knowledge	***"Same thing"*** ***"I agree with Allen"*** ***"I also think that's important"***
Imperfect and abbreviated	***" vehocle", "towars", "size off engine", "fuel econ", "noise @ idle", "SLOWwwww"***

State of the art in NLP

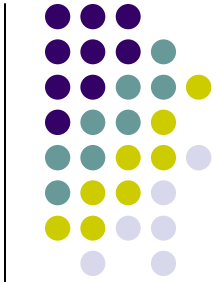


- POS tagging
  - > 95% accuracy
- Word sense
  - 75% overall, > 90% for easier cases
- Parsing
  - 90% precision/recall
- Issues remain …
  - Processing without prior training data or domain knowledge
  - Unknown words and phrases
- And fundamental problems are not yet solved
  - Semantic interpretation and reasoning

Text data in marketing research

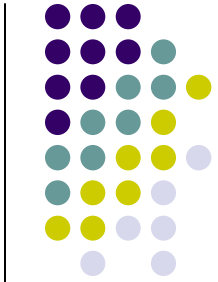
- Purposely elicited
  - Verbatim responses to open-ended questions
  - Focus group, discussion board transcripts and logs
  - IVR and telephone interview data
- “received”
  - Customer email, blog content, newsgroups, advertising content, editorial content, complaints via phone and in person

Text data analysis overview



- Data features
  - # of records/vol. of records
  - Length of records
  - # of language users/record
  - Scope of topics
  - Stability of topics
  - Availability of background/contextual data
    - Qual or quant
- Learning objectives
  - Classify topics
  - Classify authors
  - Detect particular topics
  - Detect “new” topics
  - Classify language users
  - Compare language users
- Analytic approaches
  - Pre-processing
  - Supervised
  - Unsupervised
  - Augmentation
  - Clustering and data reduction

Examples

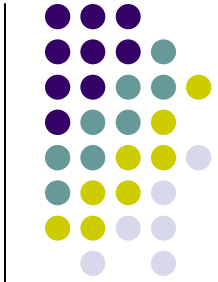


the computers main problem is it is very slow and there ore ablout only 13 icons on the main screen and 6 packs of very small memory comsuming games

Functions properly but the cost difference between a brand name PC compared with building a system yourself outweighs the benefits a brand name company can offer you if you are confident in your ability to build and maintain a pc on your own

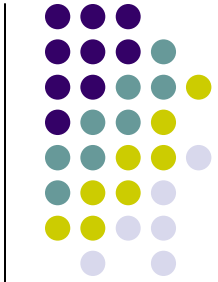
The only problem I had with this pc is, it dropped a file out of the fax-modem folder and I could only get it back by either upgrading to Windows XP or send the computer back and they choose to up gradeit, so I did.

Text analysis techniques



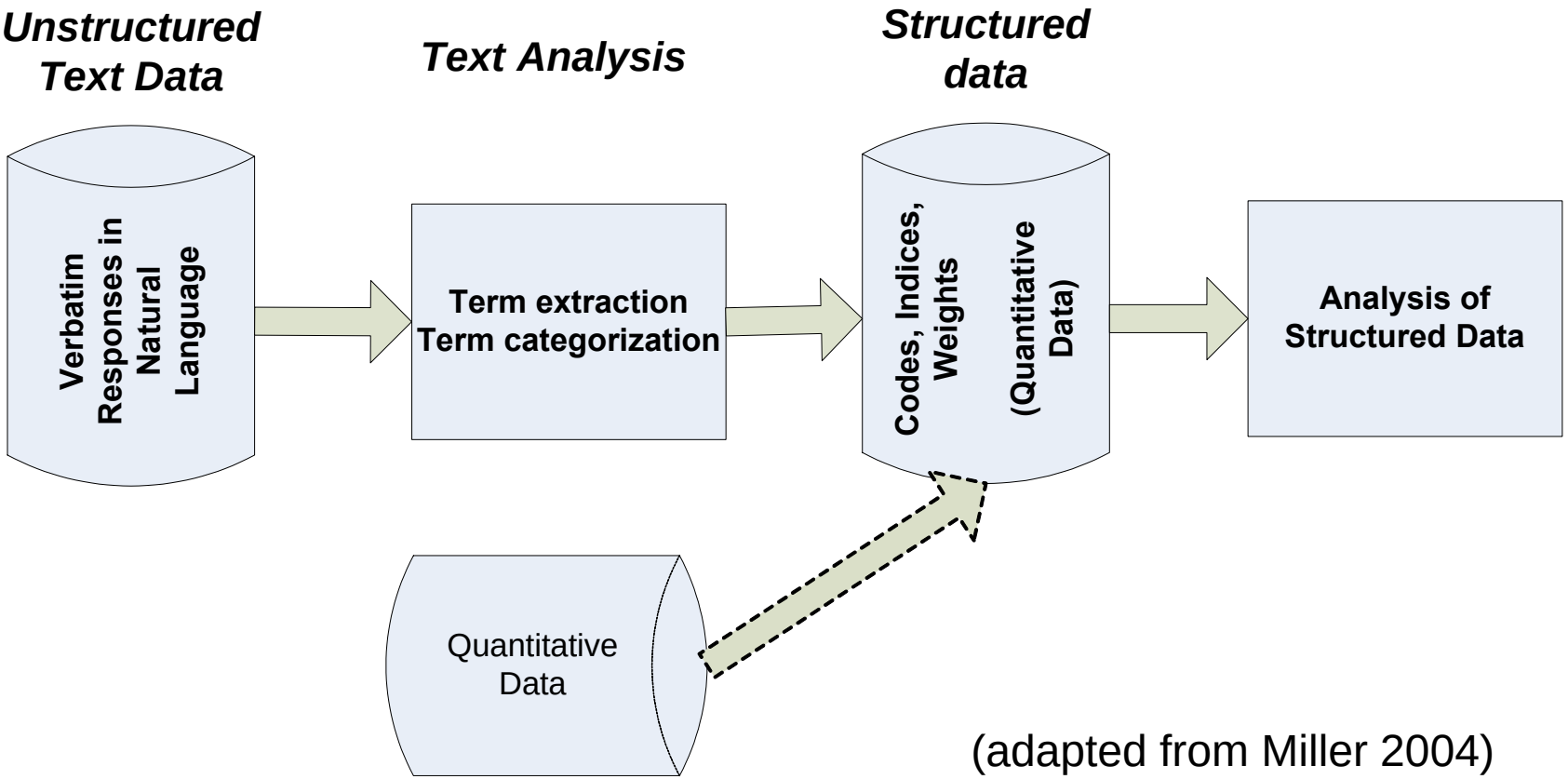
- Preprocessing
  - Sentence and text segmentation
  - Tokenization
  - Stemming/lemmatization
  - Spelling correction
- Term extraction
  - Word indexing
  - Phrase indexing
  - Dictionaries
  - Stop-lists
- Term classification
  - Into semantic classes
- Text classification
- Text clustering
- Entity extraction
  - People, places, company names
  - Times and dates
  - Monetary amounts
  - Measurements
  - Product numbers
  - Postal addresses
  - Email addresses
  - Phone numbers
- Fact extraction
  - “Tom’s number is 415-322-8244”
- Sentiment detection
- Shallow parsing

Text analysis goals



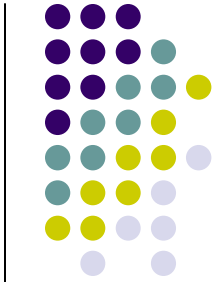
- Automated analytics, trends
- To enable a better user interface for the manual review and analysis of text data
  - Improve speed and quality of manual analysis
  - Allows more data to be considered

Integrated analysis



(adapted from Miller 2004)

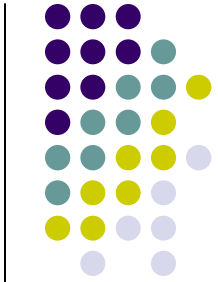
Example: consumer experiences with computers



- Study of PC owners
- Data collected on line over three months in 2003
- 30 quantitative questions
- one open-ended question at the end of the survey:

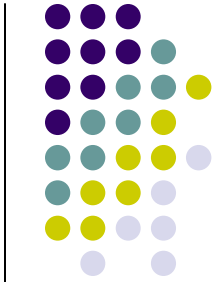
“Please use this space to add any other comments you would like to make about your PC.”

The data



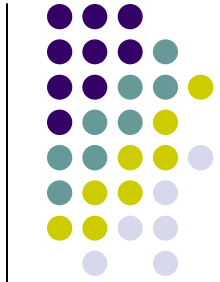
- N=9,086
- Number providing a verbatim response = 4,362 (48%)
- Total number of words = 98,785
- Number of unique words = 6,314
- Maximum number of words = 59

Text analysis process



- Extracted a subset of verbatim responses
- Preprocessed and reviewed text data:
  - Applied a stop-list to remove frequent and generally content-free words
  - Applied lemmatization
  - Reviewed words and phrases, in frequency order
- Identified distinct expressions used for
  - PC component/feature (***hard drive***, ***Windows 2000***)
  - PC event/issue (***crashed***, ***failed***)
  - Positive or negative sentiment (***would not buy***, ***junk***, ***great***)
    - Assigned a weight according to likelihood that the expression predicts positive or negative sentiment
- Assigned expressions to a term class
- Applied term classes to complete data set

Term class for "CD_Component"



	FREQUENCY	#RESPONSES	%RESPONSES
CD	7	7	0.9%
CD BURNER	1	1	0.1%
CD DRIVE	3	3	0.4%
CD WRITER	3	3	0.4%
CDROM	3	2	0.2%
CD-RW	6	6	0.7%
COMBO DRIVE	1	1	0.1%
DVD	13	11	1.4%
DVD WRITER	1	1	0.1%
DVD-ROM	1	1	0.1%

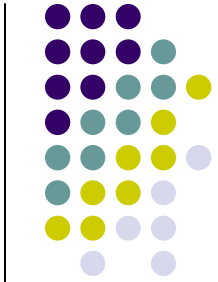
Occurrence of term classes

	FREQUENCY	#RESPONSES	%RESPONSES
MANUF	312	219	27.2%
SENTIMENT_POS	446	125	15.5%
CUSTOMER_SERVICE	146	107	13.3%
PROBLEM_EVENT	122	94	11.7%
PC	66	58	7.2%
HDD	75	51	6.3%
WINOS	52	41	5.1%
EXPANDABILITY	44	37	4.6%
CPU	32	30	3.7%
CD_COMP	41	30	3.7%
INTERACTION	33	29	3.6%
POWER	33	24	3.0%
SOFTWARE	24	22	2.7%
SETUP	26	21	2.6%
RAM	23	18	2.2%
VIDEO	20	15	1.9%
INTERNET	14	14	1.7%
INTERFACES	13	13	1.6%
SENTIMENT_NEG	39	8	1.0%
COMPUTER_BUILD	4	3	0.4%
WARRANTY	2	2	0.2%

Modeling application data

- Sub-sample of consumers who produced verbatim responses, also one or more codes taken to indicate positive or negative sentiment, and who had valid satisfaction rating data (n=669)
- Satisfaction measure
  - 7 point rating scale, very dis- (1) to very sat'd (7)
- Sentiment codes
  - Incidences ranged from 0.1% ("don't really like") to 54% ("Great")

Application data (cont.)



- "Sentiment" codes: (m=17)
  - "Positive" (12)
    - Am Satisfied, Dependable, Excellent, Extremely Satisfied, Good for Work, Great, Like, No Problems, Please, Very Easy, Very Reliable, Very Satisfied
  - "Negative" (5)
    - Don't Like, Don't Really Like, Junk, Not Buy, Would Not Buy

A parametric latent trait model

- I observations on J discrete (binary or ordinal) manifest measures
- A single unobserved continuous dimension, or trait
- I unobserved person parameters on the trait
- Assume:
  - independence of measures within persons (local independence)
  - Independence of measures across persons given person parameters

Model (cont.)

$$p(Y \mid \phi, \lambda) = \prod_{i=1}^{I} \int \prod_{j=1}^{J} p(Y_{ij} \mid \theta_i, \phi_j)\, p(\theta_i \mid \lambda) d\theta_i$$

Y - observed responses

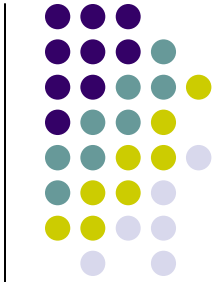
θ - person params

ϕ - variable (item) params (e.g. α_j's and β_j's)

λ - θ distribution params

$$p(\theta, \phi, \lambda \mid Y) \propto p(Y \mid \theta, \phi)\, p(\theta \mid \lambda)\, p(\phi)\, p(\lambda)$$

model (cont.)



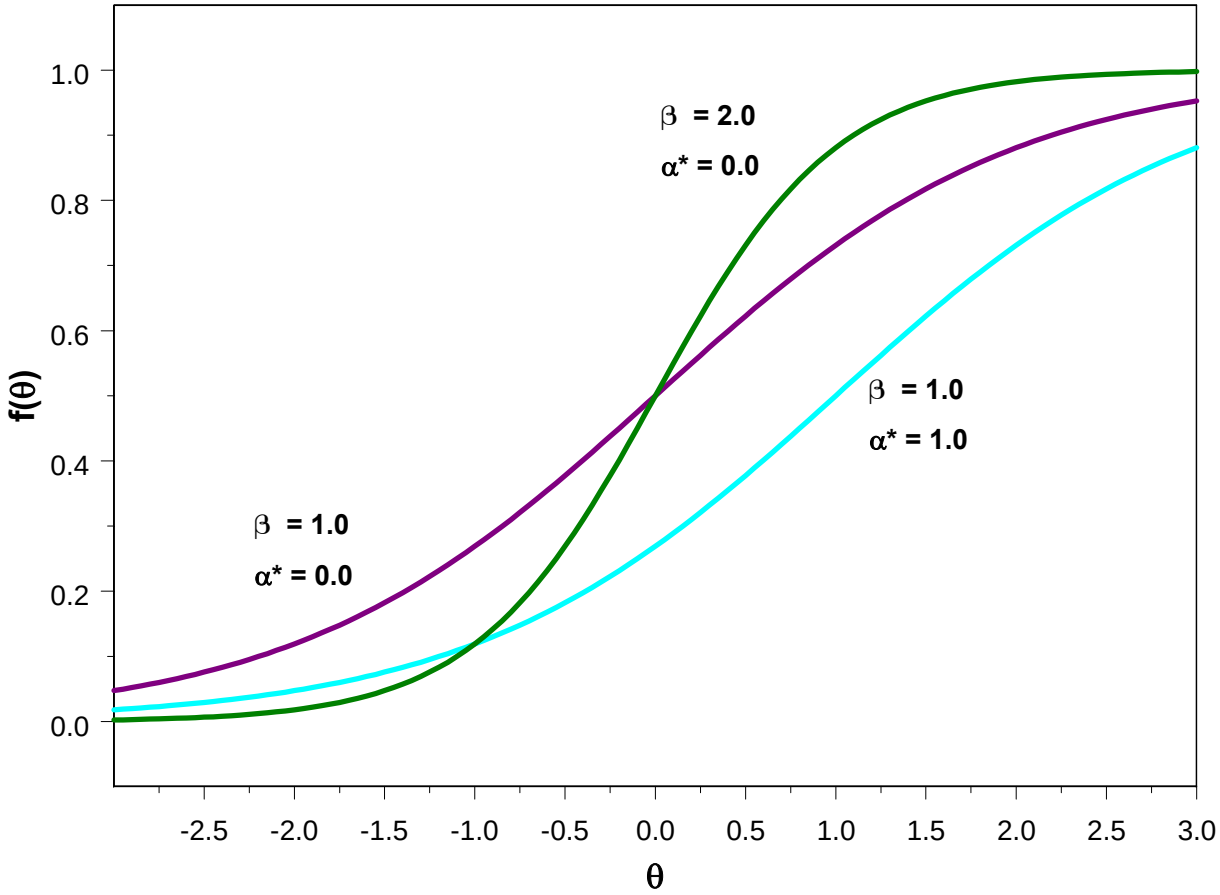
binary item response function:

$$P(Y_{ij}=1|\theta_i,\alpha_j,\beta_j)=(1+\exp(\beta_j\theta_i-a_j))^{-1}$$

multiple category response function:

$$P(Y_{ij}=k|\theta_i,\beta_j,\alpha_{\delta.j})=\frac{\exp\sum_{l=1}^{k}(\beta_j\theta_i-\alpha_{\delta lj})}{\sum_{m=1}^{k}\exp\left(\sum_{l=1}^{m}\beta_j\theta_i-\alpha_{\delta lj}\right)}$$

Example response functions



LTM (cont.)

priors:

$\theta_{\mathrm{i}} \sim N(0,1)$

$\beta_{\mathrm{j}} \sim \log normal(0, 0.1)$

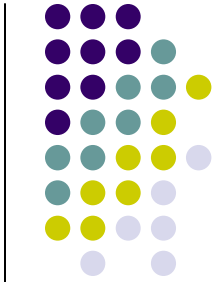
$a_j \sim normal(0, 0.1)$

identification: $\alpha_{(\delta=\delta^*)j} = 0$

location param $\alpha^* = \alpha_j / \beta_j$

for satisfaction scale, all $\beta_{\mathrm{j,\ j=1...6}} = \beta_s$

LTM estimation



- Used logical complements of negative codes
- MCMC
  - M-H w/in Gibbs
- 50,000 iterations, after burn-in of 10,000
- Examined selected chains, compared means
- Sampled every 50

LTM results

Rating Scale:

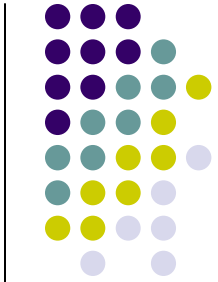
Var	discrim (post. Mean) β	location (post. Mean) α*
sat 12	1.08 (0.28)	-1.95 (0.55)
sat 23	-	-1.70 (0.30)
sat 34	-	-1.50 (0.18)
sat 45	-	-1.28 (0.08)
sat 56	-	-0.88 (0.05)
sat 67	-	*0 (f)*

log-lik	-14,944.20
no. params	732
BIC	34,650.63

Verbatim Codes:

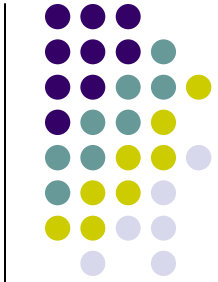
Var	discrim (post. Mean) β	location (post. Mean) α*	
Excellent	2.60 (0.33)	1.84 (0.12)	
Ext. Satisfied	2.02 (0.18)	2.63 (0.16)	
Very Satisfied	2.21 (0.04)	1.59 (0.20)	
Dependable	0.39 (0.02)	1.96 (0.32)	
Very Reliable	0.24 (0.25)	1.02 (0.22)	
No Probs	0.19 (0.20)	2.51 (0.18)	
Great	0.02 (0.05)	-0.11 (0.17)	
Very Easy	0.42 (0.20)	1.66 (0.08)	
Don't Really Like	1.20 (0.22)	-2.45 (0.14)	**rev**
Please	0.04 (0.03)	2.29 (1.80)	
Would Not Buy	1.34 (0.30)	-2.34 (0.28)	**rev**
Good for Work	0.01 (0.01)	0.28 (0.20)	
I Like	0.45 (0.03)	1.84 (0.22)	
Am Satisfied	0.73 (0.30)	2.89 (0.33)	
Not Buy	0.32 (0.10)	-3.39 (0.50)	**rev**
Don't Like	0.31 (0.08)	-3.22 (0.21)	**rev**
Junk	2.78 (0.30)	-3.45 (0.13)	**rev**

LTM issues and extensions



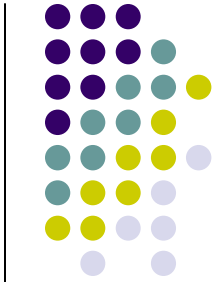
- Informative priors
- Multiple traits
- Missing data
- Covariates
- Continuous response variables
- Scalability/computational intensity

Other doings



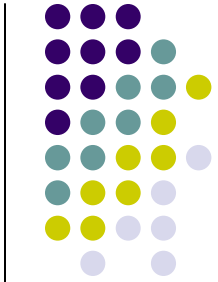
- Comparing to other methods
- Estimating information content
- Converting clustering problems into prediction problems
- Conditioning

Some take-aways



- Fully automatic processing is not currently possible
- Extent to which automation is possible depends on features of the data and context
- “Featurization” (i.e. coding) is critical step in all analysis
- Coded text data can be analyzed in combination with quant data using a range of advanced techniques

Resources



BOOKS

Jurafsky, D. & Martin, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.** Prentice Hall, 2000.

Manning, C.D. & Shultz, H. **Foundations of Statistical Natural Language processing.** Cambridge MA: MIT Press, 1999.

Mertz, D. **Text Processing in Python.** Boston, Addison-Wesley, 2003.

Miller, T.W. **Data and Text Mining: A Management Introduction.** Prentice-Hall, 2004

Sullivan, D. **Document Warehousing and Text Mining.** Wiley, 2001.

NLP TOOLS

fnTBL

Open source tools for tasks such as part-of-speech tagging and shallow parsing, using transformation-based learning.
nlp.cs.jhu.edu/~rflorian/fntbl/

WordNet

www.cogsci.princeton.edu/~wn/

TEXT MINING TOOLS

WordStat, Provalis Research

Tool for content analysis and text mining, for use with tools for quantitative data analysis.
www.simstat.com

TextAnalyst and PolyAnalyst, Megaputer Intelligence

Combined tool for data and text mining.
www.megaputer.com

Intelligent Miner for Text, IBM

Powerful suite of text mining and text processing functions.
www.ibm.com/software/data/iminer

WordSmith

Simple Windows software for concordances and other text analysis.
www.lexically.net/wordsmith/

SPSS Clementine w/ Lexiquest

www.spss.com

Insightful Corp. Insightful Miner

www.insightful.com

For more resources, go to:
www.atomicintelligence.com/ai/resources.html