From Pattern Matching to Knowledge Discovery Using ***Text Mining and Visualization*** Techniques

Seth Grimes

Alta Plana Corporation

@sethgrimes – 301-270-0795 -- *http://altaplana.com*

Special Libraries Association 2010

June 13, 2010

Introduction

Seth Grimes –

Principal Consultant, Alta Plana Corporation.

Contributing Editor, *IntelligentEnterprise.com*.

Channel Expert, *BeyeNETWORK.com*.

Instructor, The Data Warehousing Institute, *tdwi.org*.

Founding Chair, Sentiment Analysis Symposium.

Founding Chair, Text Analytics Summit.

Perspectives

Perspective #1: You support research or work in IT.

You help end users who have lots of text.

Perspective #2: You're a researcher, business analyst, or other "end user."

You have lots of text. You want an automated way to deal with it.

Perspective #3: You work for a solution provider.

Perspective #4: Other?

Perspective A: Your focus is Information Retrieval.

Perspective B: Your focus is Data Analysis.

Agenda

1. The "Unstructured" Data Challenge.
2. Text analytics for information retrieval and BI.
3. Text analysis technologies and processes.
4. Applications.
5. Software and tools.
6. Text visualization for exploratory analysis.

Note:

I will not cover the agenda in a linear fashion. Text mining and viz are intermixed.

Class coverage is for both information analysts and end users.

Text analytics ≈ text mining ≈ text data mining.

Value in Text

It's a truism that 80% of enterprise-relevant information originates in "unstructured" form:

- E-mail and messages.
- Web pages, news & blog articles, forum postings, and other social media.
- Contact-center notes and transcripts.
- Surveys, feedback forms, warranty claims.
- Scientific literature, books, legal documents.
- ...

Non-text "unstructured" forms?

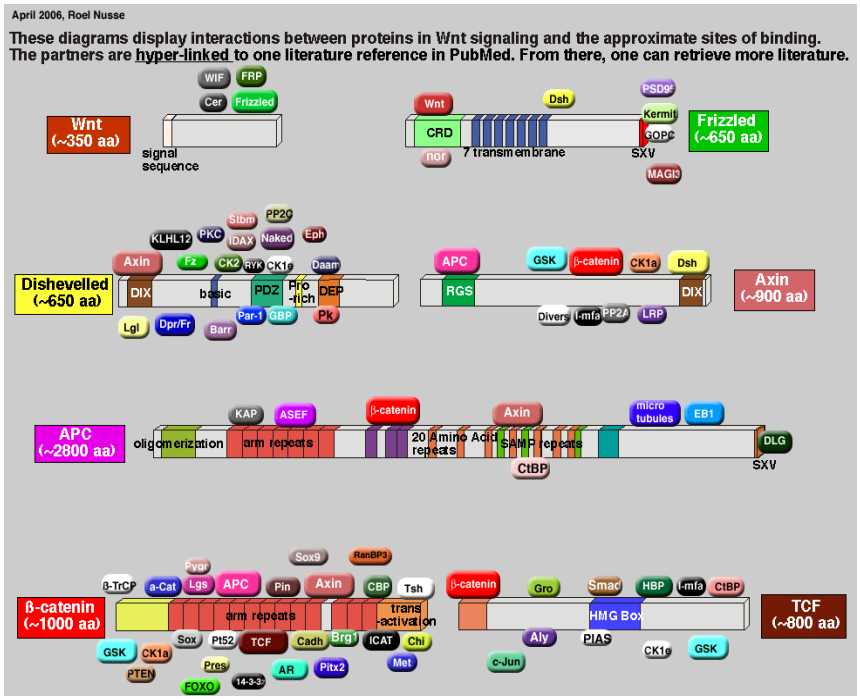
http://upload.wikimedia.org/wikipedia/commons/thumb/9/90/LOC_Brooklyn_Bridge_and_East_River_3.png/753px-LOC_Brooklyn_Bridge_and_East_River_3.png

Unstructured Sources

These sources may contain "traditional" data.

The Dow fell 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index fell 1.44, or 0.11 percent, to 1,263.85.

And they may not.



Axin and Frat1 interact with dvl and GSK, bridging Dvl to GSK in Wnt-mediated regulation of LEF-1. Wnt proteins transduce their signals through dishevelled (Dvl) proteins to inhibit glycogen synthase kinase 3beta (GSK), leading to the accumulation of cytosolic beta-catenin and activation of TCF/LEF-1 transcription factors. To understand the mechanism by which Dvl acts through GSK to regulate LEF-1, we investigated the roles of Axin and Frat1 in Wnt-mediated activation of LEF-1 in mammalian cells. We found that Dvl interacts with Axin and with Frat1, both of which interact with GSK. Similarly, the Frat1 homolog GBP binds Xenopus Dishevelled in an interaction that requires GSK.

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Retrieve&list_uids=10428961&dopt=Abstract

www.stanford.edu/%7ernusse/wntwindow.html

Unstructured Sources

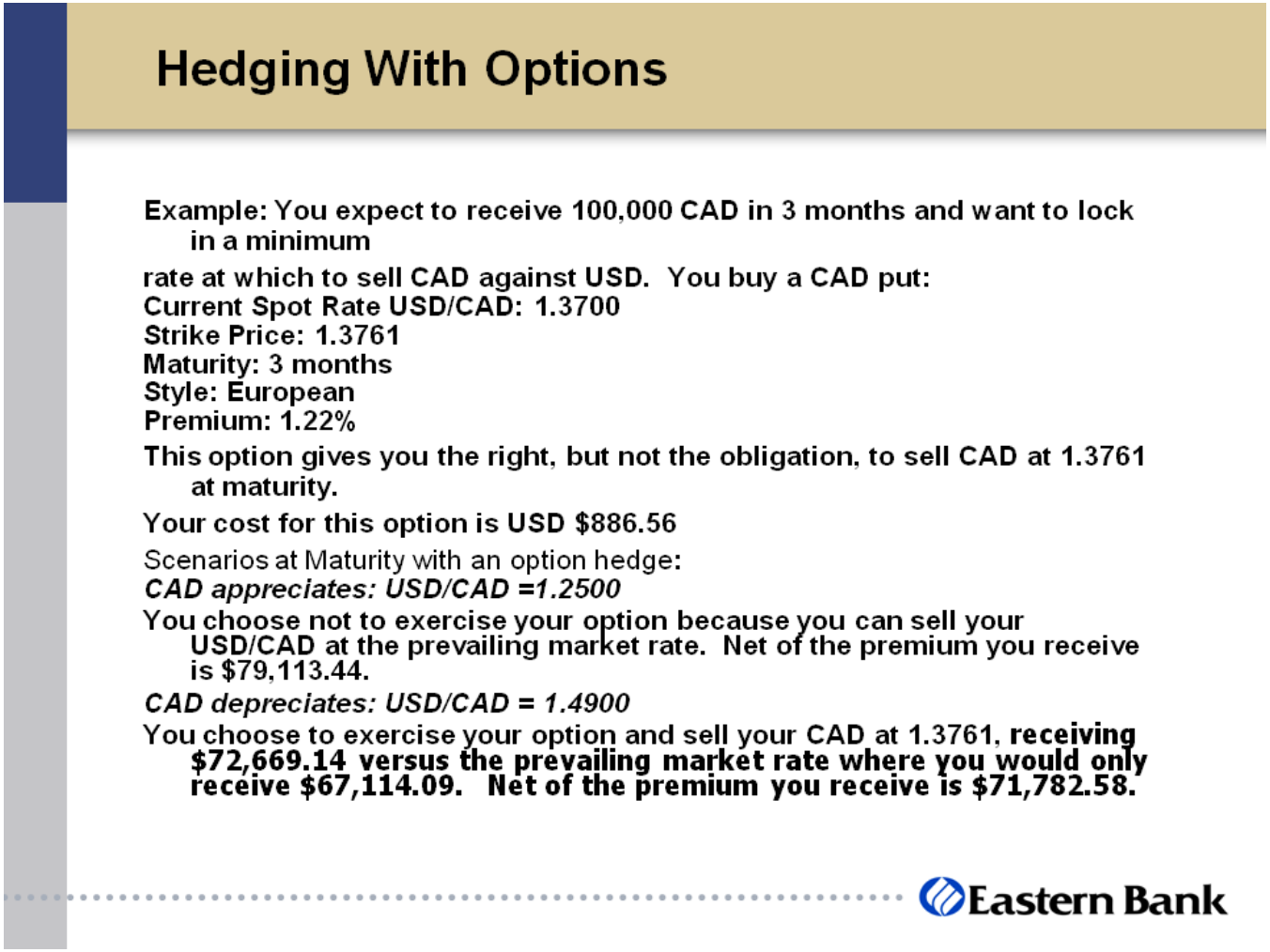
Sources may mix fact and sentiment –

> *When you walk in the foyer of the hotel it seems quite inviting but the room was very basis and smelt very badly of stale cigarette smoke, it would have been nice to be asked if we wanted a non smoking room, I know the room was very cheap but I found this very off putting to have to sleep with the smell, and it was to cold to leave the window open.*
>
> *Overall I would never sell/buy a Motorola V3 unless it is demanded. My life would be way better without this phone being around (I am being 100% serious) Motorola should pay me directly for all the problems I have had with these phones. :-(*

– and contain information that isn't either…

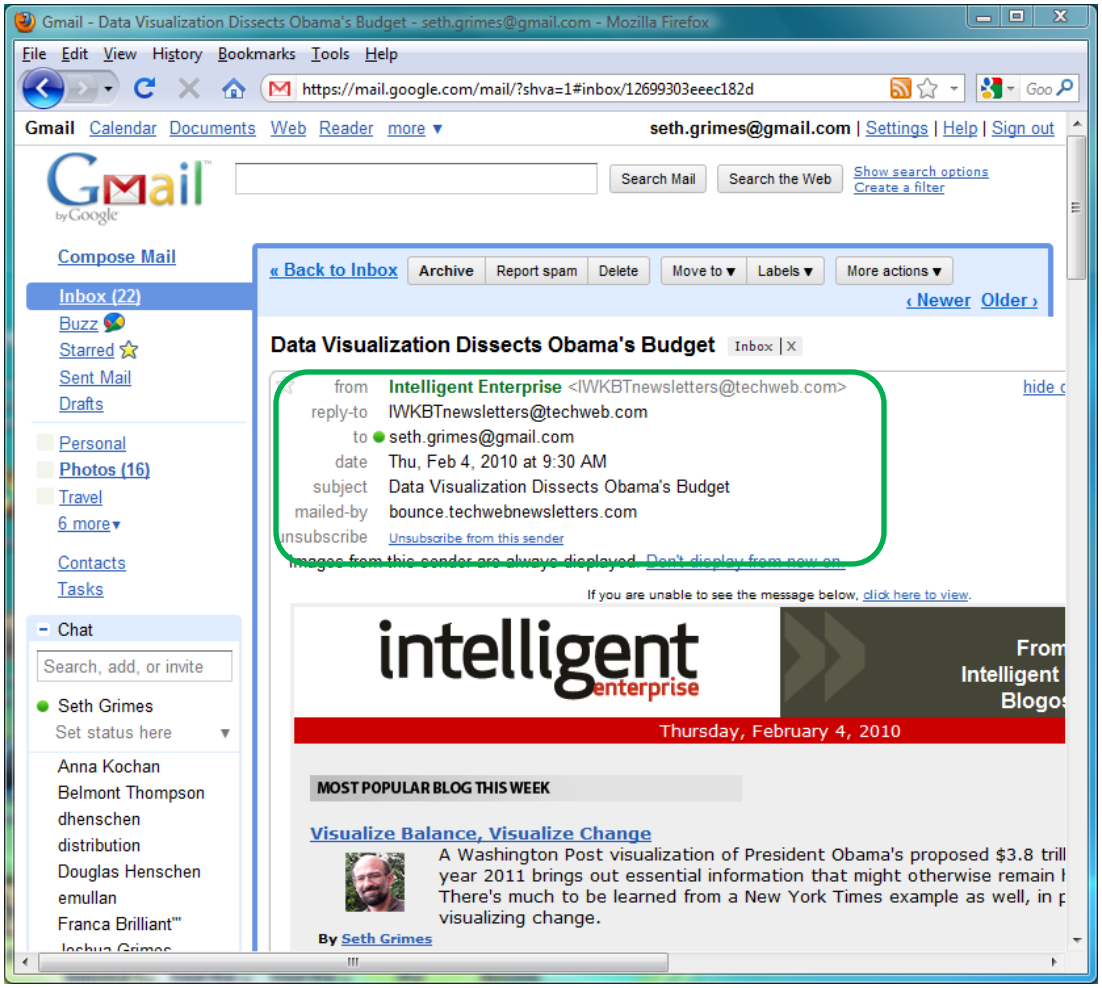
Unstructured Sources

Neither fact nor fiction, but definitely narrative:



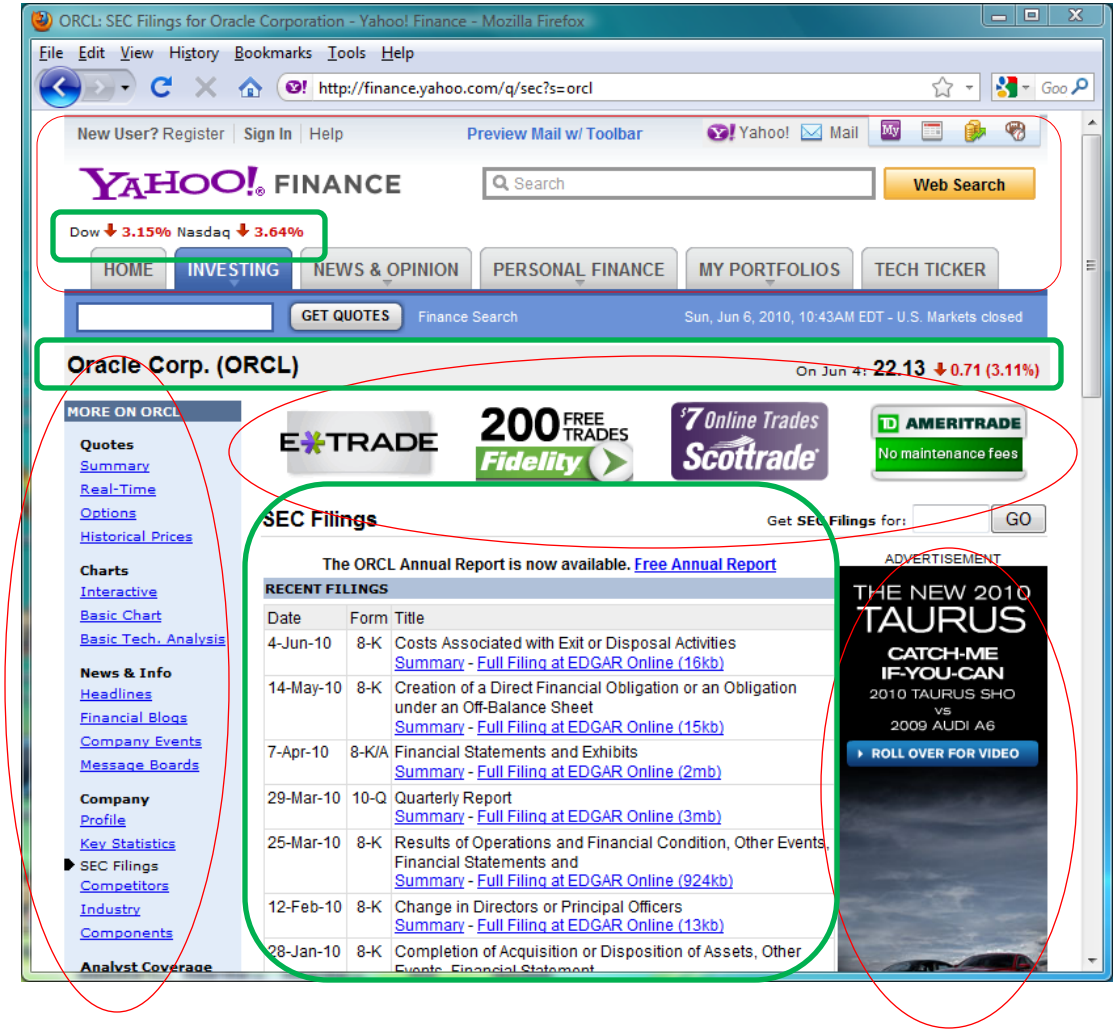
Unstructured Sources

Sources contain/provide metadata:



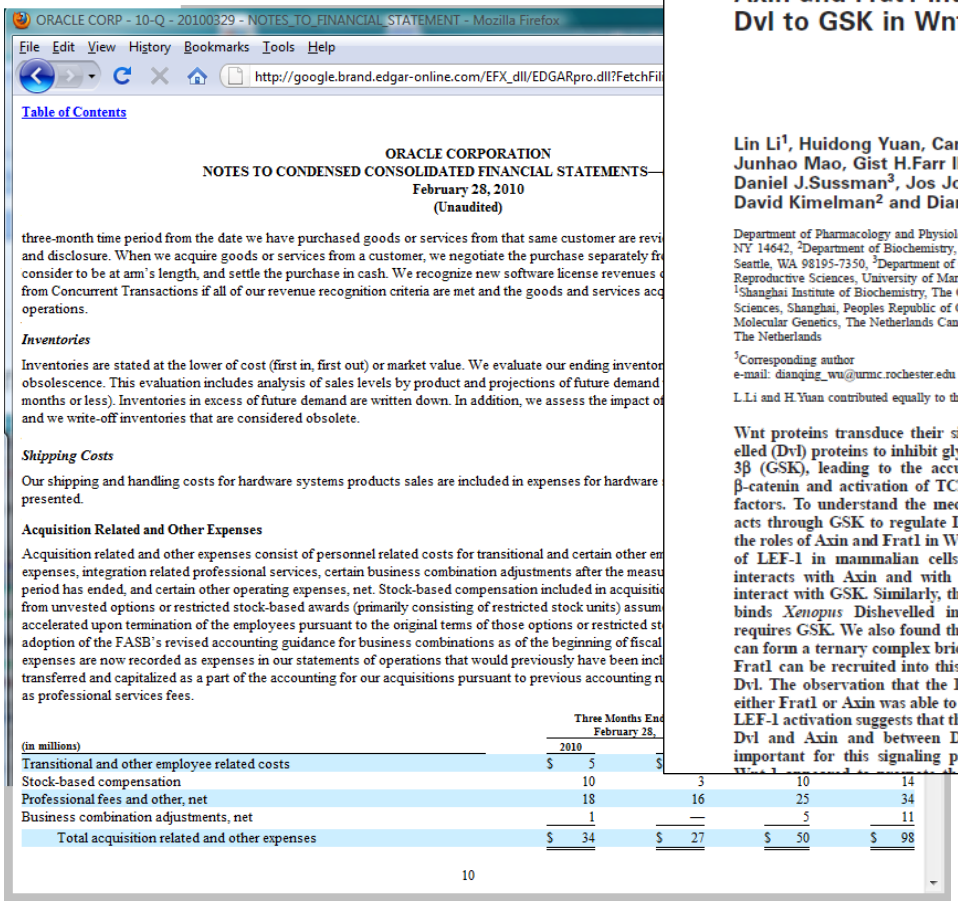
Unstructured Sources

Sources may intermix content and "noise":



Unstructured Sources

"Unstructured" materials likely contain structure:



The EMBO Journal Vol.18 No.15 pp.4233–4240, 1999

Axin and Frat1 interact with Dvl and GSK, bridging Dvl to GSK in Wnt-mediated regulation of LEF-1

Lin Li[1], Huidong Yuan, Carole D.Weaver[2], Junhao Mao, Gist H.Farr III[2], Daniel J.Sussman[3], Jos Jonkers[4], David Kimelman[2] and Dianqing Wu[5]

Department of Pharmacology and Physiology, University of Rochester, NY 14642, [2]Department of Biochemistry, University of Washington, Seattle, WA 98195-7350, [3]Department of Obstetrics, Gynecology and Reproductive Sciences, University of Maryland, Baltimore, MD, USA, [4]Shanghai Institute of Biochemistry, The Chinese Academy of Sciences, Shanghai, Peoples Republic of China and [5]Division of Molecular Genetics, The Netherlands Cancer Institute, Amersterdam, The Netherlands

[1]Corresponding author
e-mail: dianqing_wu@urmc.rochester.edu

L.Li and H.Yuan contributed equally to this work

Wnt proteins transduce their signals through dishevelled (Dvl) proteins to inhibit glycogen synthase kinase 3β (GSK), leading to the accumulation of cytosolic β-catenin and activation of TCF/LEF-1 transcription factors. To understand the mechanism by which Dvl acts through GSK to regulate LEF-1, we investigated the roles of Axin and Frat1 in Wnt-mediated activation of LEF-1 in mammalian cells. We found that Dvl interacts with Axin and with Frat1, both of which interact with GSK. Similarly, the Frat1 homolog GBP binds *Xenopus* Dishevelled in an interaction that requires GSK. We also found that Dvl, Axin and GSK can form a ternary complex bridged by Axin, and that Frat1 can be recruited into this complex probably by Dvl. The observation that the Dvl-binding domain of either Frat1 or Axin was able to inhibit Wnt-1-induced LEF-1 activation suggests that the interactions between Dvl and Axin and between Dvl and Frat may be important for this signaling pathway. Furthermore,

Pangolin (Pan) in *Wingless* (Wg) signaling. The genetic order of these signal transducers has been established, in which Wg acts through Dsh to inhibit Zw-3, thus relieving the repression of Arm by Zw-3, resulting in the up-regulation of Arm (Dickinson and McMahon, 1992; Nusse and Varmus, 1992; Klingensmith and Nusse, 1994; Perrimon, 1994; Cadigan and Nusse, 1997; Dale, 1998). The Wnt signaling pathway appears to be largely conserved in mammals. In addition to the existence of a large number of Wg homologs, there are mammalian homologs for Dsh, Zw-3, Arm and Pan.

Molecular cloning has revealed several mammalian Dsh homologs (Dvl), including three from the mouse (Sussman *et al.*, 1994; Klingensmith *et al.*, 1996; Tsang *et al.*, 1996). Amino acid sequence comparison of all known Dsh/Dvl molecules across species revealed several highly conserved regions. Most notable is the one located in the central part of the molecule referred to as the disc-large homology region or PDZ domain, which was found in a number of proteins including PSD-95, ZO-1 and Discs-large (Ponting *et al.*, 1997). Studies have shown that the PDZ domain in PSD-95 can bind to a C-terminal motif of four amino acids (X-Thr/Ser-X-Val) (Doyle *et al.*, 1996). However, ligands for the PDZ domain of Dsh/Dvl remain unknown. At the C-terminal side of the PDZ domain is located a DEP (dishevelled, egl-10 and pleckstrin) domain. Similar DEP motifs are also found in a number of other proteins (Ponting and Bork, 1996). The N-terminal conserved domain shares homology with a newly identified protein, Axin, which was shown to antagonize Wnt signaling (Zeng *et al.*, 1997; Behrens *et al.*, 1998; Ikeda *et al.*, 1998; Sakanaka *et al.*, 1998; Hamada *et al.*, 1999). This N-terminal domain is referred to as the DIX (dishevelled and axin) domain. Dvl as well as its *Drosophila* homolog Dsh recently have been shown to regulate two independent

The "Unstructured" Data Challenge

The task:

1. Find the right information.
2. Use inherent structure and latent semantics to –
   Infer meaning.
   Discern information content.
   Structure content for machine use.
3. Apply analytical methods to generate insight.
4. Present findings.

From the Analytics/Business Perspective

1. If you are not analyzing text – if you're analyzing only transactional information – you're missing opportunity or incurring risk.
2. Text analytics can boost business results –

"Organizations embracing text analytics all report having an epiphany moment when they suddenly knew more than before."

-- Philip Russom, the Data Warehousing Institute

– via established BI / data-mining programs, or independently.

From the Analytics/Business Perspective

Some folks may need to expand their views of what BI and business analytics are about.

Others can do text analytics without worrying about BI or data mining.

Let's deal with text-BI first...

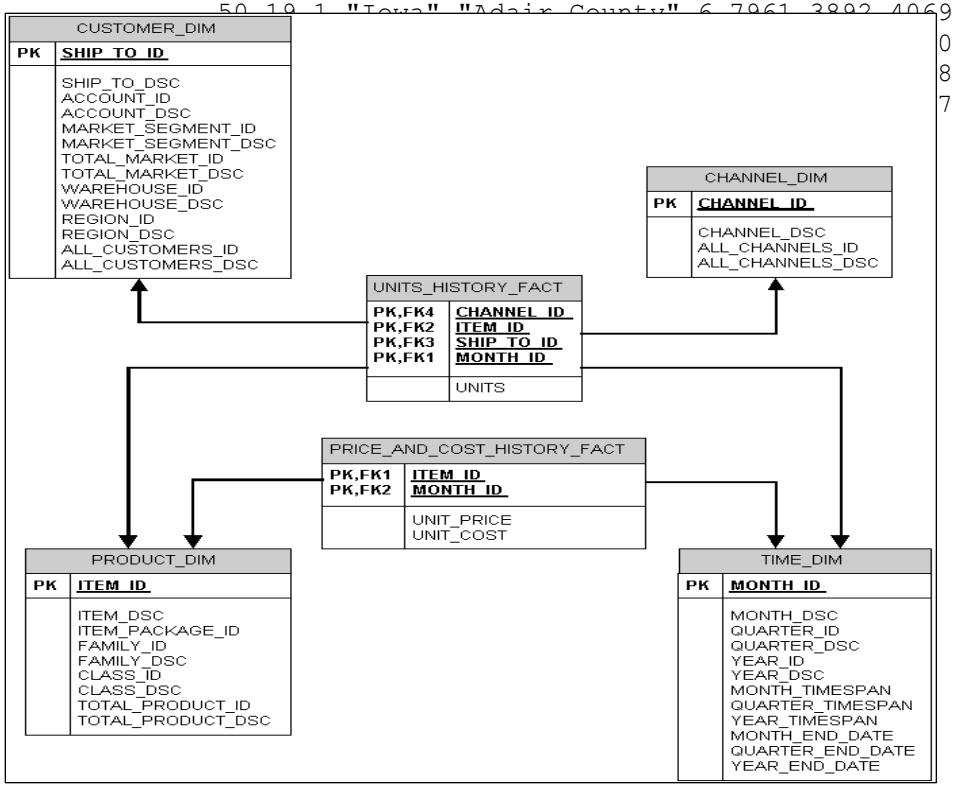
> *"The bulk of information value is perceived as coming from data in relational tables. The reason is that data that is structured is easy to mine and analyze."*
>
> -- Prabhakar Raghavan, Yahoo Research

Business Intelligence

Conventional BI feeds off:

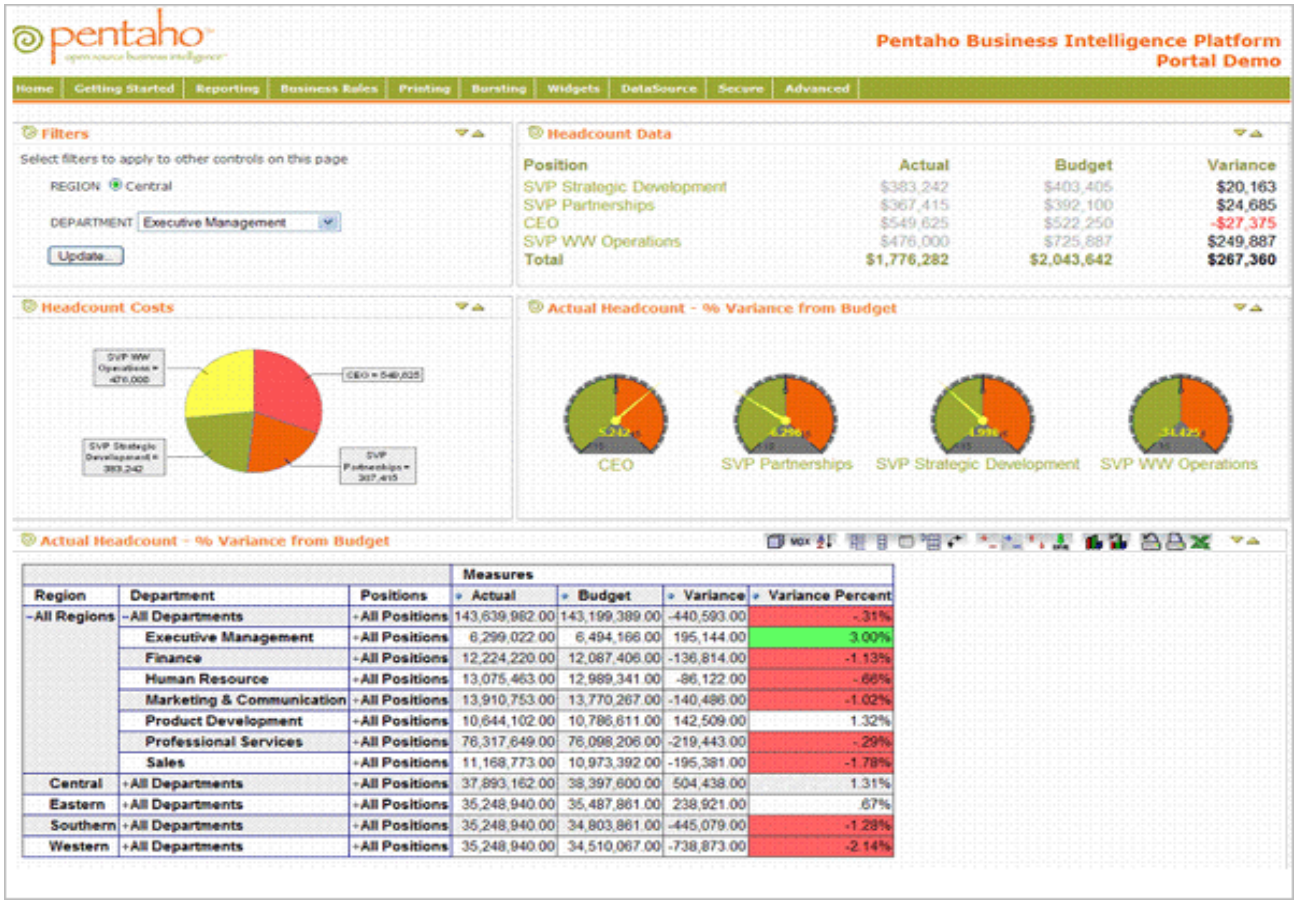
It runs off:

```
"SUMLEV","STATE","COUNTY","STNAME","CTYNAME","YEAR","POPESTIMATE",
50,19,1,"Iowa","Adair County",1,8243,4036,4207,446,225,221,994,509
50,19,1,"Iowa","Adair County",2,8243,4036,4207,446,225,221,994,509
50,19,1,"Iowa","Adair County",3,8212,4020,4192,442,222,220,987,505
50,19,1,"Iowa","Adair County",4,8095,3967,4128,432,208,224,935,488
50,19,1,"Iowa","Adair County",5,8003,3924,4079,405,186,219,928,495
50,19,1,"Iowa","Adair County",6,7961,3892,4069,384,183,201,907,472
                                                  0,366,179,187,871,454
                                                  8,343,162,181,841,439
                                                  7,338,159,179,805,417
```



Business Intelligence

Conventional BI produces:



Text-BI: Back to the Future

Note that business intelligence (BI) was first defined in 1958:

> *"In this paper, **business is a collection of activities** carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera... The **notion of intelligence** is also defined here... as 'the ability to apprehend the **interrelationships of presented facts** in such a way as to **guide action towards a desired goal**.'"*
>
> -- Hans Peter Luhn
>
> "A Business Intelligence System"
>
> IBM Journal, October 1958

What was IT like in the '50s?

Document input and processing

Knowledge handling is key

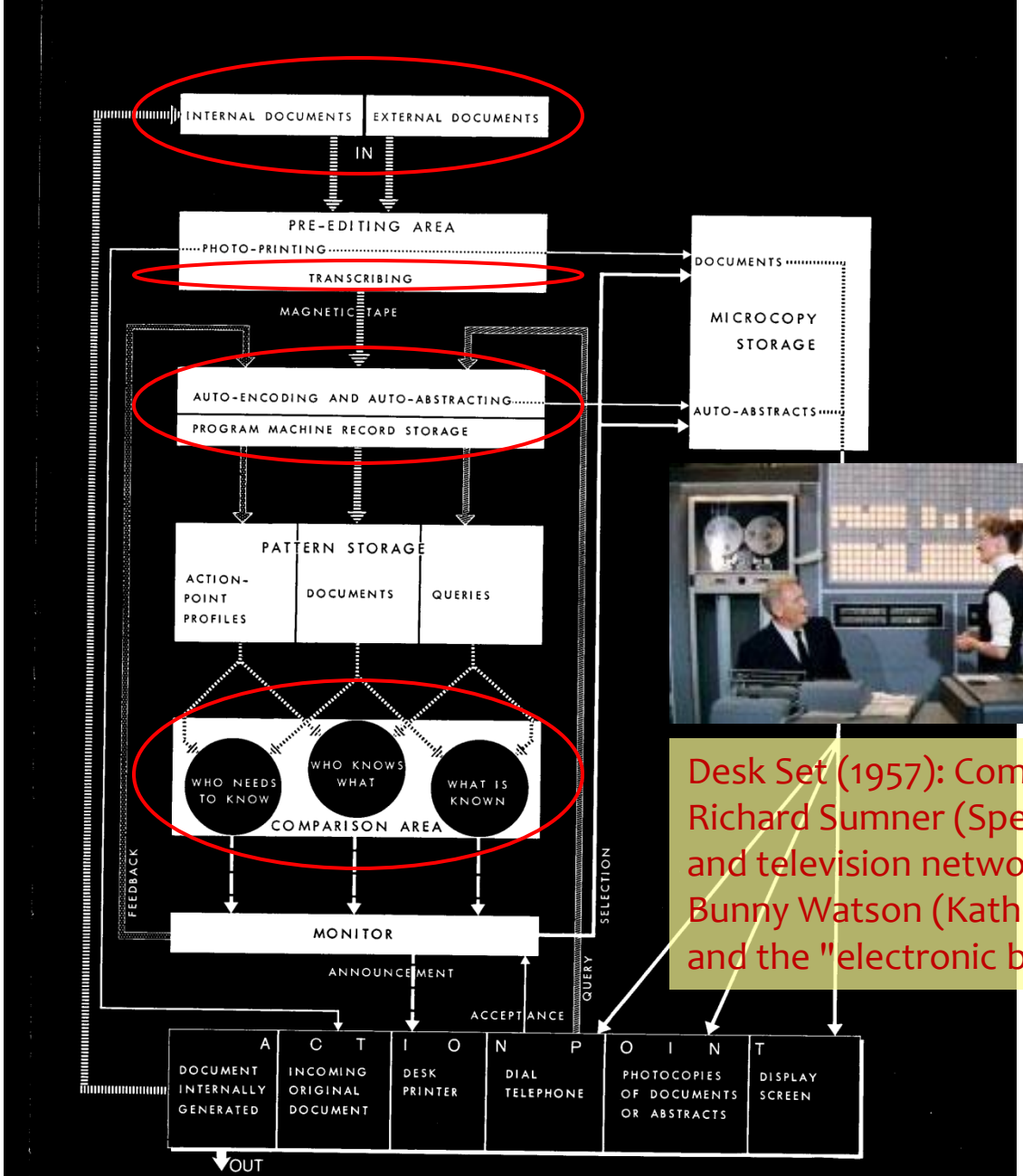


Figure 1 **A Business Intelligence System**

Desk Set (1957): Computer engineer Richard Sumner (Spencer Tracy) and television network librarian Bunny Watson (Katherine Hepburn) and the "electronic brain" EMERAC.

From the Information Retrieval Perspective

What do people do with electronic documents?

1. Publish, Manage, and Archive.
2. Index and Search.
3. Categorize and Classify according to metadata & contents.
4. Information Extraction.

For textual documents, text analytics enhances #1 & #2 and enables #3 & #4.

You need linguistics to do #1 & #4 well, to deal with meaning (a.k.a. *semantics*).

Search is not enough...

From the Information Retrieval Perspective

Keyword search is just a start, the dumbest form of pattern matching.

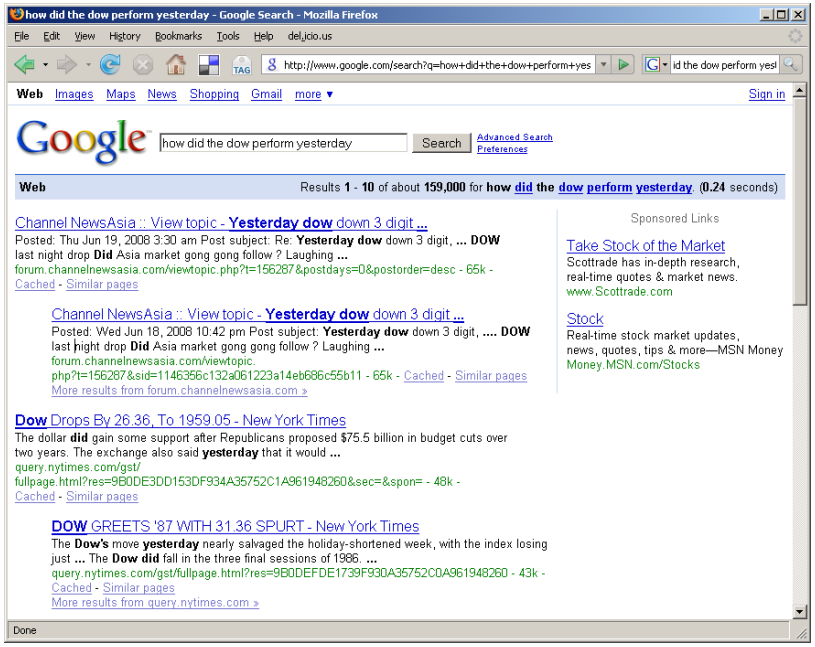
*It doesn't help you **discover** things you're unaware of.*

*Results often lack **relevance**.*

*Basic search finds documents, not **knowledge**.*

Articles from a forum site

Articles from 1987



Semantics

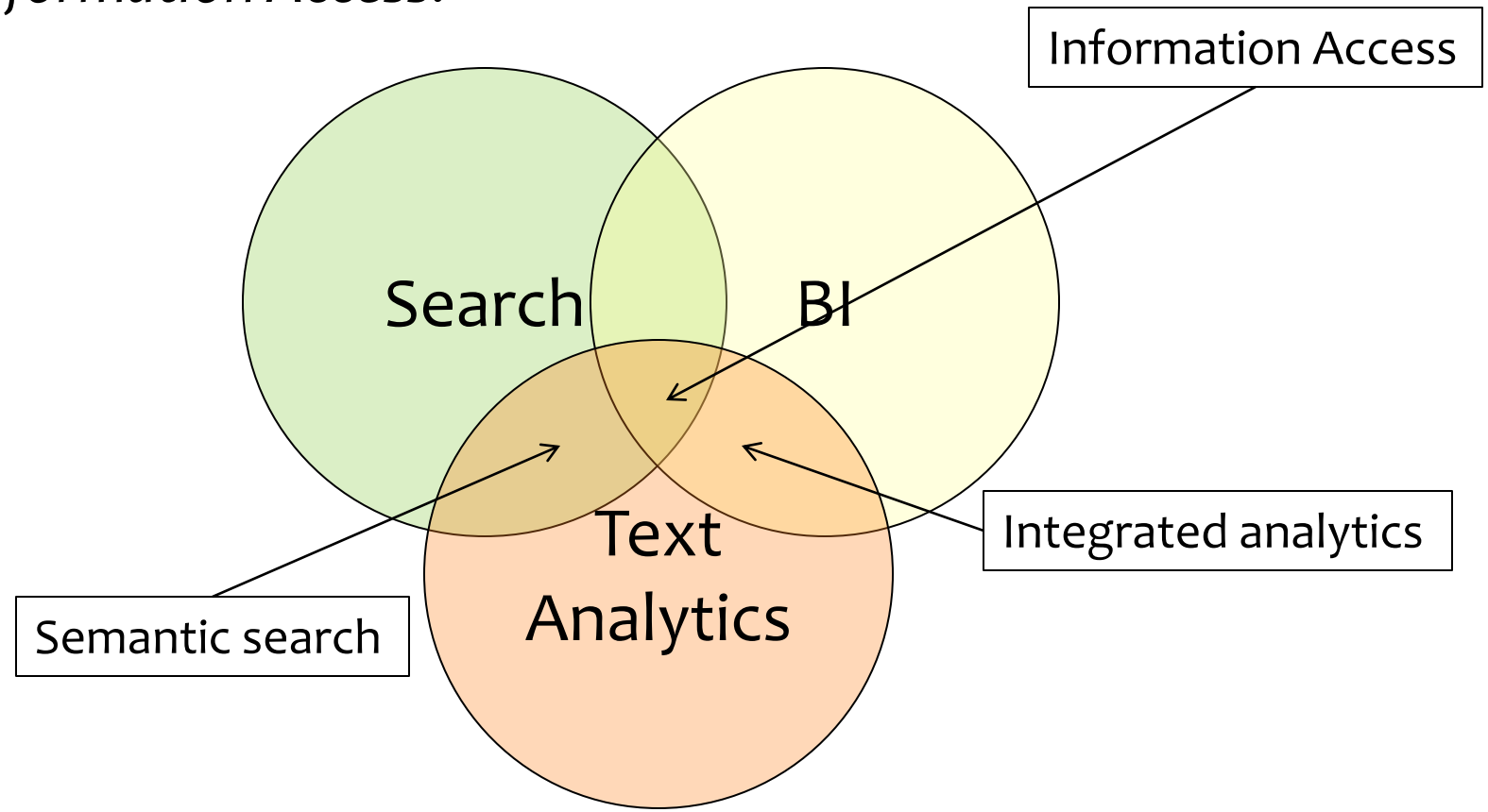
Text analytics adds ***semantic*** understanding of –

- Named entities: people, companies, places, etc.
- Pattern-based entities: e-mail addresses, phone numbers, etc.
- Concepts: abstractions of entities.
- Facts and relationships.
- Concrete and abstract attributes (e.g., 10-year, expensive, comfortable).
- Subjectivity in the forms of opinions, sentiments, and emotions: attitudinal data.

Call these elements, collectively, *features*.

Semantics, Analytics, and IR

In a sense, text analytics, by generating semantics, bridges search and BI to turn *Information Retrieval* into *Information Access*.



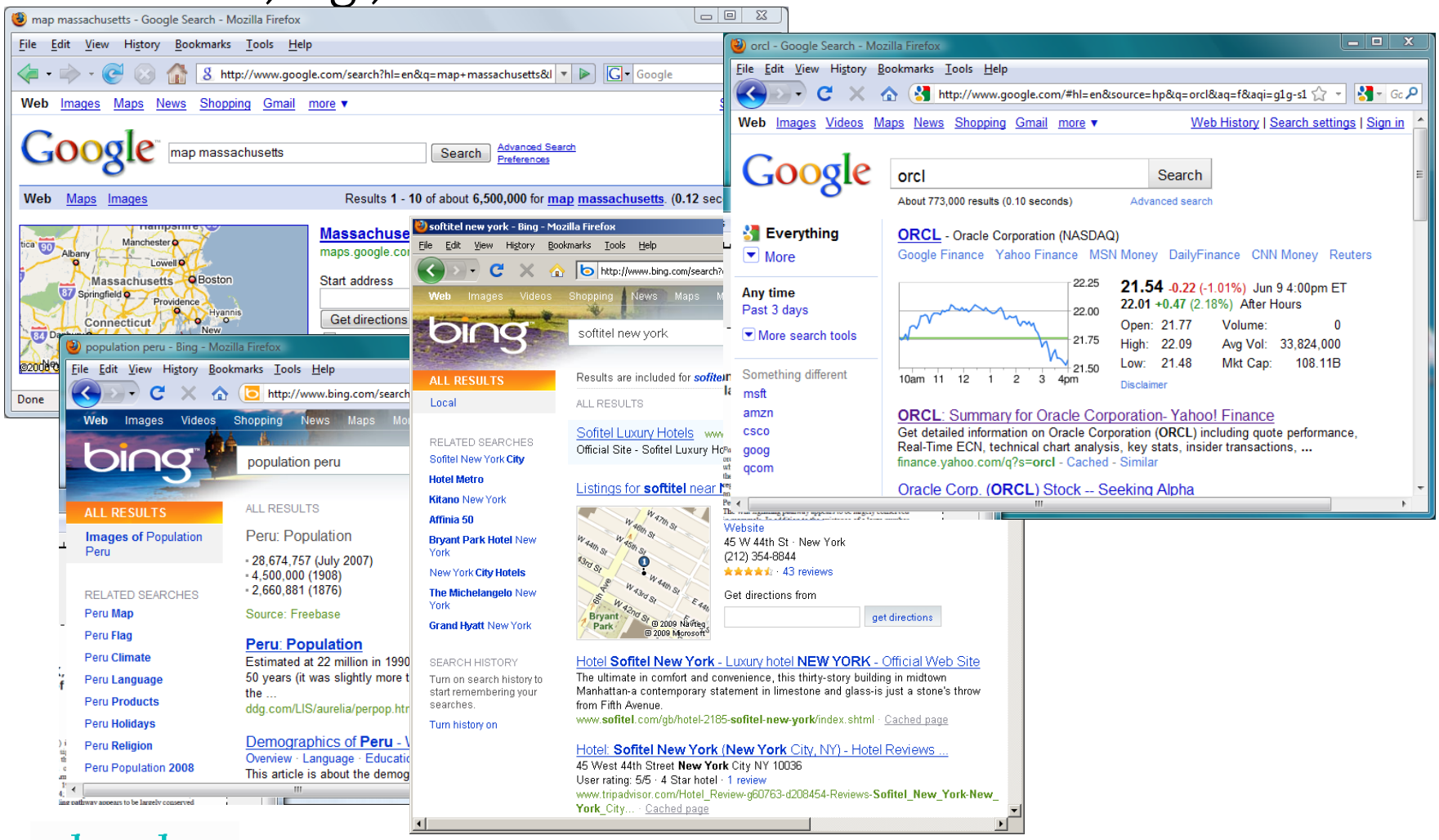
Information Access

Text analytics transforms Information Retrieval (IR) into Information Access (IA).

- Search terms become queries.
- Retrieved material is mined for larger-scale structure.
- Retrieved material is mined for features such as entities and topics or themes.
- Retrieved material is mined for smaller-scale structure such as facts and relationships.
- Results are presented intelligently, for instance, grouping on mined topics-themes.
- Extracted information may be visualized and explored.

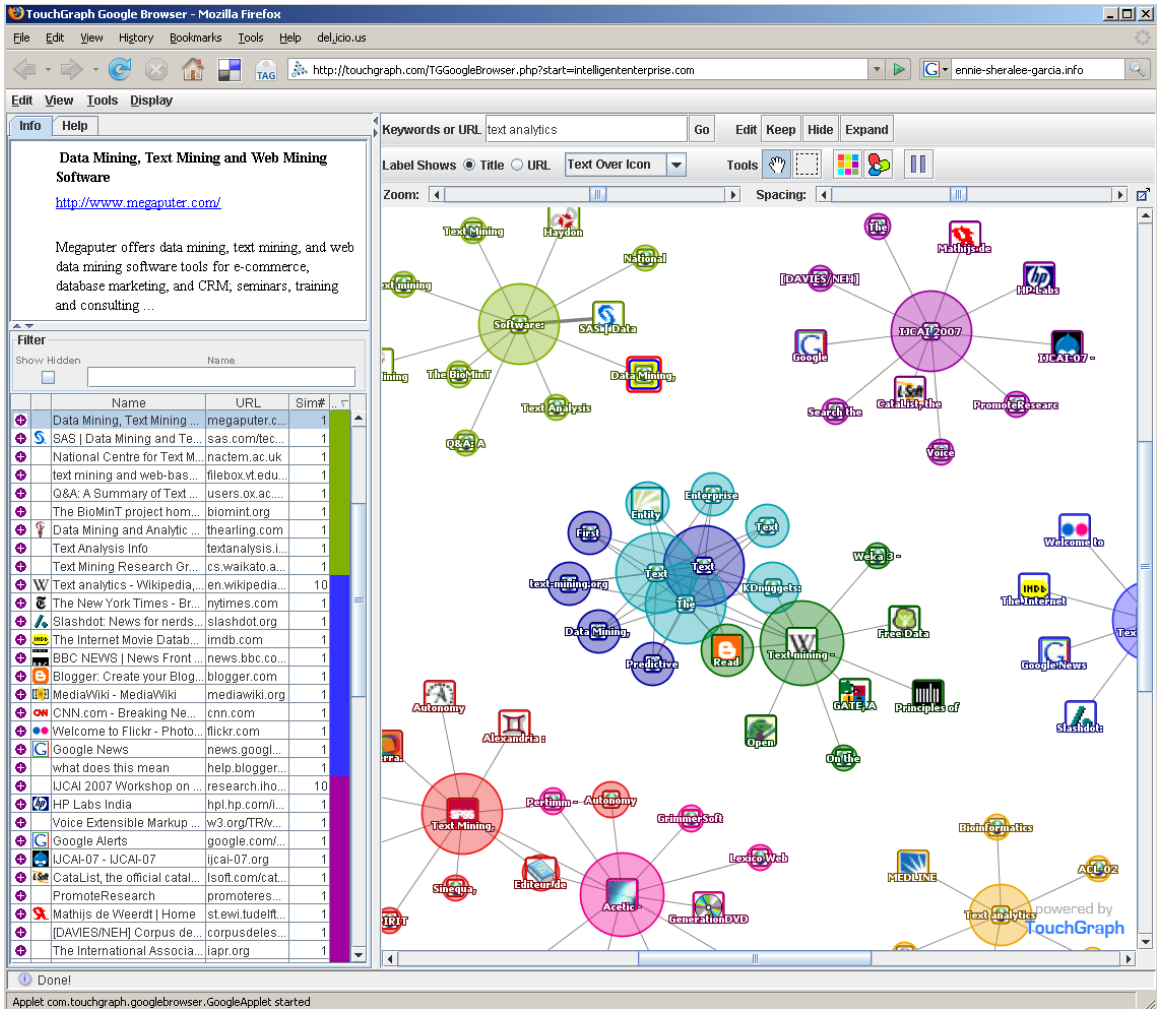
Information Access

Text analytics enables results that suit the information and the user, e.g., answers –



Intelligent Results Presentation

E.g., results clustering on extracted topics:



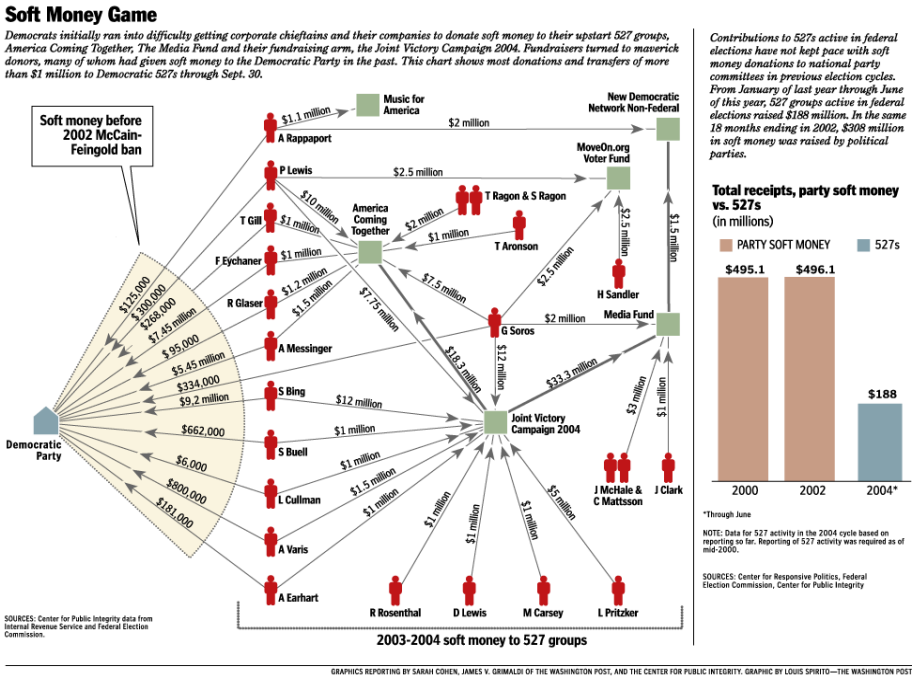
http://touchgraph.com/TGGoogleBrowser.php?start=text%20analytics

Text Data Mining

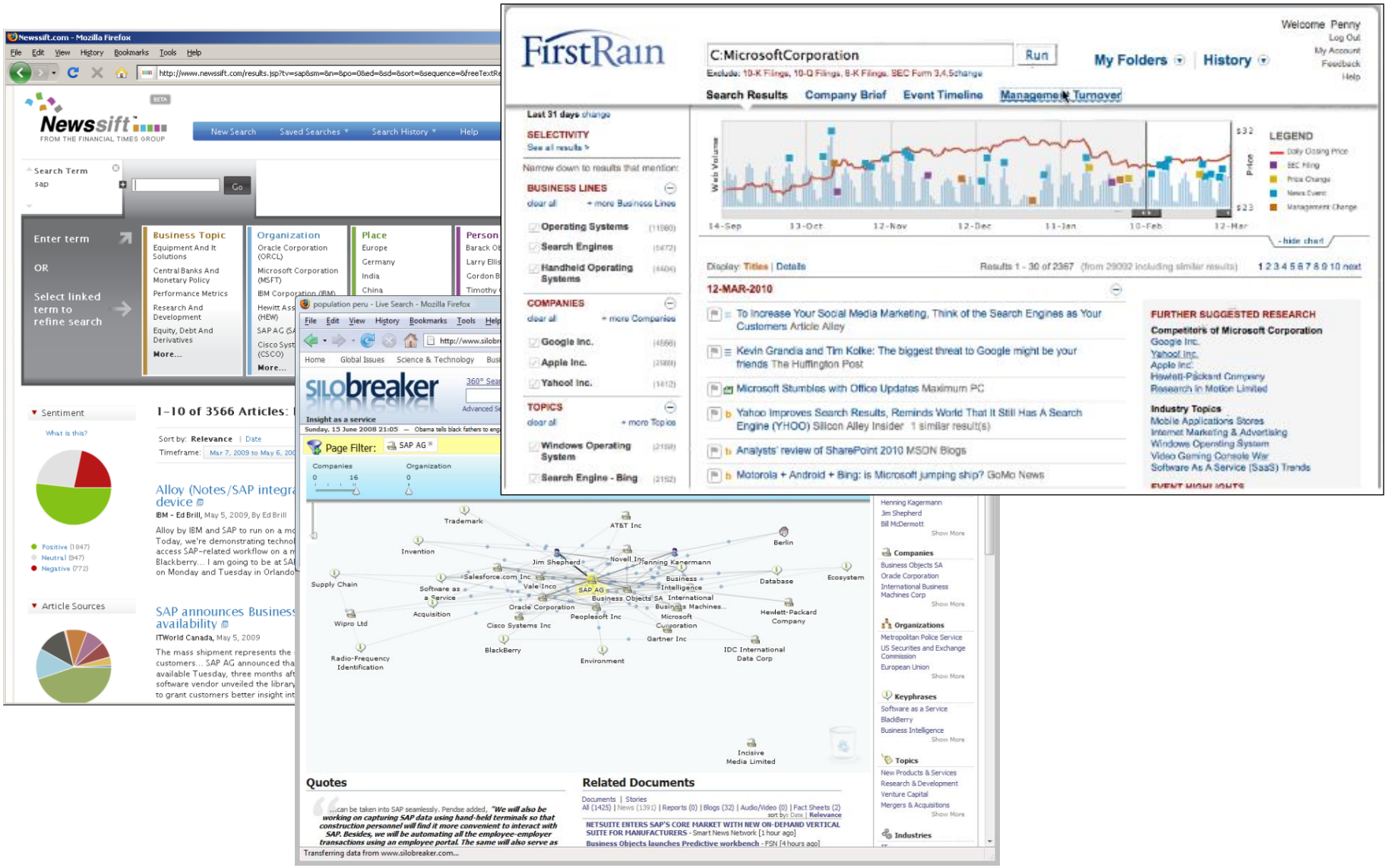
Data Mining = Knowledge Discovery in Data.

Text Mining = Data Mining of textual sources.

- Clustering and Classification.
- Link Analysis.
- Association Rules.
- Predictive Modelling.
- ~~Regression~~.
- ~~Forecasting~~.



Text Data Mining Enables Content Exploration



Text Analytics Definition

Text analytics automates what researchers, writers, scholars, and all the rest of us have been doing for years. Text analytics –

- Applies linguistic and/or statistical techniques to **extract concepts and patterns** that can be applied to categorize and classify documents, audio, video, images.
- **Transforms "unstructured" information into data** for application of traditional analysis techniques.
- Discerns **meaning and relationships** in large volumes of information that were previously unprocessable by computer.

Glossary: Information in Text

Information Extraction (IE) involves pulling features – entities & their attributes, facts, relationships, etc. – out of textual sources.

Entity: Typically a name (person, place, organization, etc.) or a patterned composite (phone number, e-mail address).

Concept: An abstract entity or collection of entities.

Co-reference: Multiple expressions that describe the same thing. ***Anaphora*** including pronoun use is an example:

John pushed *Max*. *He* fell.

John pushed Max. *He* laughed.

-- Laure Vieu and Patrick Saint-Dizier

Feature: An element of interest, e.g., an entity, concept, topic, event, etc.

Glossary : Information in Text

Fact: A relationship between two entities or an entity and an attribute.

Sentiment: A valuation at the entity or higher level. Polarity/valence/tone and intensity are sentiment attributes.

Opinion: A statement that involves a sentiment. Opinion holder is typically different from the opinion object.

Semantics: Meaning, typically contextually dependent and hinted at by…

Syntax: The arrangement of words and terms.

Glossary: Methods

Natural Language Processing (NLP): Computers hear humans.

Parsing: Evaluating the content of a document or text.

Tokenization: Identification of distinct elements, e.g., words, punctuation marks, ***n-grams***.

Stemming/Lemmatization: Reducing word variants (conjugation, declension, case, pluralization) to bases.

Term reduction: Use of synonyms, taxonomy, similarity measures to group like terms.

Tagging: Wrapping XML tags around distinct features, a.k.a. ***text augmentation***. May involve ***text enrichment***.

POS Tagging: Specifically identifying parts of speech.

Glossary: Organizing and Structuring

Categorization: Specification of ***feature*** groupings.

Clustering: Creating categories according to outcome-similarity criteria.

Taxonomy: An exhaustive, hierarchical categorization of entities and concepts, either specified or generated by clustering.

Classification: Assigning an item to a category, perhaps using a taxonomy.

Ontology : In practice, a classification of a set of items in a way that represents knowledge.

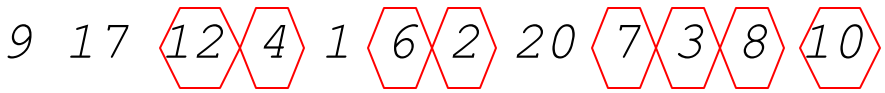
> *An oak is a tree. A rose is a flower. A deer is an animal. A sparrow is a bird. Russia is our fatherland. Death is inevitable.*
>
> -- P. Smirnovskii, A Textbook of Russian Grammar

Glossary: Evaluation

Precision: The proportion of decisions (e.g., classifications) that are correct.

Recall: The proportion of actual correct decisions (e.g., classifications) relative to the total number of correct decisions.

Find the even numbers:



What is my Precision? What is my Recall?

Accuracy: How well an IE or IR task has been performed, computed as an ***F-score*** weighting ***Precision*** & ***Recall***, typically:

f = 2(precision * recall) / (precision + recall)*

Relevance: Do results match the individual user's needs?

Text Analytics Pipeline

Typical steps in text analytics include –

1. Identify and **retrieve** documents for analysis.
2. Apply statistical &/ linguistic &/ structural techniques to **discern, tag, and extract** entities, concepts, relationships, and events (features) within document sets.
3. Apply statistical pattern-matching & similarity techniques to **classify** documents and organize extracted features according to a specified or generated categorization / taxonomy.

– via a *pipeline* of statistical & linguistic steps.

Let's look at them, at steps to model text...

Modelling Text

Metadata.

E.g., title, author, date.

Statistics.

Typically via vector space methods.

E.g., term frequency, co-occurrence, proximity.

Linguistics.

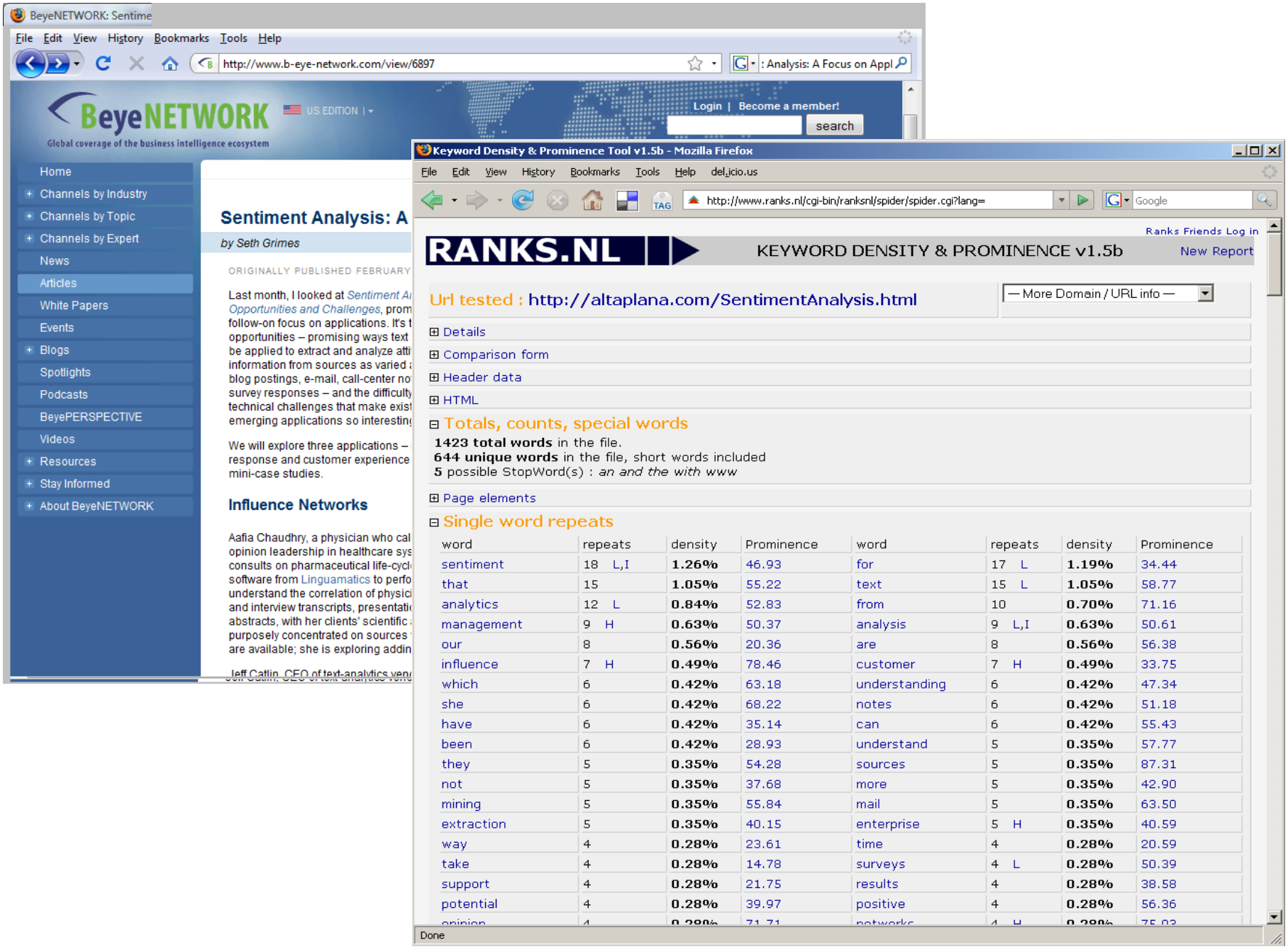
Lexicons, gazetteers, phrase books.

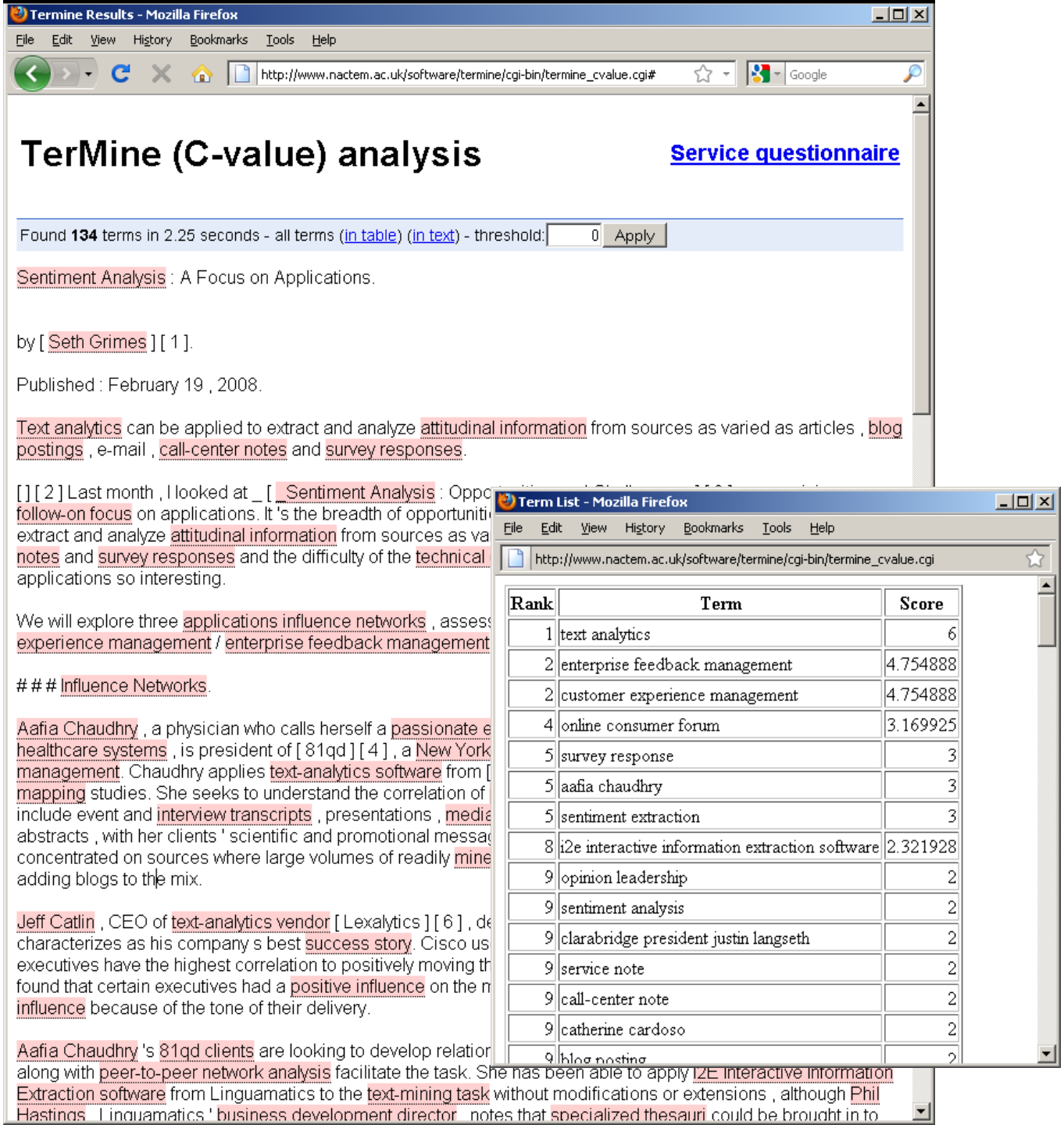
Word morphology, parts of speech, syntactic rules.

Semantic networks.

Larger-scale structure including discourse.

Machine learning.


BeyeNETWORK: Sentime
File Edit View History Bookmarks Tools Help
http://www.b-eye-network.com/view/6897
: Analysis: A Focus on Appl
BeyeNETWORK
US EDITION
Global coverage of the business intelligence ecosystem
Login | Become a member!
search
Home
Channels by Industry
Channels by Topic
Channels by Expert
News
Articles
White Papers
Events
Blogs
Spotlights
Podcasts
BeyePERSPECTIVE
Videos
Resources
Stay Informed
About BeyeNETWORK
Sentiment Analysis: A
by Seth Grimes
ORIGINALLY PUBLISHED FEBRUARY
Last month, I looked at Sentiment A
Opportunities and Challenges, prom
follow-on focus on applications. It's t
opportunities – promising ways text
be applied to extract and analyze atti
information from sources as varied
blog postings, e-mail, call-center no
survey responses – and the difficulty
technical challenges that make exist
emerging applications so interestin
We will explore three applications –
response and customer experience
mini-case studies.
Influence Networks
Aafia Chaudhry, a physician who cal
opinion leadership in healthcare sys
consults on pharmaceutical life-cycl
software from Linguamatics to perfo
understand the correlation of physici
and interview transcripts, presentati
abstracts, with her clients' scientific
purposely concentrated on sources
are available; she is exploring addin
Jeff Catlin, CEO of text-analytics ven
Keyword Density & Prominence Tool v1.5b - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
http://www.ranks.nl/cgi-bin/ranksnl/spider/spider.cgi?lang=
Google
Ranks Friends Log in
RANKS.NL
KEYWORD DENSITY & PROMINENCE v1.5b
New Report
Url tested : http://altaplana.com/SentimentAnalysis.html
— More Domain / URL info —
Details
Comparison form
Header data
HTML
Totals, counts, special words
1423 total words in the file.
644 unique words in the file, short words included
5 possible StopWord(s) : an and the with www
Page elements
Single word repeats
word repeats density Prominence word repeats density Prominence
sentiment 18 L,I 1.26% 46.93 for 17 L 1.19% 34.44
that 15 1.05% 55.22 text 15 L 1.05% 58.77
analytics 12 L 0.84% 52.83 from 10 0.70% 71.16
management 9 H 0.63% 50.37 analysis 9 L,I 0.63% 50.61
our 8 0.56% 20.36 are 8 0.56% 56.38
influence 7 H 0.49% 78.46 customer 7 H 0.49% 33.75
which 6 0.42% 63.18 understanding 6 0.42% 47.34
she 6 0.42% 68.22 notes 6 0.42% 51.18
have 6 0.42% 35.14 can 6 0.42% 55.43
been 6 0.42% 28.93 understand 5 0.35% 57.77
they 5 0.35% 54.28 sources 5 0.35% 87.31
not 5 0.35% 37.68 more 5 0.35% 42.90
mining 5 0.35% 55.84 mail 5 0.35% 63.50
extraction 5 0.35% 40.15 enterprise 5 H 0.35% 40.59
way 4 0.28% 23.61 time 4 0.28% 20.59
take 4 0.28% 14.78 surveys 4 L 0.28% 50.39
support 4 0.28% 21.75 results 4 0.28% 38.58
potential 4 0.28% 39.97 positive 4 0.28% 56.36
Done


Termine Results - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.nactem.ac.uk/software/termine/cgi-bin/termine_cvalue.cgi#
Google
TerMine (C-value) analysis
Service questionnaire
Found 134 terms in 2.25 seconds - all terms (in table) (in text) - threshold: 0 Apply
Sentiment Analysis : A Focus on Applications.
by [Seth Grimes] [1].
Published : February 19 , 2008.
Text analytics can be applied to extract and analyze attitudinal information from sources as varied as articles , blog postings , e-mail , call-center notes and survey responses.
[ ] [2] Last month , I looked at _ [_Sentiment Analysis : Oppo
follow-on focus on applications. It 's the breadth of opportuniti
extract and analyze attitudinal information from sources as va
notes and survey responses and the difficulty of the technical
applications so interesting.
We will explore three applications influence networks , assess
experience management / enterprise feedback management
Influence Networks.
Aafia Chaudhry , a physician who calls herself a passionate e
healthcare systems , is president of [81qd] [4] , a New York
management. Chaudhry applies text-analytics software from [
mapping studies. She seeks to understand the correlation of
include event and interview transcripts , presentations , media
abstracts , with her clients ' scientific and promotional messag
concentrated on sources where large volumes of readily mine
adding blogs to the mix.
Jeff Catlin , CEO of text-analytics vendor [Lexalytics] [6] , de
characterizes as his company s best success story. Cisco us
executives have the highest correlation to positively moving th
found that certain executives had a positive influence on the n
influence because of the tone of their delivery.
Aafia Chaudhry 's 81qd clients are looking to develop relation
along with peer-to-peer network analysis facilitate the task. She has been able to apply I2E Interactive Information
Extraction software from Linguamatics to the text-mining task without modifications or extensions , although Phil
Hastings , Linguamatics ' business development director , notes that specialized thesauri could be brought in to
Term List - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.nactem.ac.uk/software/termine/cgi-bin/termine_cvalue.cgi
Rank Term Score
1 text analytics 6
2 enterprise feedback management 4.754888
2 customer experience management 4.754888
4 online consumer forum 3.169925
5 survey response 3
5 aafia chaudhry 3
5 sentiment extraction 3
8 i2e interactive information extraction software 2.321928
9 opinion leadership 2
9 sentiment analysis 2
9 clarabridge president justin langseth 2
9 service note 2
9 call-center note 2
9 catherine cardoso 2
9 blog posting 2

Significant words in descending order of frequency (common words omitted).

46	*nerve*	12	*body*	6	*disturbance*	4	*accumulate*
40	*chemical*	12	*effects*	6	*related*	4	*balance*
28	*system*	12	*electrical*	5	*control*	4	*block*
22	*communication*	12	*mental*	5	*diagram*	4	*disorders*
19	*adrenalin*	12	*messengers*	5	*fibers*	4	*end*
18	*cell*	10	*signals*	5	*gland*	4	*excitation*
18	*synapse*	10	*stimulation*	5	*mechanisms*	4	*health*
16	*impulses*	8	*action*	5	*mediators*	4	*human*
16	*inhibition*	8	*ganglion*	5	*organism*	4	*outgoing*
15	*brain*	7	*animal*	5	*produce*	4	*reaching*
15	*transmission*	7	*blood*	5	*regulate*	4	*recording*
13	*acetylcholine*	7	*drugs*				
13	*experiment*	7	*normal*				
13	*substances*						

Total word occurrences in the document: . . . . . .

Different words in document:

Total of different words . . . . . .

Less different common words . . . . . .

Different non-common words . . . . . .

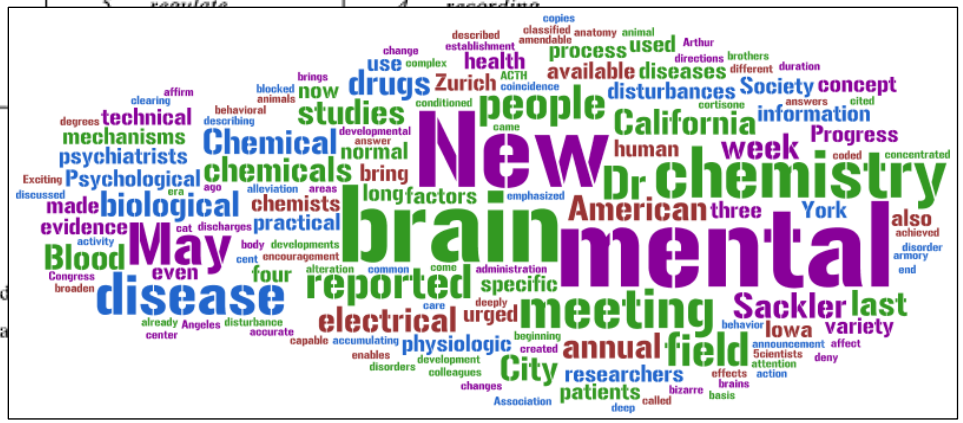
Ratio of all word occurrences to different non-common word

Non-common words having a frequency of occurrence of 5 a

Total occurrences . . . . . .

Different words . . . . . .

http://wordle.net



IBM JOURNAL • APRIL 1958

"Statistical information derived from word frequency and distribution is used by the machine to compute a relative measure of significance, first for individual words and then for sentences. Sentences scoring highest in significance are extracted and printed out to become the auto-abstract."

-- H.P. Luhn, *The Automatic Creation of Literature Abstracts*, IBM Journal, 1958.

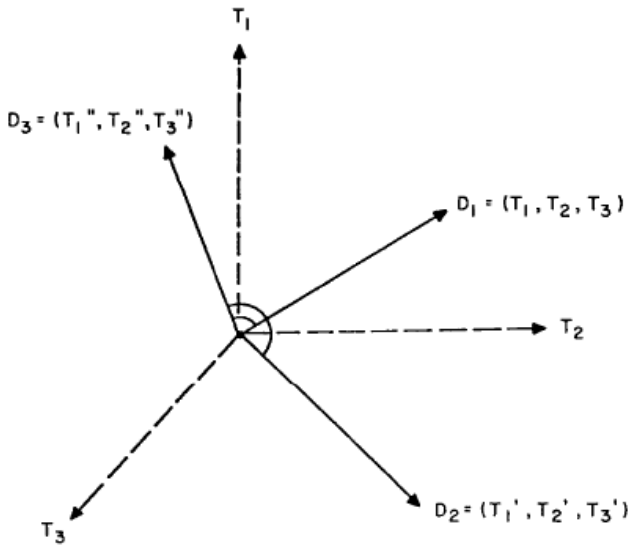
Modelling Text

The text content of a document can be considered an unordered "bag of words."

Particular documents are points in a high-dimensional vector space.

Salton, Wong & Yang, "A Vector Space Model for Automatic Indexing," November 1975.

Fig. 1. Vector representation of document space.



Modelling Text

We might construct a document-term matrix...

D1 = "I like databases"

D2 = "I hate hate databases"

	I	like	hate	databases
D1	1	1	0	1
D2	1	0	2	1

http://en.wikipedia.org/wiki/Term-document_matrix

and use a weighting such as TF-IDF (term frequency–inverse document frequency)...

in computing the cosine of the angle between weighted doc-vectors to determine similarity.

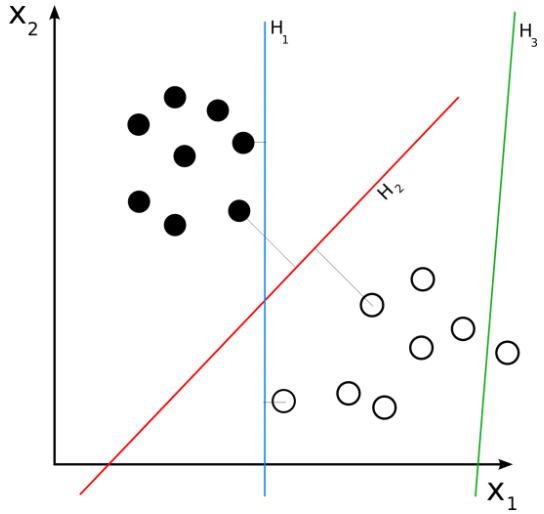
Modelling Text

Analytical methods make text tractable.

- Latent semantic indexing utilizing singular value decomposition for term reduction / feature selection.
  - Creates a new, reduced concept space.
  - Takes care of synonymy, polysemy, stemming, etc.

Classification technologies / methods:

- Naive Bayes.
- Support Vector Machine.
- K-nearest neighbor.



Modelling Text

In the form of *query-document similarity*, this is Information Retrieval 101.

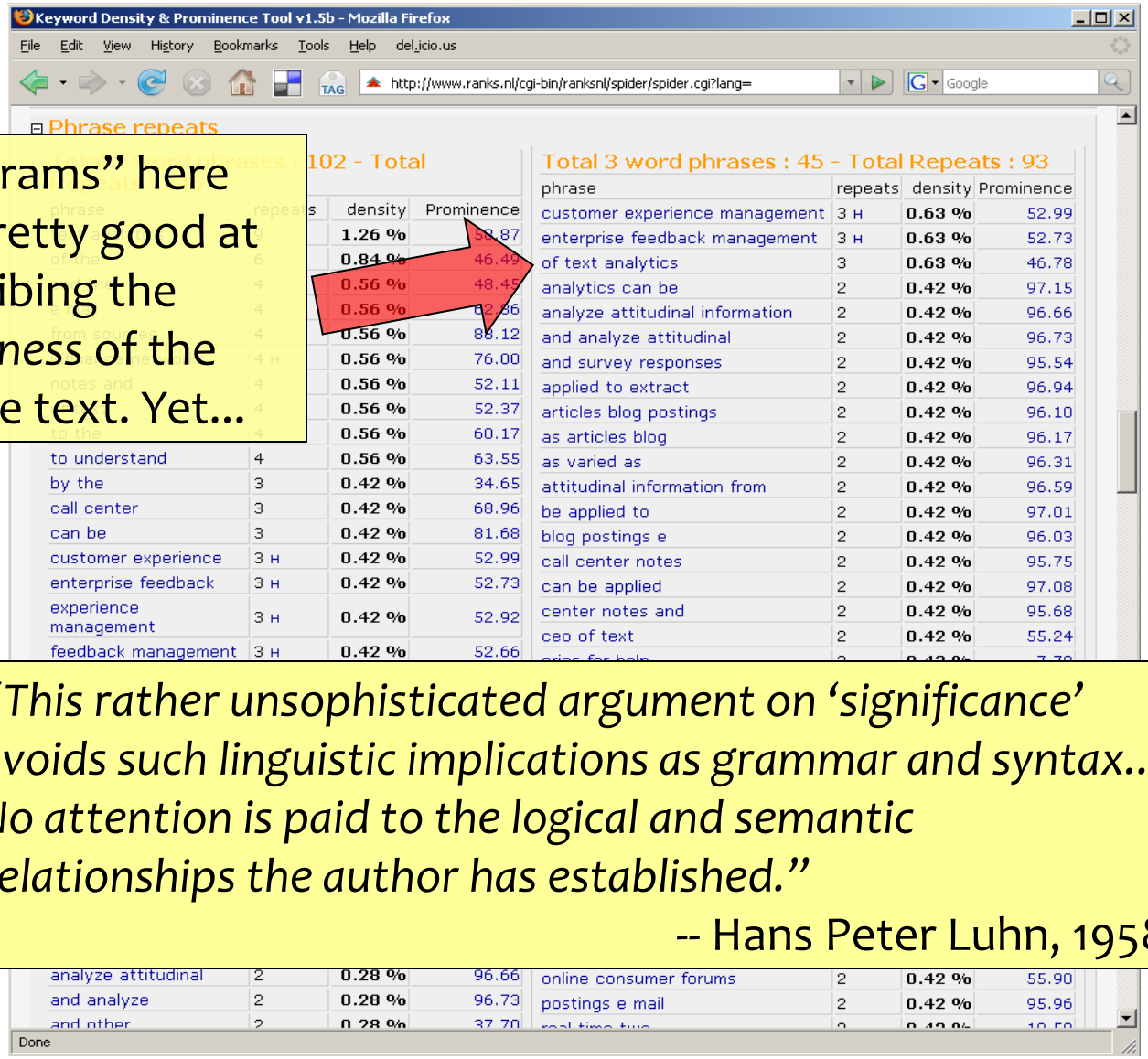
See, for instance, Salton & Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," 1988.

A useful basic tech paper: Russ Albright, SAS, "Taming Text with the SVD," 2004.

Given the complexity of human language, statistical models may fall short.

"Reading from text in general is a hard problem, because it involves all of common sense knowledge."

-- Expert systems pioneer Edward A. Feigenbaum



"Tri-grams" here are pretty good at describing the *Whatness* of the source text. Yet...

"This rather unsophisticated argument on 'significance' avoids such linguistic implications as grammar and syntax... No attention is paid to the logical and semantic relationships the author has established."

-- Hans Peter Luhn, 1958

New York Times,
September 8, 1957

Anaphora / coreference:
"They"

SCIENCE IN REVIEW

Chemistry Is Employed in a Search for New Methods to Conquer Mental Illness

By ROBERT K. PLUMB

By coincidence this week-end in New York City marks the end of the annual meeting of the American Psychological Association and the begining of the annual meeting of the American Chemical Society.

Psychologists and chemists have never had so much in common as they now have in new studies of the chemical basis for human behavior. Exciting new finds in this field were also discussed last week in Iowa City, Iowa, at the annual meeting of the American Physiological Society and at Zurich, Switzerland, at the Second International Congress for Psychiatry.

Two major recent developments have called the attention of chemists, physiologists, physicists and other scientists to mental diseases: It has been found that extremely minute quantities of chemicals can induce hallucinations and bizarre psychic disturbances in normal people, and mood-altering drugs (tranquilizers, for instance) have made long-institutionalized people amenable to therapy.

Money to finance research on the physical factors in mental illness is being made available. Progress has been achieved toward the understanding of the chemistry of the brain. New goals are in sight.

At the psychiatrists meeting in Zurich last week, four New York City physicians urged their colleagues to broaden their concept of "mental disease," and to probe more deeply into the chemistry and metabolism of the human body for answers to mental disorders and their prevention.

Blood May Tell

Dr. Felix Marti-Ibanez and three brothers, Dr. Mortimer D. Sackler, Dr. Raymond R. Sackler and Dr. Arthur M. Sackler, cited evidence that the blood chemistry of victims of schizophrenia is different from that of normal people. Perhaps multiple biological factors are responsible for this chemical change, they suggested.

Mental disease is a "developmental process" and long duration of a disorder may result in "permanent alteration of anatomy and physiology," they said. They urged that trials of new drugs which affect the brain should be concentrated on complex studies of the mechanism of action of the drugs. The variety of substances capable of producing profound mental effects is a new armory of weapons for use in investigating biological mechanisms underlying mental disease, they said.

The sources of behavioral disturbance are many and they may come from external as well as internal forces, the four reported. This concept has already proven practical, for instances, when it enabled psychiatrists to predict that the administration of ACTH and cortisone could produce psychosis.

"It led some years ago to the development of a blood test which was 80 per cent accurate in the identification of schizophrenic patients," they said. "It permitted us on physiologic grounds to deny that the psychoneuroses and the psychoses were lesser and greater degrees of the same disease process, and, in fact, to affirm that they represented opposite and even mutually exclusive directions of physiologic disturbances," they said.

Chemicals now available should be used not only to bring relief to the mentally sick but also to uncover the biological mechanisms of the disease processes themselves. "Only then will the metabolic era mature and bring to fruition man's long hoped for salvation from the ravages of mental disease," they reported.

Chemistry of the Brain

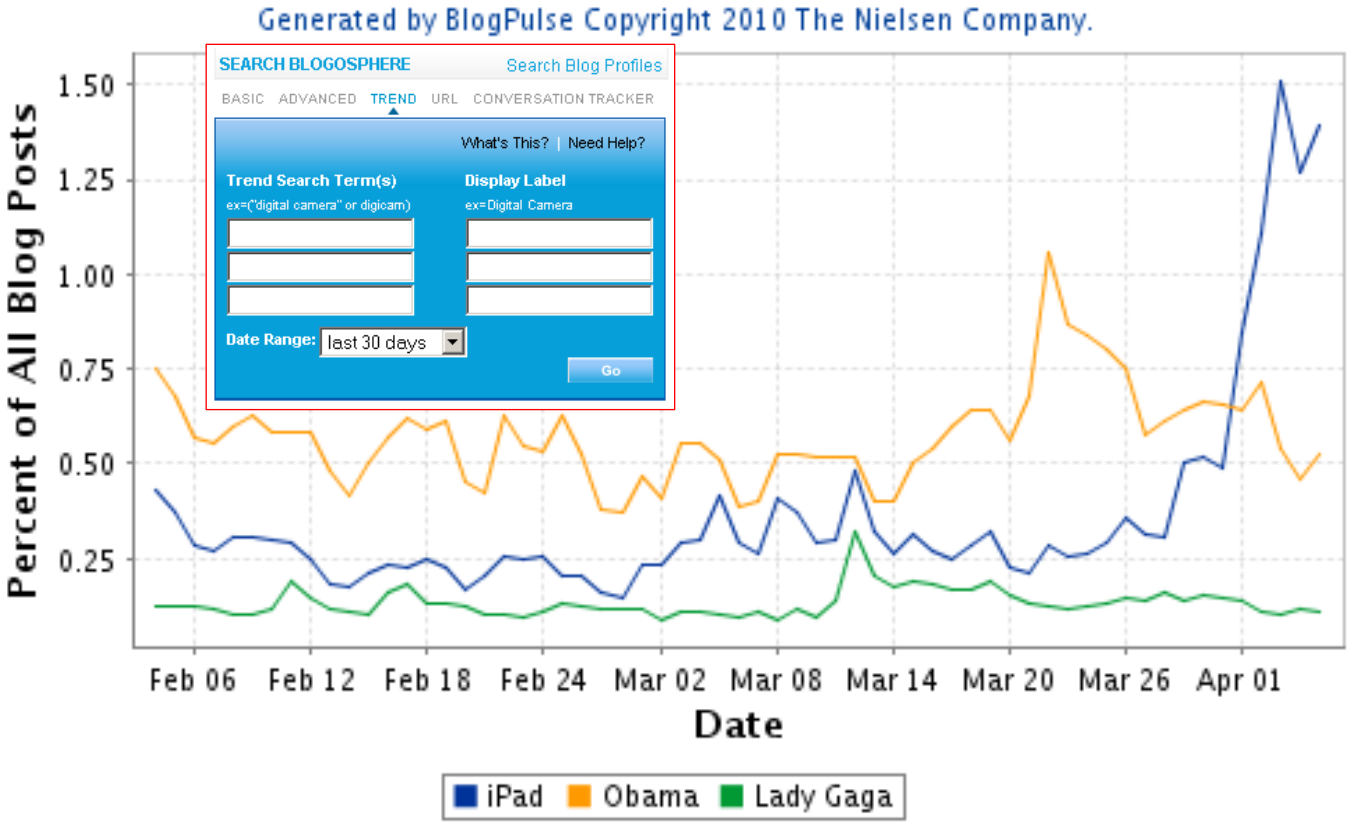
At the psychologist's meeting here, a technique for tracing electrical activity in specific portions of the animal brain was described by researchers of California ... reported in cat ... electric ... animals ... which ... In this ... reported ... sequence ... its variation ... S... the brain may be located. Furthermore, the electrical pathways so traced out can be blocked temporarily by the use of chemicals. This poses new possibilities for studying brain ... and sig... the Ca... sized.

The ... try hav... peutic ... courage ... for me... that kn... ary fiel... last we... Washin...

This ... the Na... Health ... formati... Literat... fied an... bers ca... technic... People ... vited to ... or othe... letters ... have in... Informa... Silver ... Spring, Md.

Advanced Term Counting

Counting term hits, in one source, at the doc level, doesn't take you far…



Good or bad? What's behind the posts?

Why Do We Need Linguistics?

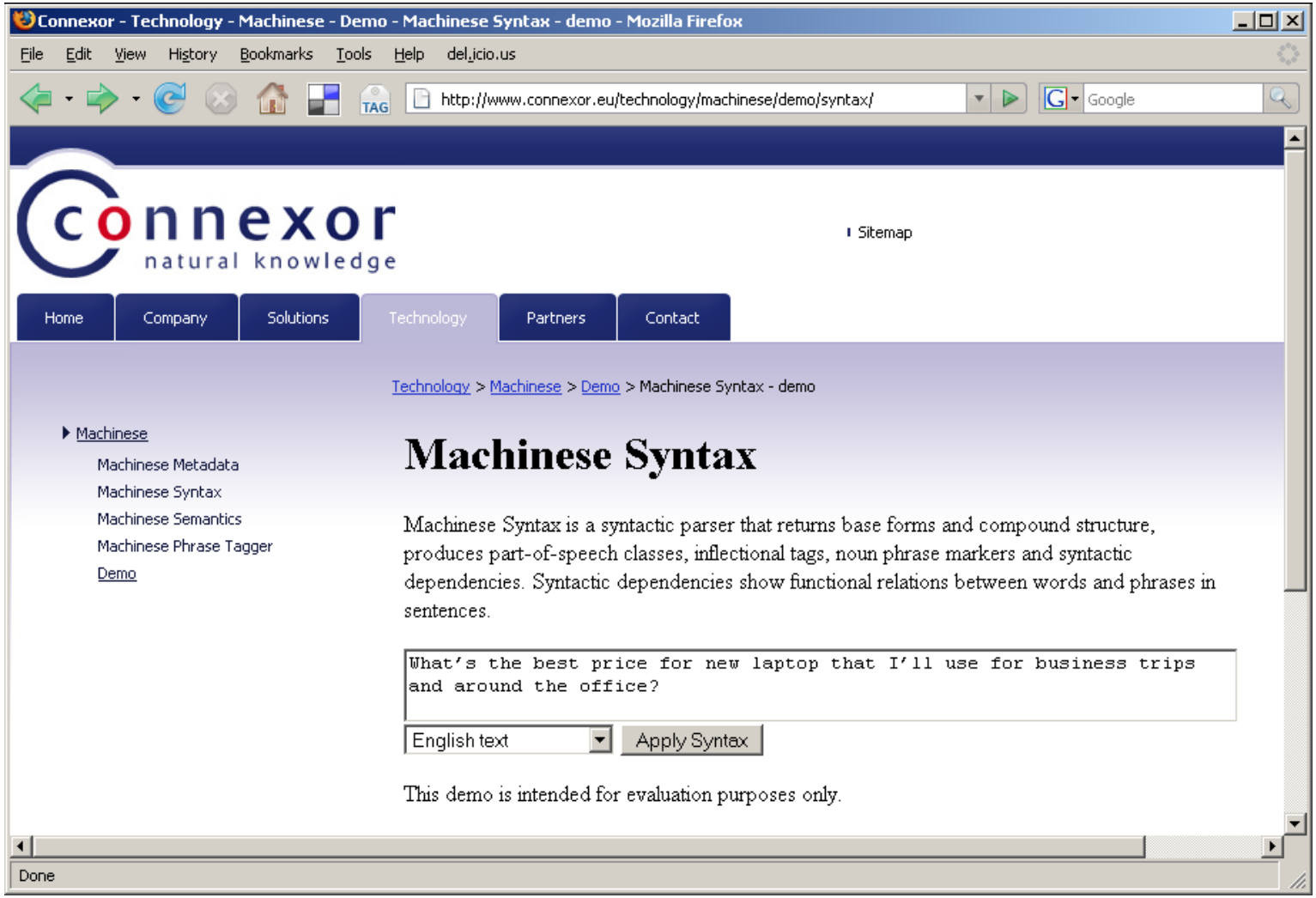
To get more out of text than can be delivered by a bag/vector of words and term counting.

> The Dow *fell* 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index *gained*1.44, or 0.11 percent, to 1,263.85.
> The Dow *gained* 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index *fell* 1.44, or 0.11 percent, to 1,263.85.
>
> -- Luca Scagliarini, Expert System

> Time *flies like* an arrow. Fruit *flies like* a banana.
>
> -- Groucho Marx

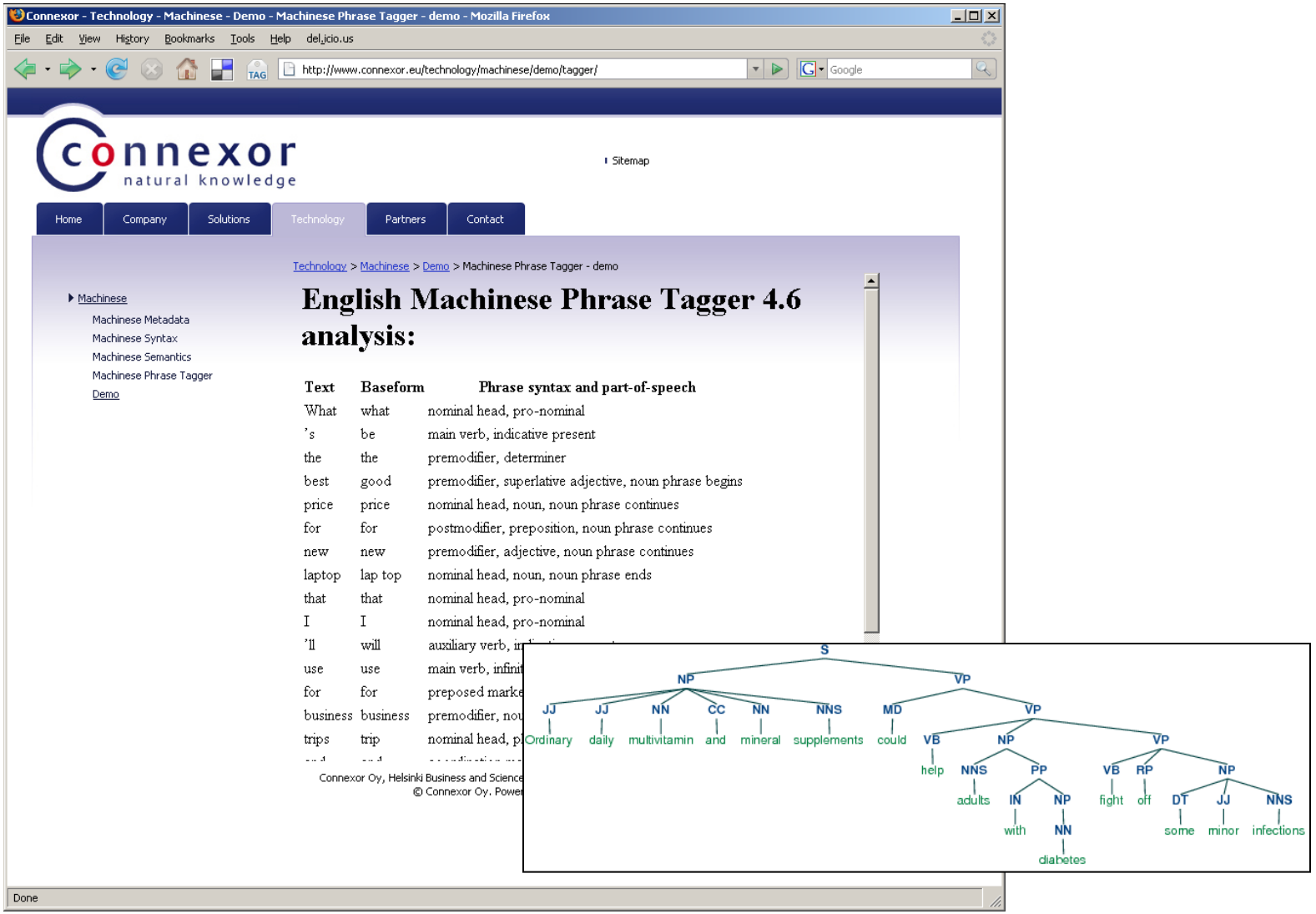
(Statistical co-occurrence to build a model for analysis of text such as these is possible but still limited.)

Parts of Speech



Parts of Speech

Parts of Speech



From POS to Relationships

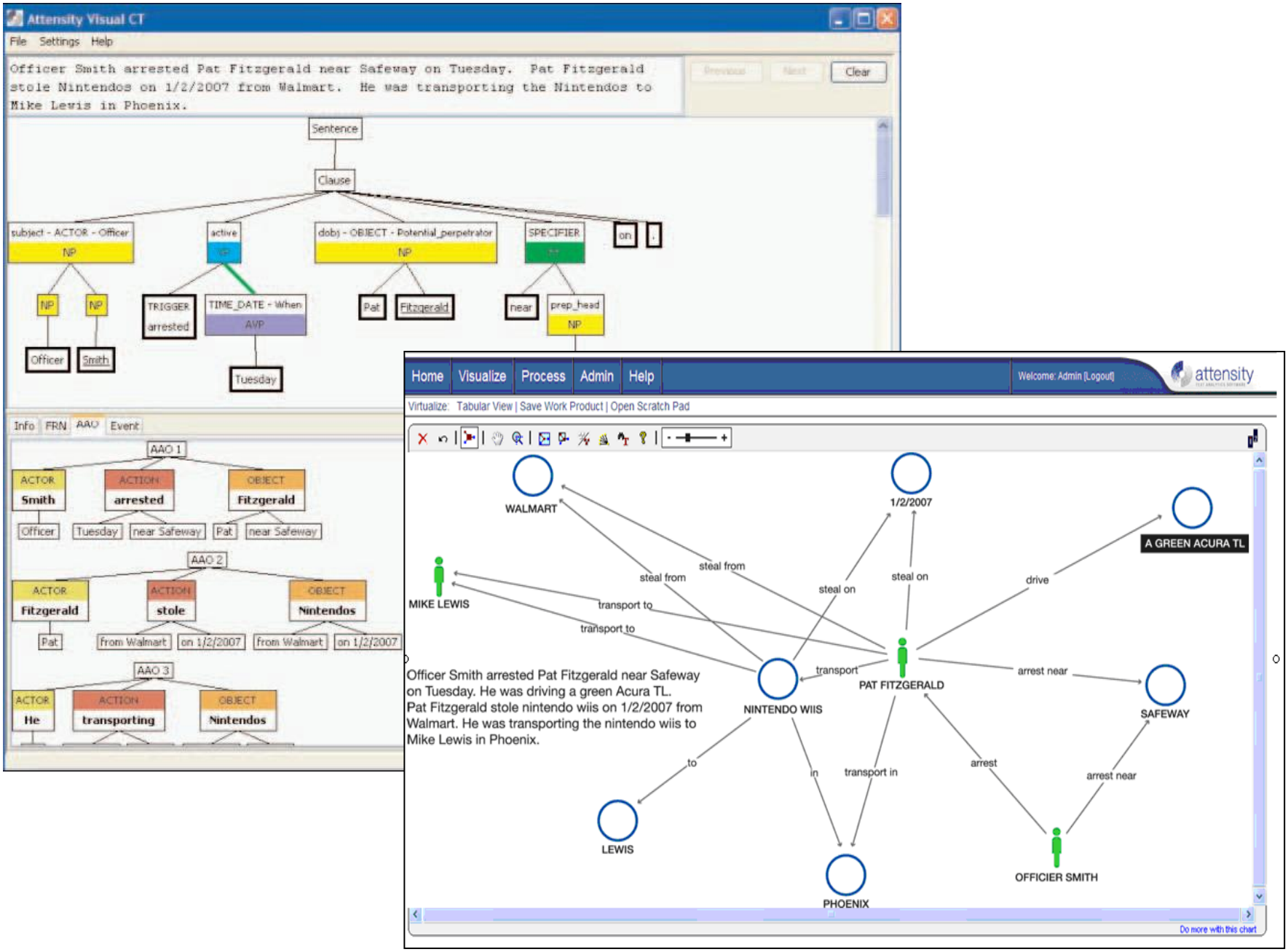
When we understand, for instance, parts of speech (POS), e.g. –

<subject> <verb> <object>

– we're in a position to discern *facts* and *relationships*...

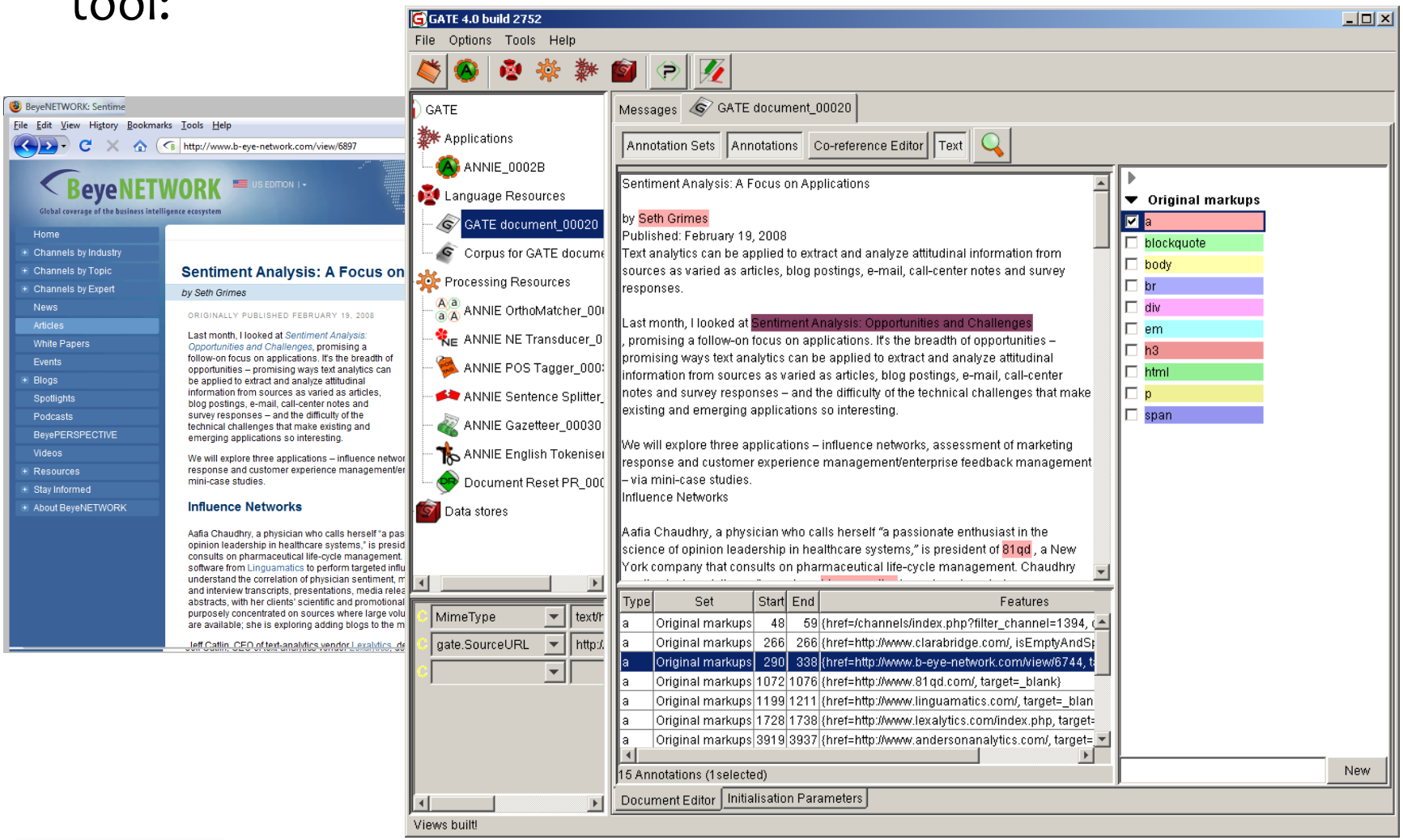
Semantics networks such as WordNet are an asset for word-sense disambiguation.

> *"WordNet is a large lexical database of English... Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations… WordNet's structure makes it a useful tool for computational linguistics and natural language processing."*
>
> *http://wordnet.princeton.edu/*

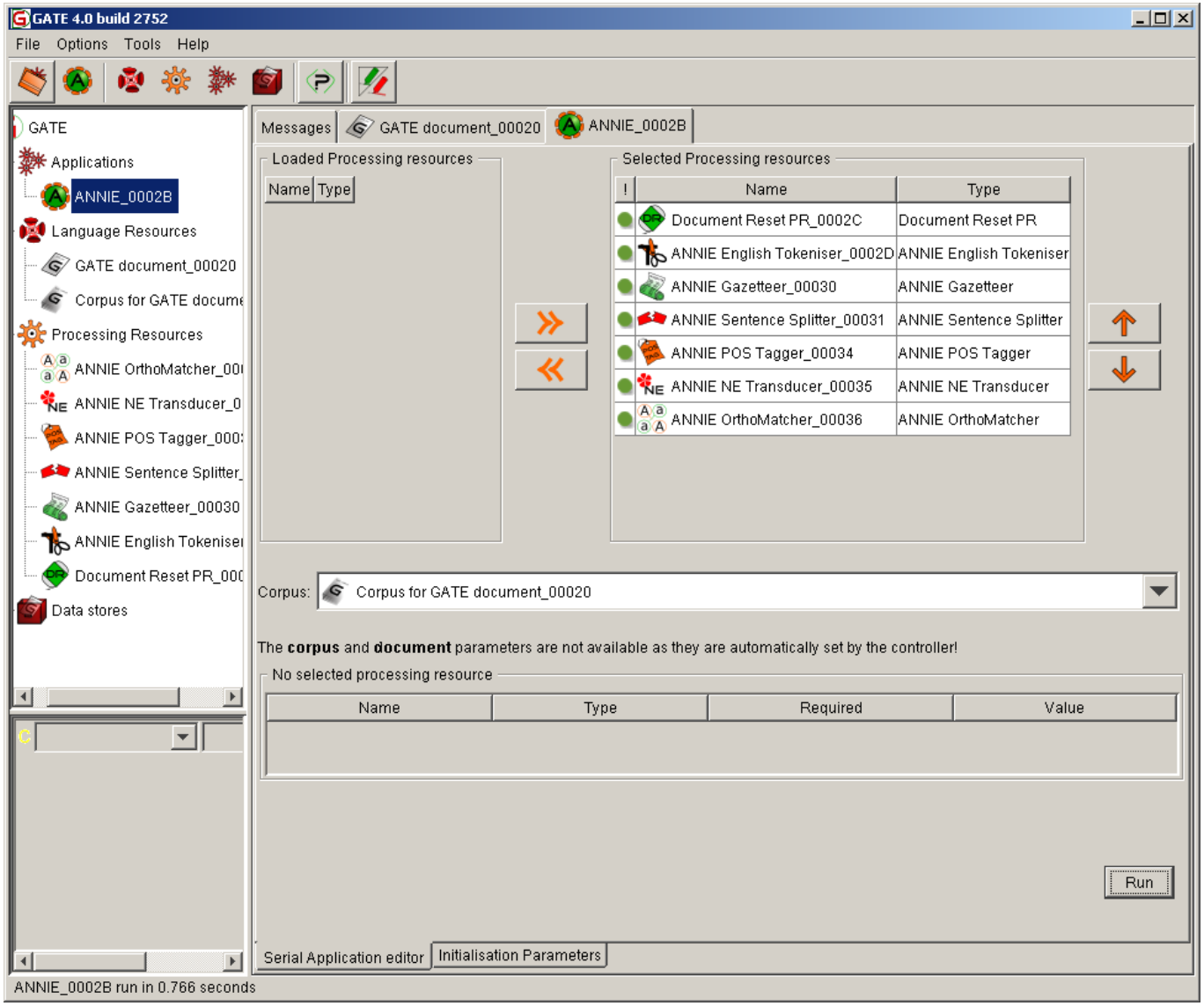

Attensity Visual CT
File Settings Help
Officer Smith arrested Pat Fitzgerald near Safeway on Tuesday. Pat Fitzgerald stole Nintendos on 1/2/2007 from Walmart. He was transporting the Nintendos to Mike Lewis in Phoenix.
Clear
Sentence
Clause
subject - ACTOR - Officer
active
dobj - OBJECT - Potential_perpetrator
SPECIFIER
on
.
NP
VP
NP
NP
NP
TRIGGER
arrested
TIME_DATE - When
AVP
Pat
Fitzgerald
near
prep_head
NP
Officer
Smith
Tuesday
Info FRN AAO Event
AAO 1
ACTOR
Smith
ACTION
arrested
OBJECT
Fitzgerald
Officer
Tuesday
near Safeway
Pat
near Safeway
AAO 2
ACTOR
Fitzgerald
ACTION
stole
OBJECT
Nintendos
Pat
from Walmart
on 1/2/2007
from Walmart
on 1/2/2007
AAO 3
ACTOR
He
ACTION
transporting
OBJECT
Nintendos
Home Visualize Process Admin Help
Welcome: Admin [Logout]
attensity
Virtualize: Tabular View | Save Work Product | Open Scratch Pad
WALMART
1/2/2007
A GREEN ACURA TL
MIKE LEWIS
steal from
steal from
steal on
steal on
drive
transport to
transport to
transport
arrest near
PAT FITZGERALD
NINTENDO WIIS
SAFEWAY
Officer Smith arrested Pat Fitzgerald near Safeway on Tuesday. He was driving a green Acura TL. Pat Fitzgerald stole nintendo wiis on 1/2/2007 from Walmart. He was transporting the nintendo wiis to Mike Lewis in Phoenix.
to
in
transport in
arrest
arrest near
LEWIS
PHOENIX
OFFICIER SMITH
Do more with this chart

Tagging with GATE

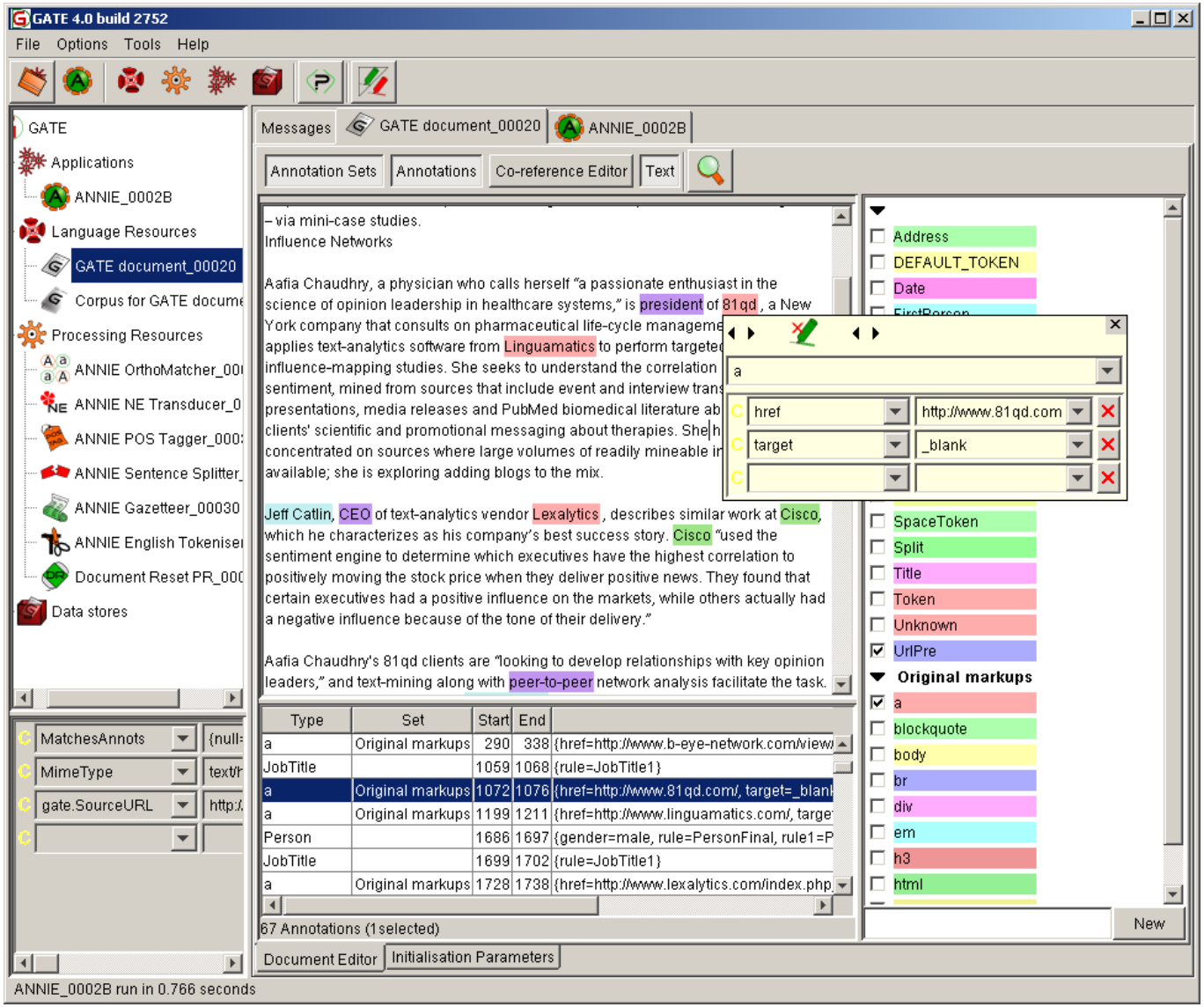
Annotation (tagging) in action via GATE, an open-source tool:



GATE Language Processing Pipeline



GATE Text Annotation



GATE Exported XML

```
<?xml version='1.0' encoding='windows-1252'?>
<GateDocument>
<!-- The document's features-->
    <GateDocumentFeatures>
    <Feature>
      <Name className="java.lang.String">MimeType</Name>
      <Value className="java.lang.String">text/html</Value>
    </Feature>
    <Feature>
      <Name className="java.lang.String">gate.SourceURL</Name>
      <Value className="java.lang.String">http://altaplana.com/SentimentAnalysis.html</Value>
    </Feature>
    </GateDocumentFeatures>
<!-- The document content area with serialized nodes -->
    <TextWithNodes><Node id="0" />Sentiment<Node id="9" /> <Node id="10" />Analysis<Node id="18"
    />:<Node id="19" /> <Node id="20" />A<Node id="21" /> <Node id="22" />Focus<Node id="27" /> <Node
    id="28" />on<Node id="30" /> <Node id="31" />Applications<Node id="43" />
    <Node id="44" />
    <Node id="45" />by<Node id="47" /> <Node id="48" />Seth<Node id="52" /> <Node id="53"
    />Grimes<Node id="59" />                                                          [material cut]
    </TextWithNodes>
<!-- The default annotation set -->
    <AnnotationSet>                                                                  [material cut]
      <Annotation Id="67" Type="Token" StartNode="48" EndNode="52">
          <Feature><Name className="java.lang.String">length</Name><Value
              className="java.lang.String">4</Value></Feature>
          <Feature><Name className="java.lang.String">category</Name><Value
              className="java.lang.String">NNP</Value></Feature>
          <Feature><Name className="java.lang.String">orth</Name><Value
              className="java.lang.String">upperInitial</Value></Feature>
          <Feature><Name className="java.lang.String">kind</Name><Value
              className="java.lang.String">word</Value></Feature>
          <Feature><Name className="java.lang.String">string</Name><Value
              className="java.lang.String">Seth</Value></Feature>
      </Annotation>                                                                  [material cut]
    </AnnotationSet>
</GateDocument>
```

Information Extraction

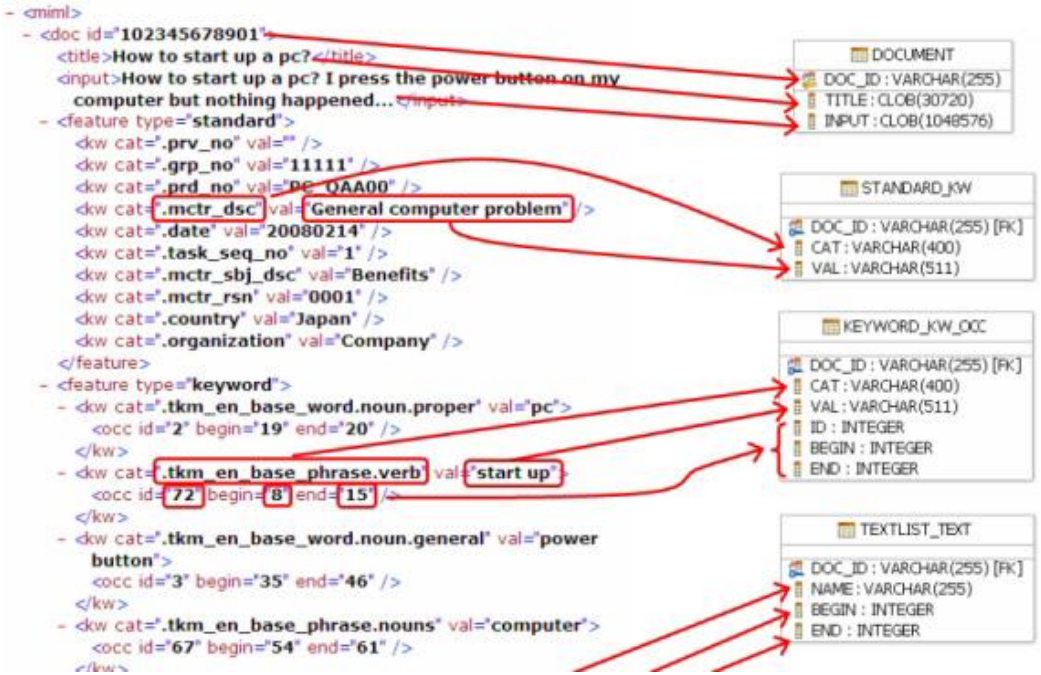
For content analysis, key in on extracting information.

- Text features are typically marked up (annotated) in-place with XML.
- Entities and concepts may correspond to dimensions in a standard BI model.
  - Both classes of object are hierarchically organized and have attributes.
  - We can have both discovered and predetermined classifications (taxonomies) of text features.
- Dimensional modelling facilitates extraction to databases...

Database Insert

Illustrated via an IBM example:

> *"The standard features are stored in the STANDARD_KW table, keywords with their occurrences in the KEYWORD_KW_OCC table, and the text list features in the TEXTLIST_TEXT table. Every feature table contains the DOC_ID as a reference to the DOCUMENT table."*



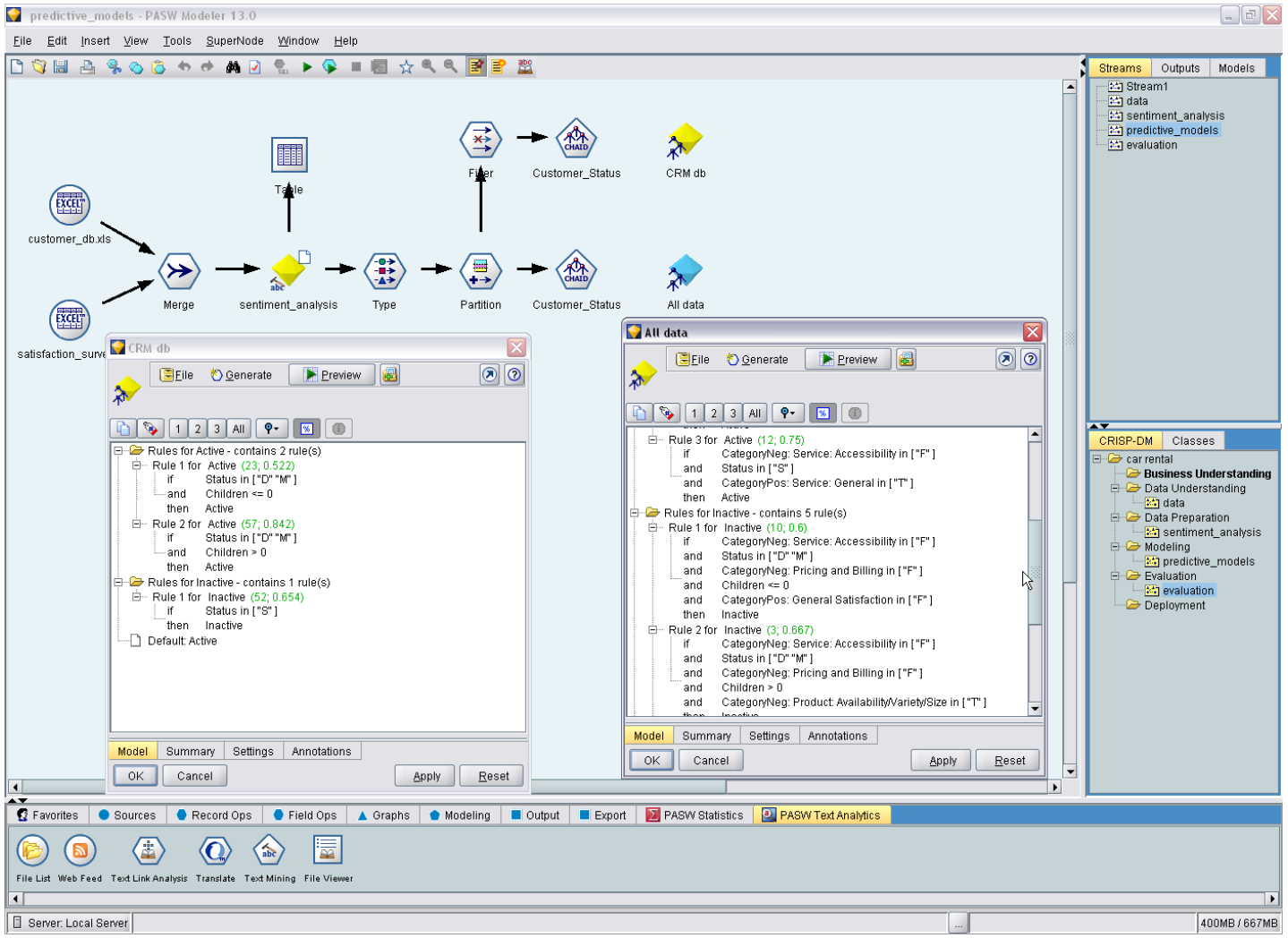
Sophisticate Pattern Matching

Lexicons and language rules boost accuracy. An example – a bit complicated – GATE Extension via JAPE Rules...

```
/* locationcontext2.jape
 * Dhavalkumar Thakker, Nottingham Trent University/PA Photos 15 Sept 2008*/
Phase: locationcontext2
Input: Lookup Token
Options: control = all debug = false
//Manchester, UK
Rule: locationcontext2
Priority:50
({Token.string == "at"})
    ( ({Token.string =~ "[Tt]he"})?
       (
          (
       {Token.kind == word, Token.category == NNP, Token.orth == upperInitial}
           ({Token.kind == punctuation})?
       {Token.kind == word, Token.category == NNP, Token.orth == upperInitial}
           ({Token.kind == punctuation})?
       {Token.kind == word, Token.category == NNP, Token.orth == upperInitial} )  |
    ( {Token.kind == word, Token.category == NNP, Token.orth == upperInitial}
           ({Token.kind == punctuation})?
    ( {Token.kind == word, Token.category == NNP, Token.orth == allCaps} |
           {Token.kind == word, Token.category == NNP, Token.orth == upperInitial} )
           ) |
...
```

Predictive Modeling

Another processing pipeline and more rules…



Predictive Modeling

In the text context, predictive analytics is mostly about classification and automated processing.

Modeling also helps, operationally, with:

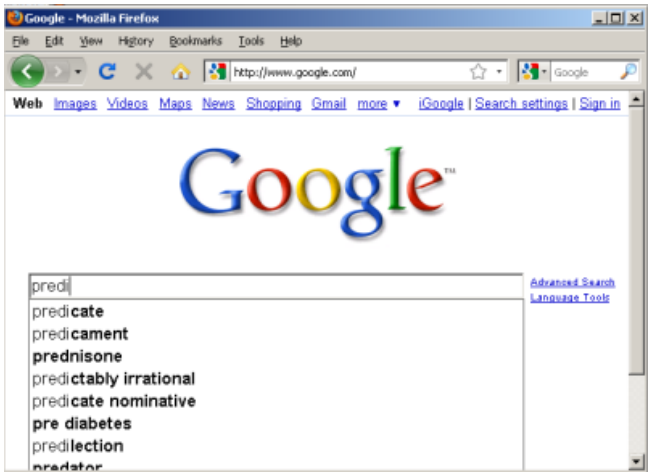
Completion

http://en.wikipedia.org/wiki/File:ITap_on_Motorola_C350.jpg

Disambiguation

: use dictionaries, context

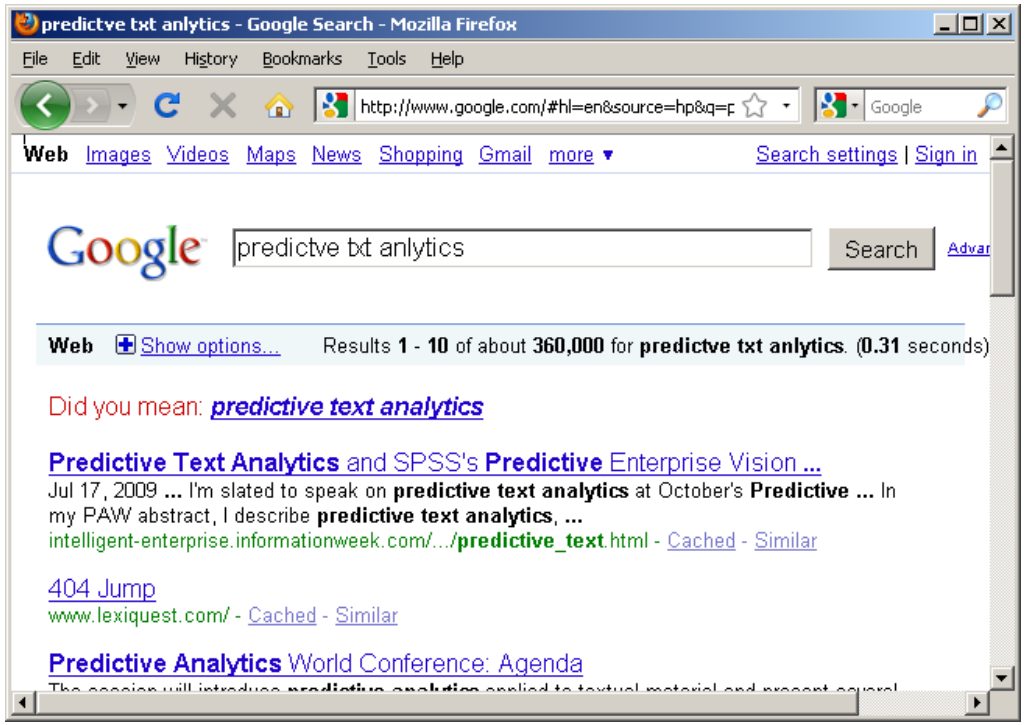
Error correction



Error Correction

"Search logs suggest that from 10-15% of queries contain spelling or typographical errors. Fittingly, one important query reformulation tool is spelling suggestions or corrections."

-- Marti Hearst, Search User Interfaces



Accuracy and Semi-Structured Sources

An e-mail message is "semi-structured," which facilitates extracting *metadata* --

```
Date: Sun, 13 Mar 2005 19:58:39 -0500
From: Adam L. Buchsbaum <alb@research.att.com>
To: Seth Grimes <grimes@altaplana.com>
Subject: Re: Papers on analysis on streaming data
seth, you should contact divesh srivastava,
divesh@research.att.com regarding at&t labs data streaming
technology.
Adam
```

"Reading from text in structured domains I don't think is as hard."

-- Edward A. Feigenbaum

Surveys are also typically s-s in a different way...

Structured &'Unstructured'

The respondent is invited to explain his/her attitude:

My overall experience was positive.	○	○	○	○	○
Please complete the section below if your contact with us involved permitting/licensing/registration assistance.					
The regulations were understandable.	○	○	○	○	○
The application instructions were understandable.	○	○	○	○	○
The terms and conditions of the permit, license, or registration were understandable.	○	○	○	○	○

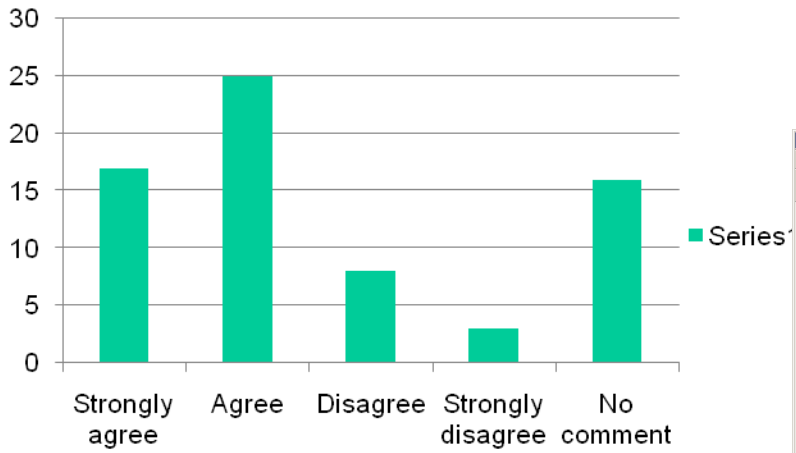
Please indicate the name(s) of any staff person you would like to commend:

Comments:

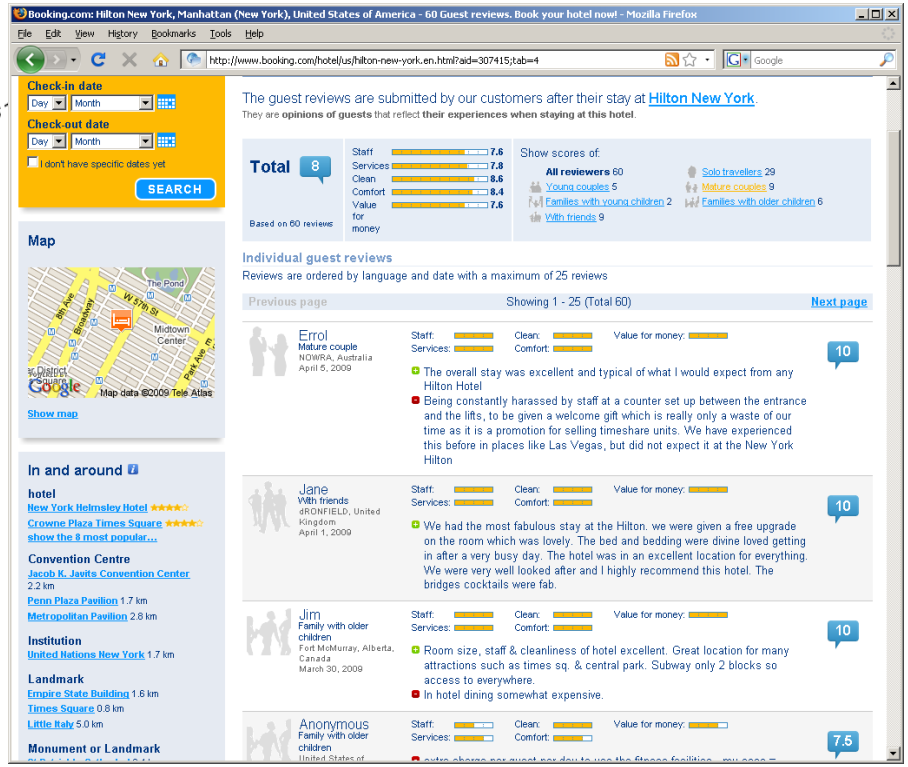
If you feel we fell short in meeting your service expectations, please describe the situation, including name of the staff person involved and the date the incident occurred:

Structured &'Unstructured'

We typically look at frequencies and distributions of coded-response questions:



Linkage of responses to coded ratings helps in analyses of free text:



Sentiment Analysis

"Sentiment analysis is the task of identifying positive and negative opinions, emotions, and evaluations."

-- Wilson, Wiebe & Hoffman, 2005, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis"

"Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in text… An opinion on a feature f is a positive or negative view, attitude, emotion or appraisal on f from an opinion holder."

-- Bing Liu, 2010, "Sentiment Analysis and Subjectivity," in Handbook of Natural Language Processing

"Dell really... REALLY need to stop overcharging... and when i say overcharing... i mean atleast double what you would pay to pick up the ram yourself."

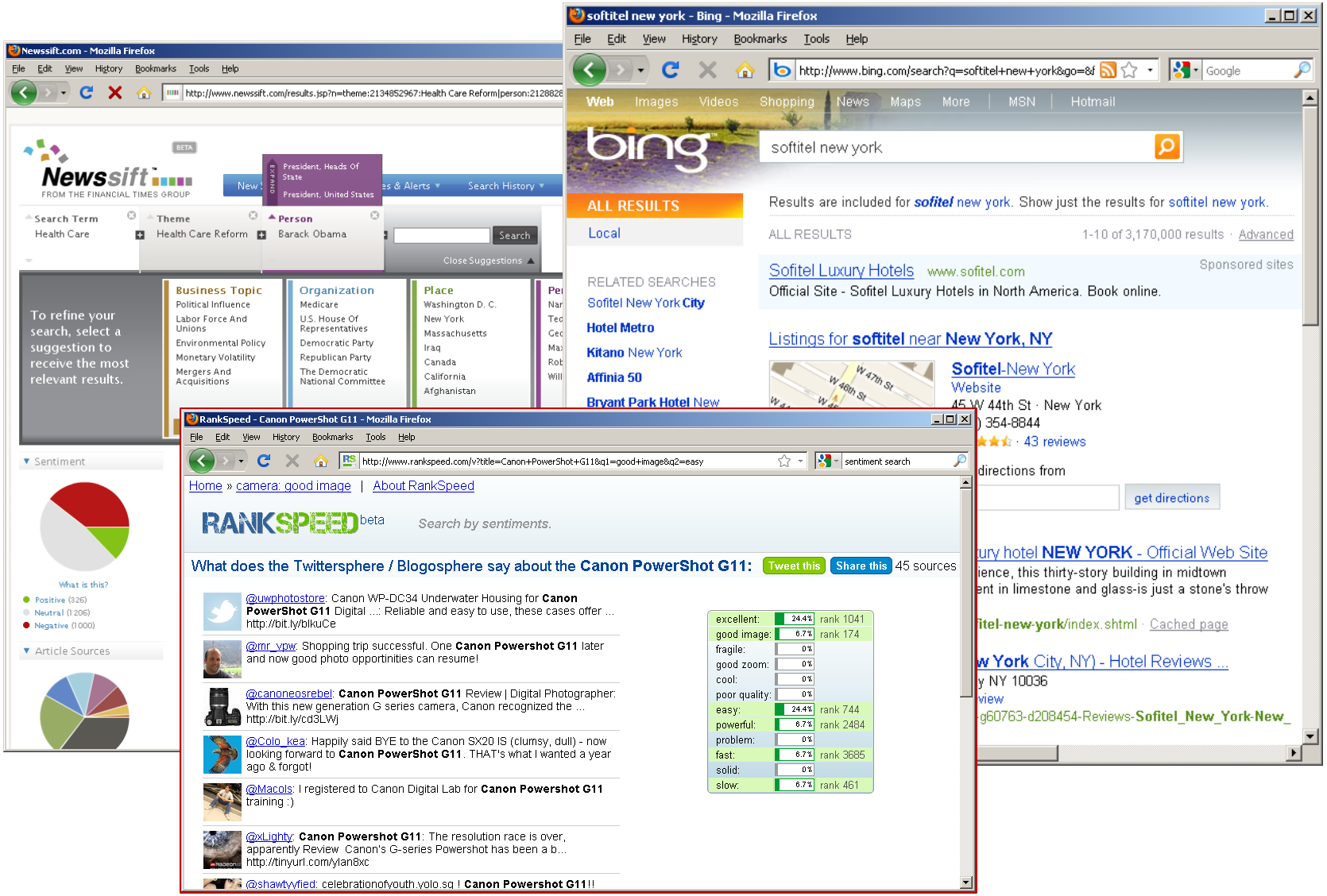
-- From Dell's *IdeaStorm.com*

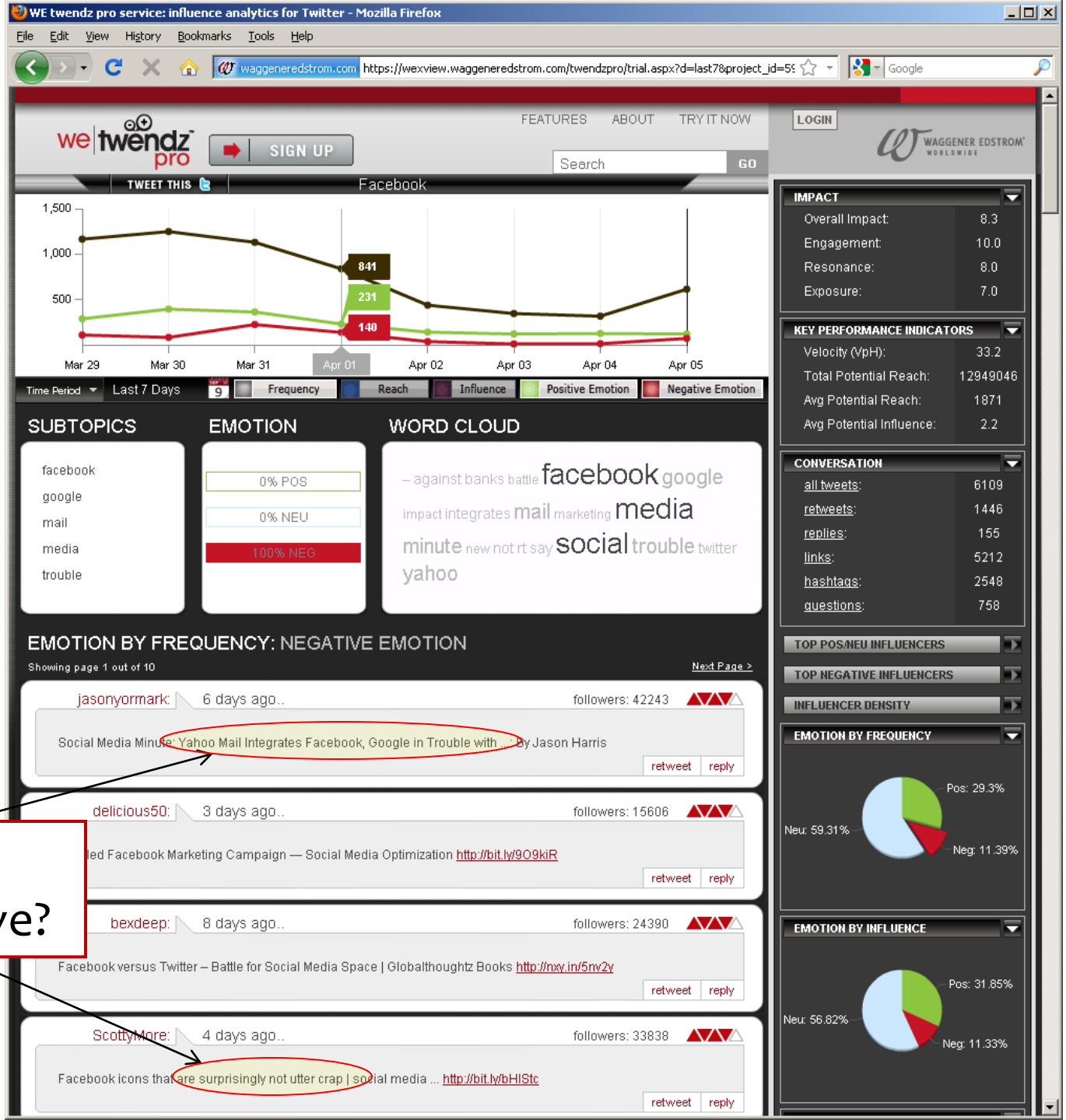
Sentiment Analysis

Applications include:

- Brand / Reputation Management.
- Competitive intelligence.
- Customer Experience Management.
- Enterprise Feedback Management.
- Quality improvement.
- Trend spotting.

Steps in the Right Direction





Rated negative?

Jodange | Top of Mind Service

Welcome, demo | Log out

Keyword: alternative energy (271 Total Opinions)

alternative energy Search

Home

Search results can be viewed in List, Chart or Heatmap views. Search Result Filters, available on the left, provide you with live filters to show you only the results you want to see. (hide)

List View | Opinion Index | Doppler View

Showing only opinions matching:
(remove all)

Topic: Alternative Energies, Urban Planning/Development

Opinion Holder: International Energy Agency, Boone Pickens, Energy Information Administration

Create an alert

Look Back: 1d 1w 1m 3m 6m

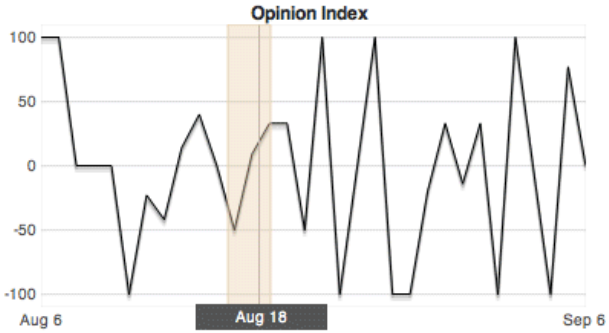
Opinion Holder (4440)
clear selected | show all

- [ ] Barack Obama (331)
- [ ] John McCain (305)
- [x] International Energy Agency (132)
- [ ] Nancy Pelosi (78)
- [ ] Sarah Palin (76)
- [x] Boone Pickens (65)
- [ ] Congress (62)
- [x] Energy Information Administration (53)
- [ ] Energy Department (47)
- [ ] George W Bush (44)
- [ ] BRITISH ENERGY GROUP PLC (33)
- [ ] International Atomic Energy Agency (33)
- [ ] Gordon Brown (31)
- [ ] Government (31)
- [ ] Reuters (31)
- [ ] Angelo Reyes (29)
- [ ] Department of Energy (29)
- [ ] Bloomberg (27)
- [ ] Vladimir Putin (27)
- [ ] U.S. Department of Energy (26)

Topic (3357)
clear selected | show all

- [x] Alternative Energies (8682)

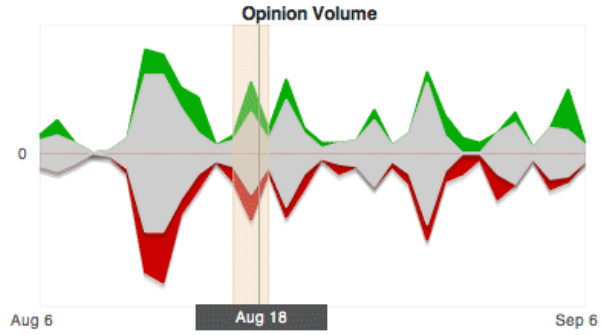
August 06, 2008 - September 06, 2008



Show Aggregate JOI

	JOI	9%		

Add Ticker Symbol Add Reset



Total	Pos	Neu	Neg
27	6	16	5

Opinion Summary
Aug 17 - Aug 19

By **Boone Pickens** in Herald-Sun on 2008-08-19

Since he formed Pickens Fuel Corp in 1997, Pickens has been arguing that gas is the best alternative fuel for cars.

Topics: Alternative Energies, Analyst Comment/Recommendation, National/Presidential Elections, Urban Planning/Development

By **International Energy Agency** in Energy Bulletin on 2008-08-19

After assessing all aspects of the situation, the International Energy Agency warned last week that while oil consumption in the US is expected to fall by 3.1 percent this year and 2 percent next year, Chinese oil consumption, while only a third that of the US, is expected to grow by about 5.6-5.7 percent in the next two years, thereby offsetting much of the drop in US consumption.

Topics: Alternative Energies, Climate Change

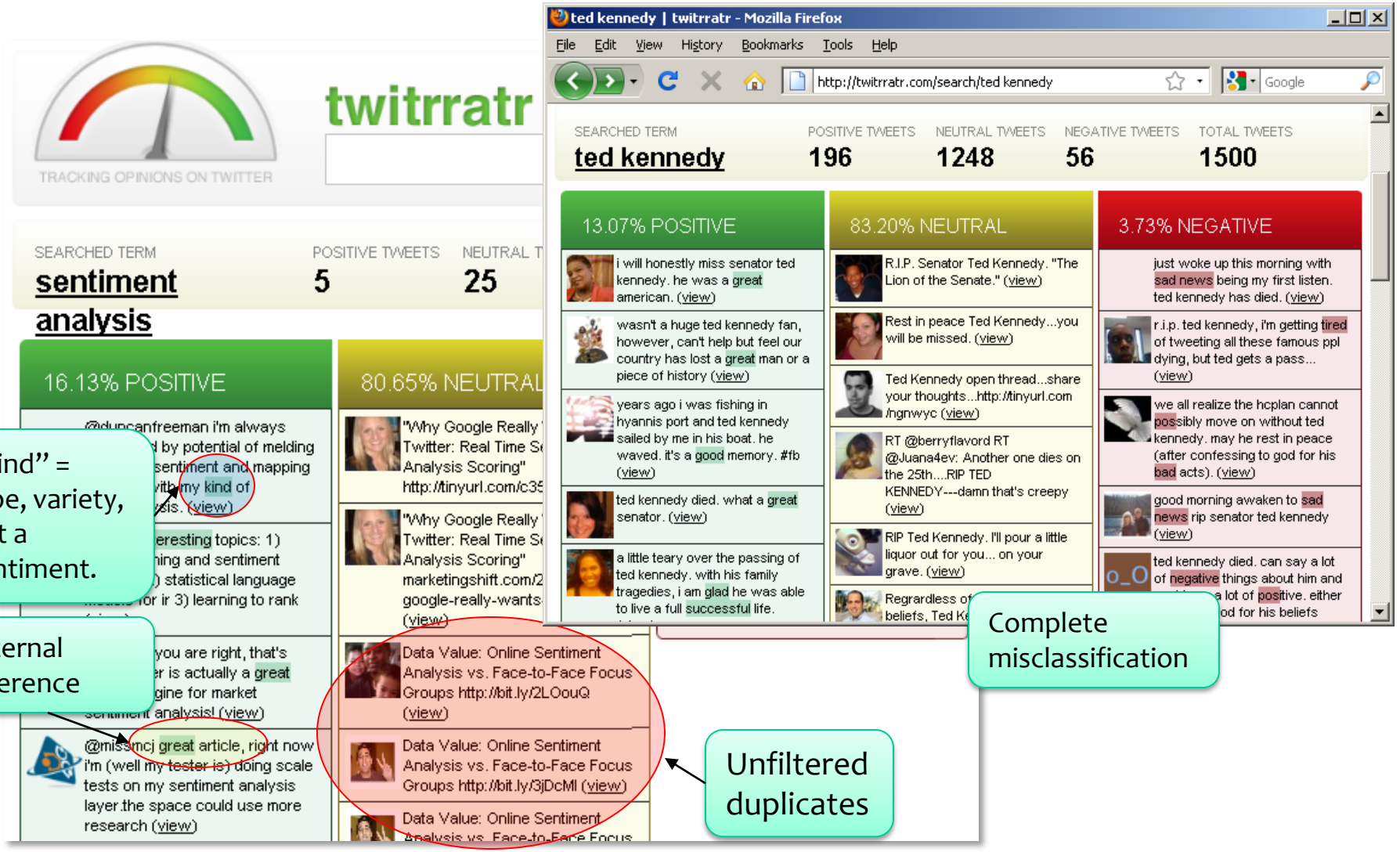
By **Energy Information Administration** in Pollution Online on 2008-08-19

In May 2008 the Energy Information Administration, which provides official energy statistics from the U S Government, stated that in 2004 the United States produced about 22 percent of global carbon dioxideemissions, primarily because the U Secnomy is the

Correlate the Opinion Index and Volumes to stock price changes

CLOSE

… and Missteps



Sentiment Complications

There are many complications.

- Sentiment may be of interest at multiple levels.
  - Corpus / data space, i.e., across multiple sources.
  - Document.
  - Statement / sentence.
  - Entity / topic / concept.
- Human language is noisy and chaotic!
  - Jargon, slang, irony, ambiguity, anaphora, polysemy, synonymy, etc.
  - Context is key. Discourse analysis comes into play.
- Must distinguish the sentiment holder from the object:
  - *"Geithner said the recession may worsen."*

Beyond Polarity

We present a system that adds an emotional dimension to an activity that Internet users engage in frequently, search."

-- Sood, Vasserman & Hoffman, 2009, "ESSE: Exploring Mood on the Web"

Happy	Sad	Angry
Energetic	Confused	Aggravated
Bouncy	Crappy	Angry
Happy	Crushed	Bitchy
Hyper	Depressed	Enraged
Cheerful	Distressed	Infuriated
Ecstatic	Envious	Irate
Excited	Gloomy	Pissed off
Jubilant	Guilty	
Giddy	Intimidated	
Giggly	Jealous	
Lonely		
Rejected		
Sad		
Scared		

The three prominent mood groups that emerged from K-Means Clustering on the set of* LiveJournal *mood labels.

Applications

Text analytics has applications in –

- Intelligence & law enforcement.
- Life sciences.
- Media & publishing including social-media analysis and contextual advertizing.
- Competitive intelligence.
- Voice of the Customer: CRM, product management & marketing.
- Legal, tax & regulatory (LTR) including compliance.
- Recruiting.

Online Commerce

Text analytics is applied for marketing, search optimization, competitive intelligence.

- Analyze social media and enterprise feedback to understand the Voice of the Market:
  - Opportunities
  - Threats
  - Trends
- Categorize product and service offerings for on-site search and faceted navigation and to enrich content delivery.
- Annotate pages to enhance Web-search findability, ranking.
- Scrape competitor sites for offers and pricing.
- Analyze social and news media for competitive information.

Voice of the Customer

Text analytics is applied to enhance customer service and satisfaction.

Analyze customer interactions and opinions –

- E-mail, contact-center notes, survey responses
- Forum & blog posting and other social media

– to –

- Address customer product & service issues
- Improve quality
- Manage brand & reputation

If you can link qualitative information from text you can –

- Link feedback to transactions
- Assess customer value
- Understand root causes
- Mine data for measures such as churn likelihood

E-Discovery and Compliance

Text analytics is applied for compliance, fraud and risk, and e-discovery.

Regulatory mandates and corporate practices dictate –

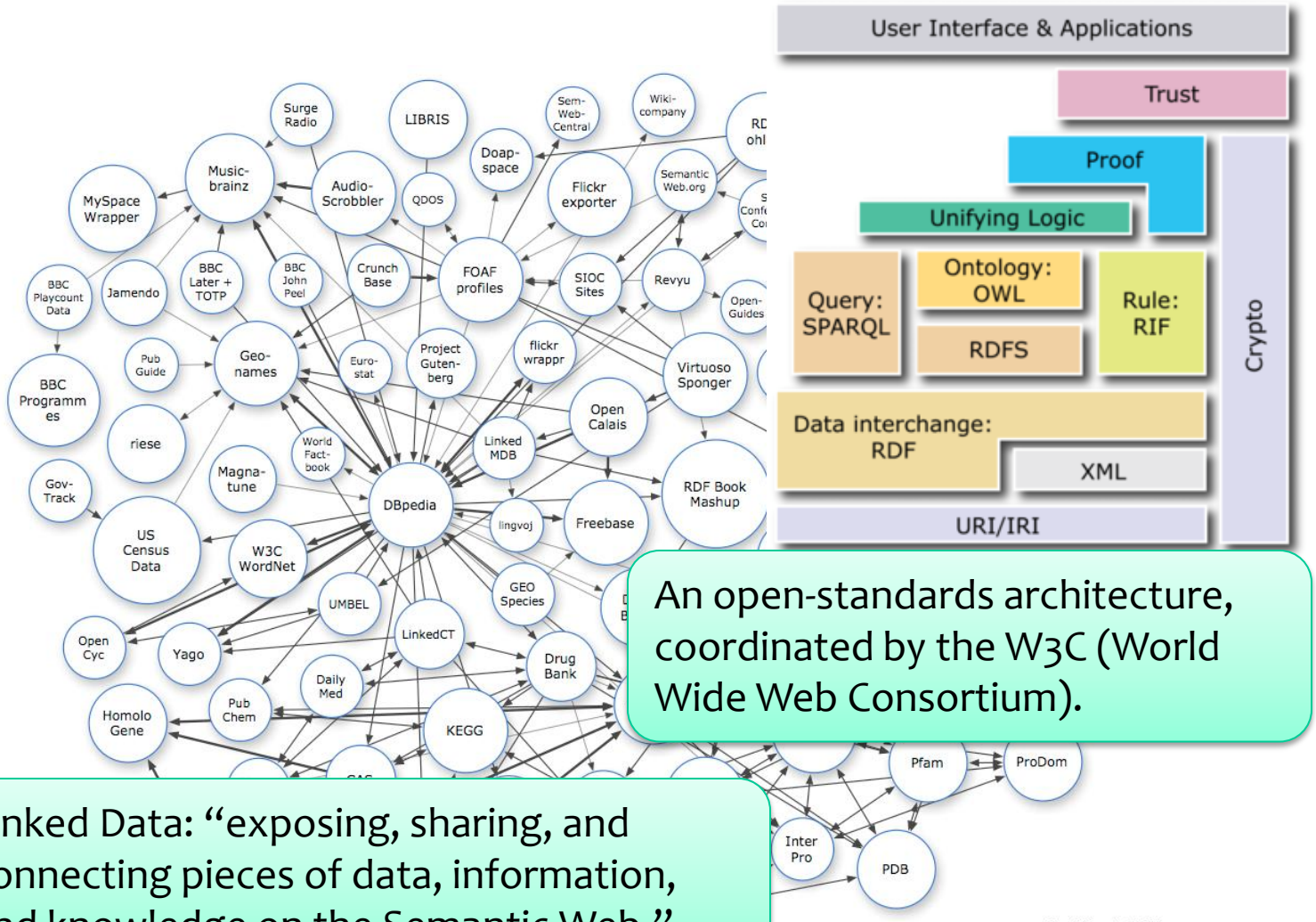
- Monitoring corporate communications
- Managing electronic stored information for production in event of litigation

Sources include e-mail (!!), news, social media

Risk avoidance and fraud detection are key to effective decision making

- Text analytics mines critical data from unstructured sources
- Integrated text-transactional analytics provides rich insights

The Semantic Web vision



An open-standards architecture, coordinated by the W3C (World Wide Web Consortium).

Linked Data: "exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web."

Getting Started

A best practices approach…

Assess:

- Assess business goals.
- Understand information sources.
- Consult and educate stakeholders.

Evaluate:

- Evaluate installed, hosted/SaaS, database-integrated options.
- Determine performance and business requirements.
- Match methods to goals, sources, and work practices.

Implement:

- Start with basic functions such as search, modest goals, or with a single information source.
- Go for clear wins to gain support.
- Build out applications, capacity, BI/research integration.

Users' Perspectives

I estimate a $425 million global market in 2009, up from $350 in 2008. I foresee 25% growth in 2010.

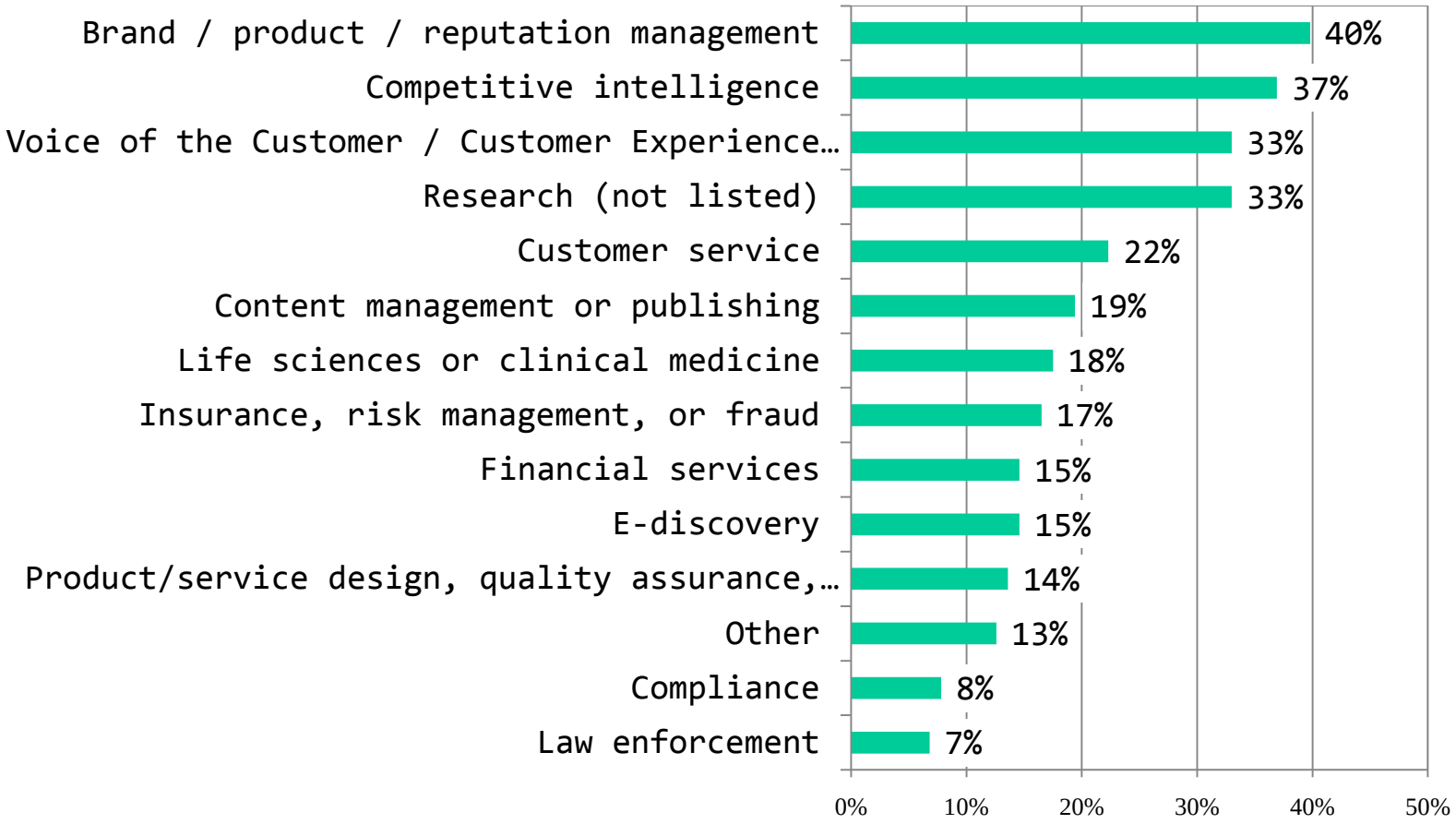
Last year, I published a study report, "Text Analytics 2009: User Perspectives on Solutions and Providers."

http://www.slideshare.net/SethGrimes/text-analytics-2009-user-perspectives-on-solutions-and-providers

I relayed findings from a survey that asked…

Primary Applications

What are your primary applications where text comes into play?



Analyzed Textual Information

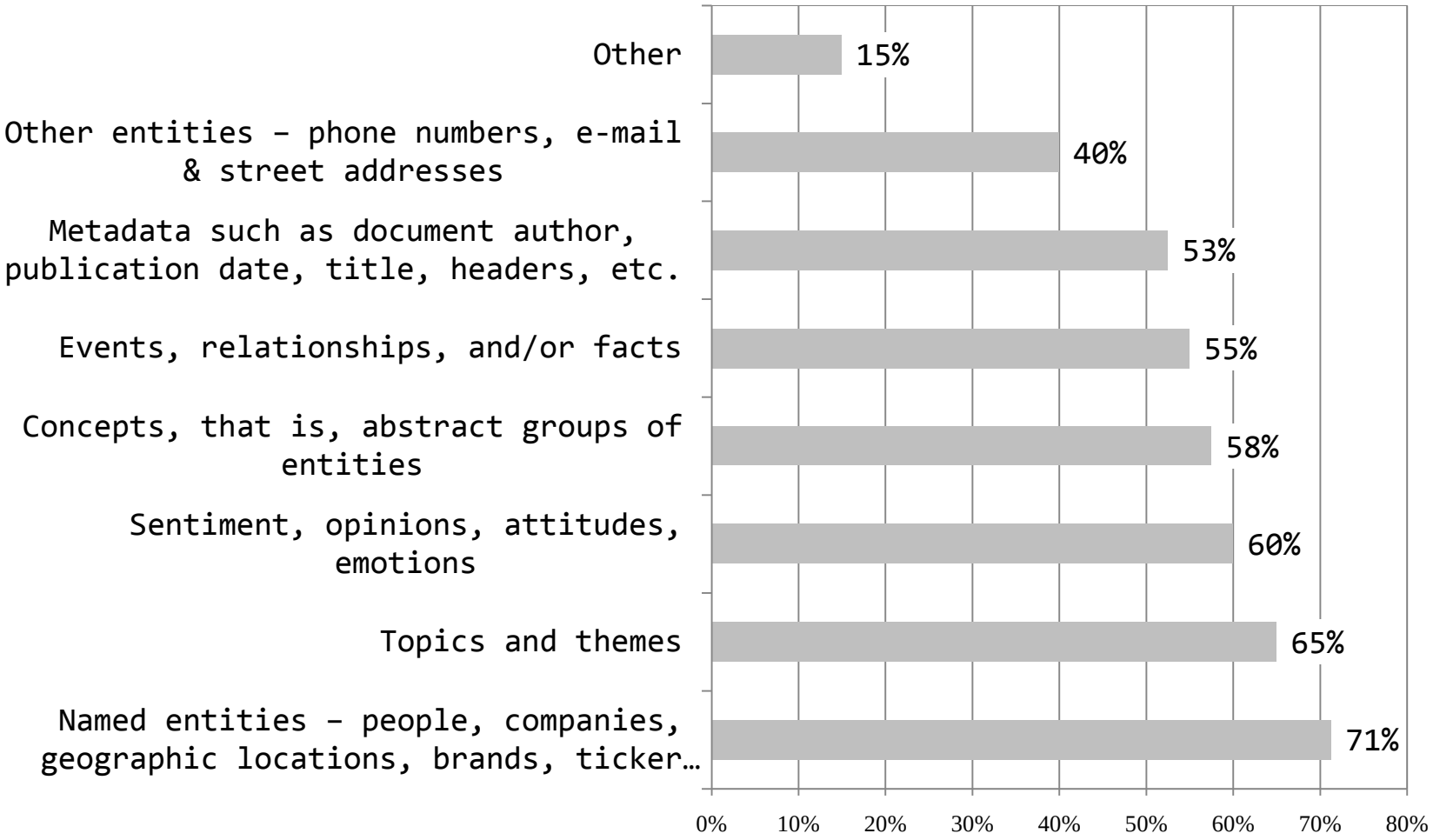
What textual information are you analyzing or do you plan to analyze?

Current users responded:

blogs and other social media (twitter, social-network sites, etc.)	62%
news articles	55%
on-line forums	41%
e-mail and correspondence	38%
customer/market surveys	35%

Extracted Information

Do you need (or expect to need) to extract or analyze:



Software & Platform Options

Text-analytics options may be grouped in general classes.

- Installed text-analysis application, whether desktop or server or deployed in-database.
- Data mining workbench.
- Hosted.
- Programming tool.
- As-a-service, via an application programming interface (API).
- Code library or component of a business/vertical application, for instance for CRM, e-discovery, search.

The slides that follow next will present leading options in each category except Hosted…

Text Analysis Applications

Vendors:

Attensity, Clarabridge, IBM Cognos, Linguamatics, Provalis Research, Nstein (Open Text), SAP, SAS Teragram, SRA NetOwl, TEMIS Luxid.

Typical uses:

Customer experience management (CEM), survey analysis, social-media analysis, law enforcement.

Typical characteristics:

Interface that allows the user to configure a processing pipeline.

Interface for text exploration and visualization.

Export to databases

Data Mining Workbench

Vendors:

IBM SPSS Modeler, Megaputer PolyAnalyst, Rapid-I RapidMiner, SAS Text Miner.

Typical uses:

Customer experience management (CEM), marketing analytics, survey analysis, social-media analysis, law enforcement.

Predictive modeling.

Typical characteristics:

Same as text-analysis applications, but with more sophisticated modeling and analysis capabilities.

Programming/Development Tool

Vendors:

- GATE, Python NLTK, R – open source.
- NooJ – free, non-open source.
- IBM LanguageWare.

Typical uses:

- Language modeling.
- Data exploration.
- Up to the programmer.

Typical characteristics:

- Text is an add-in to a programming language/environment.

As a Service, API

Vendors:

- Lexalytics, Open Amplify, Orchest8 Alchemy API, Thomson Reuters Open Calais.
- GeeYee, Jodange, Sentimentrix

Typical uses:

- Annotation and content enrichment with the application domain up to the user.

Typical characteristics:

- Relies on remote, server-resident processing resources.
- May or, more likely, may not be end-user customizable.

Code Library or Application Component

Vendors:

Alias-I LingPipe, GATE.

Basis Technology Rosette, SAP Inxight, SAS Teragram, TEMIS.

Typical uses:

Information extraction in support of business applications.

Typical characteristics:

Same as text-analysis applications, but with more sophisticated modeling and analysis capabilities.

Text Visualizations

There are many text-visualization options.

We've seen –

- Structured search-generated information, slide 24.
- Clustered search results, slide 25.
- Document-set exploration via tabular & graphically rendered facets, slide 27.
- Mined feature-relationship network, slides 27, 51.
- Annotated documents, slides 37, 54.
- Word cloud, slide 38.
- Trend lines, extracted terms, topics, and sentiment, slides 45, 68, 69.
- Sentence diagrams with parts of speech, slides 48, 49.

Text Visualizations

Text visualizations operate at –

- The document level, visualizations *in-situ*.
- The document-set level, organizing documents based on metadata and/or extracted information.
- At the feature level, organizing extracted information.

Examples to this point have been –

- Freely usable on the Web.
- Part of products or solutions.

IBM's Many Eyes is a great, free resource, supporting text viz types:

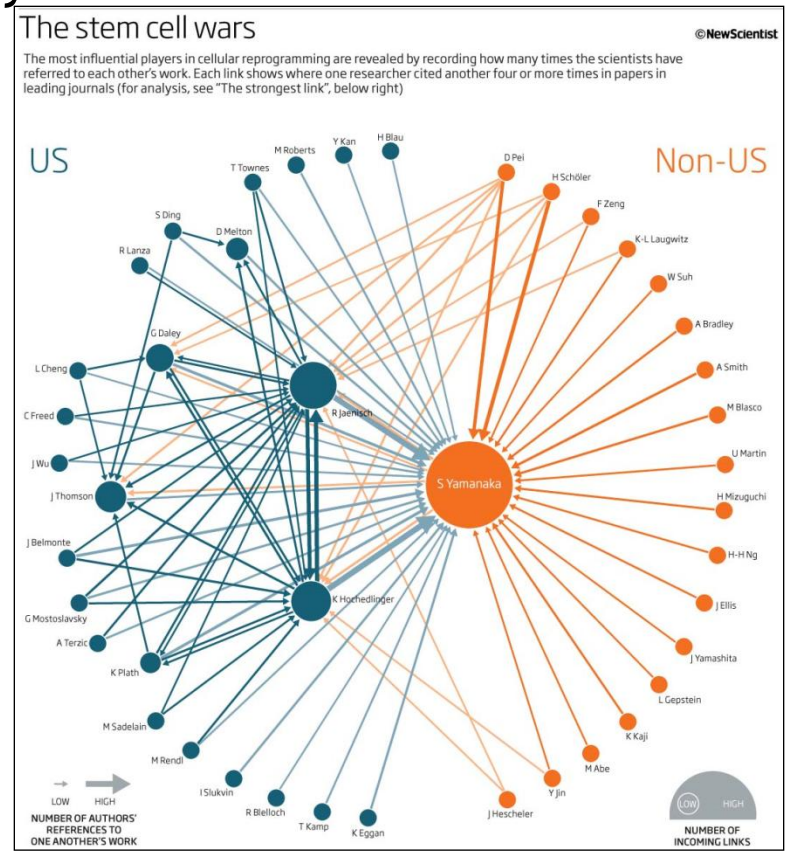
- Word Tree, Tag Cloud, Phrase Net, Word Cloud Generator

The illustrations that follow complement the ones we've seen…

Visualization: Relationship Network

"Using a program called Sitkis, we extracted data on citations from the Web of Science… To map the network of citations, we imported the data into a Microsoft Excel spreadsheet running a social-network analysis extension called NodeXL."

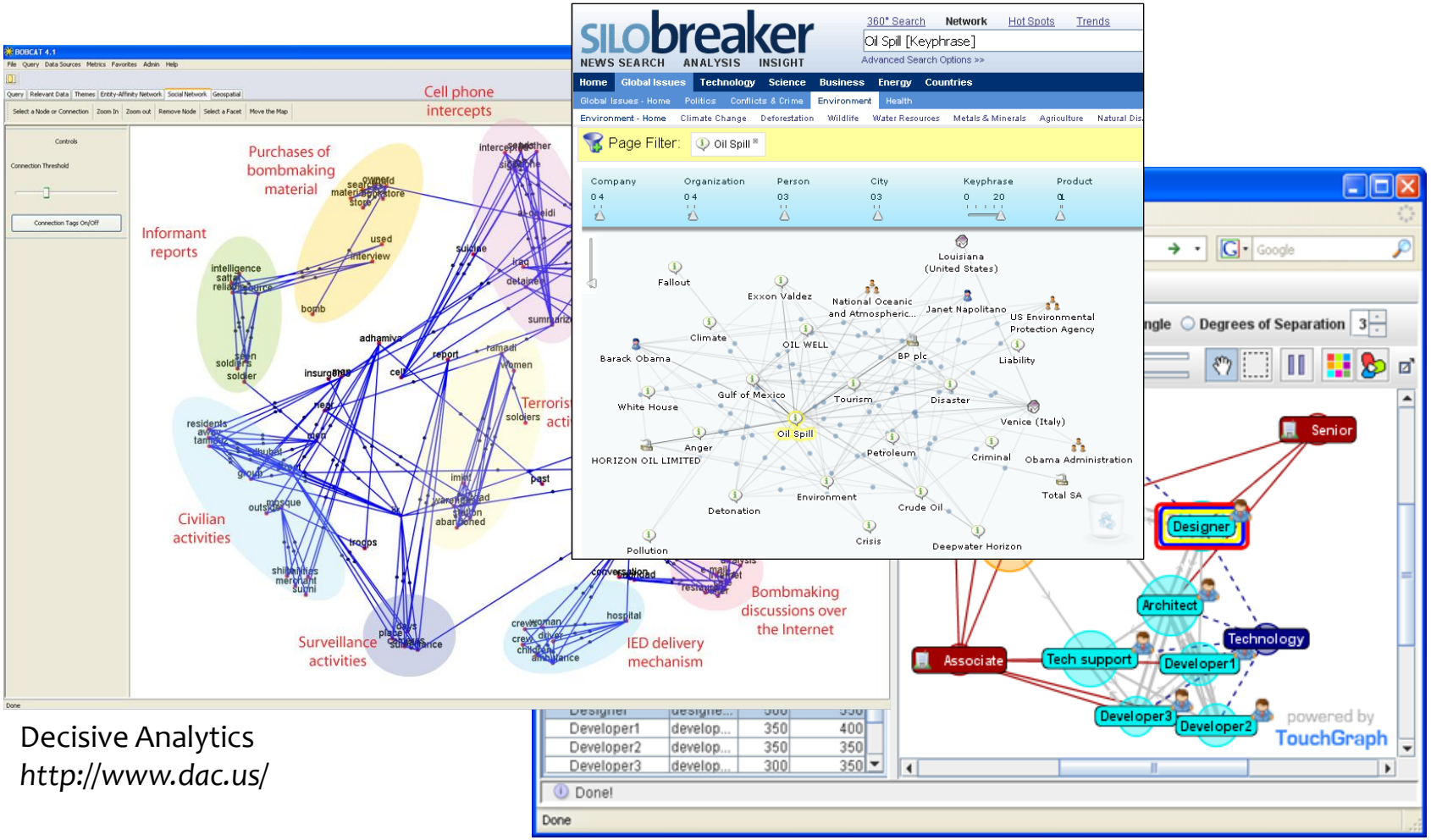
-- Peter Aldhous, "The stem cell wars: data, methods and results," New Scientist, 2 June 2010.



http://www.newscientist.com/article/dn18996-the-stem-cell-wars-data-methods-and-results.html

Visualization: Content Network

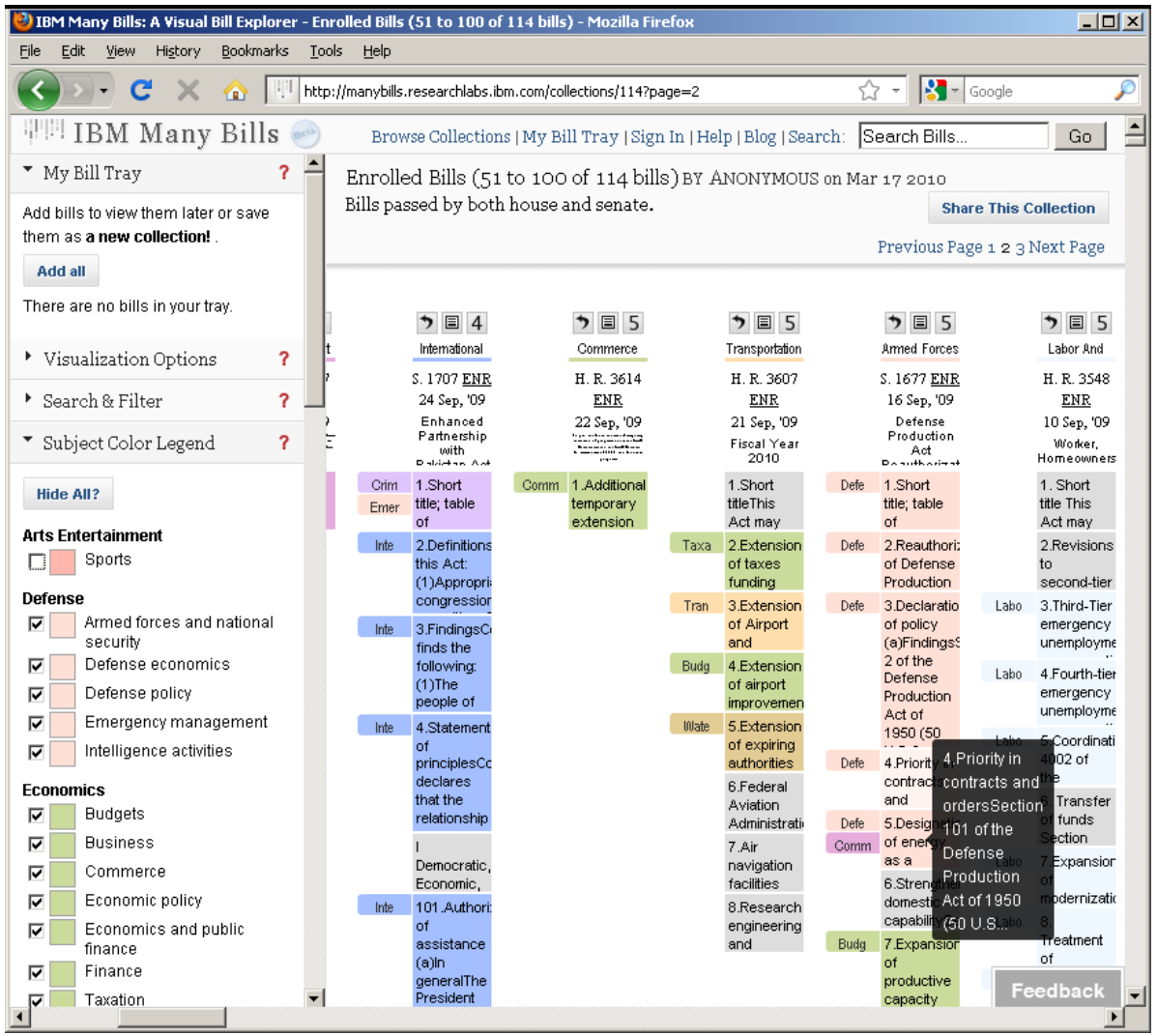
Operating on text-extracted features & themes…



Decisive Analytics
http://www.dac.us/

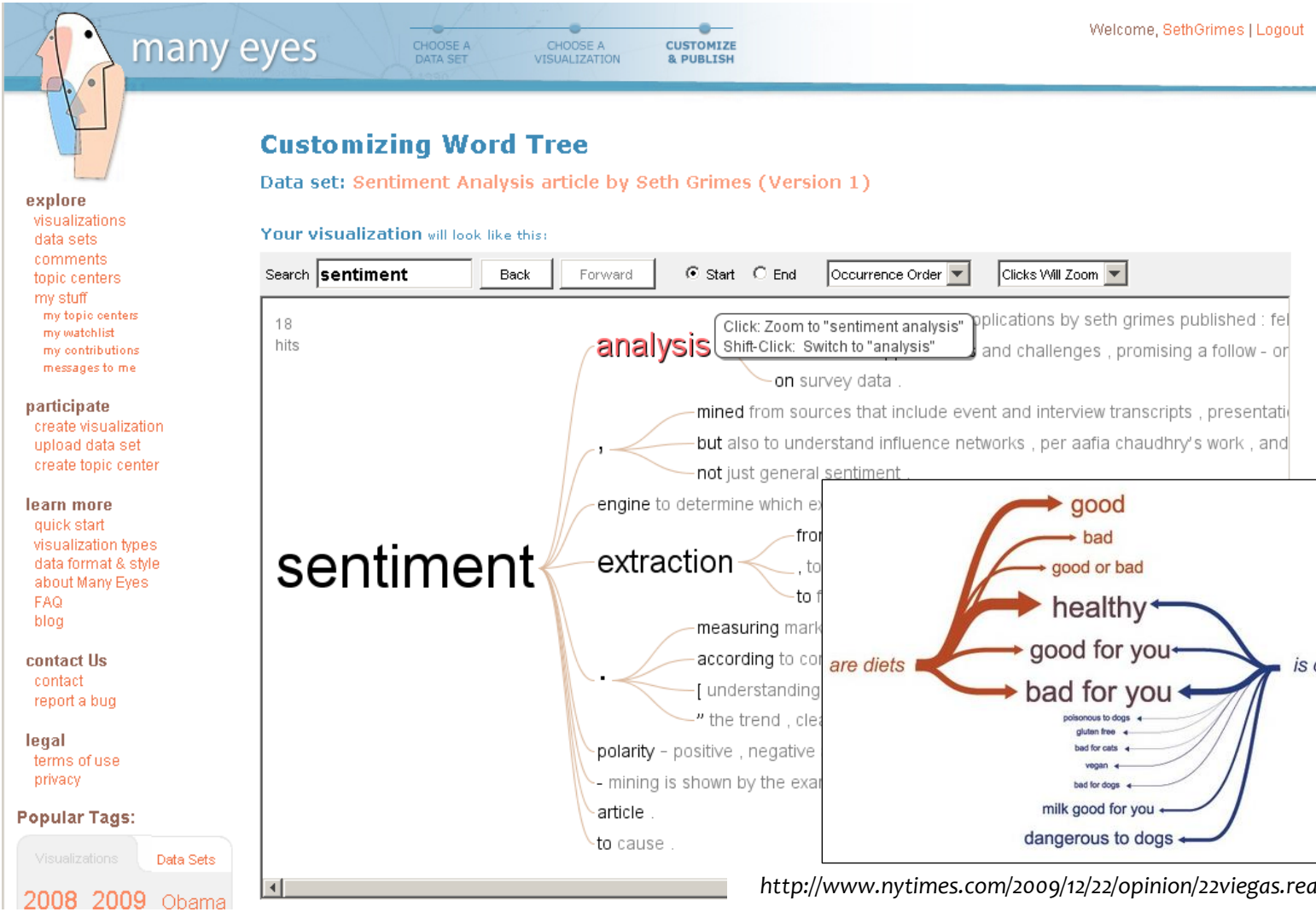
TouchGraph Navigator (commercial),
http://www.touchgraph.com/navigator.html

Visualization: Document Set Categorization



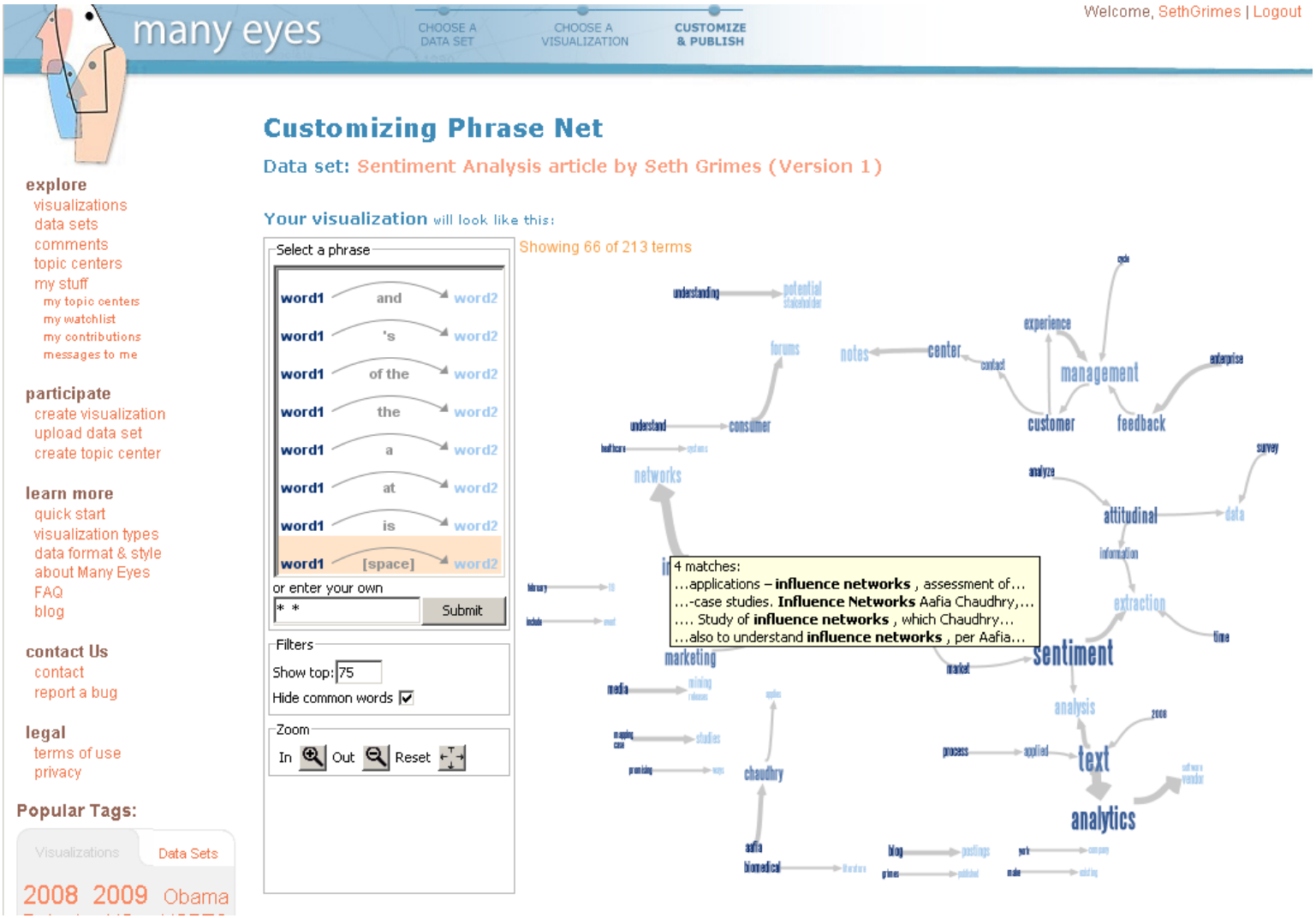
http://manybills.researchlabs.ibm.com/collections/114

Visualization: Word Tree

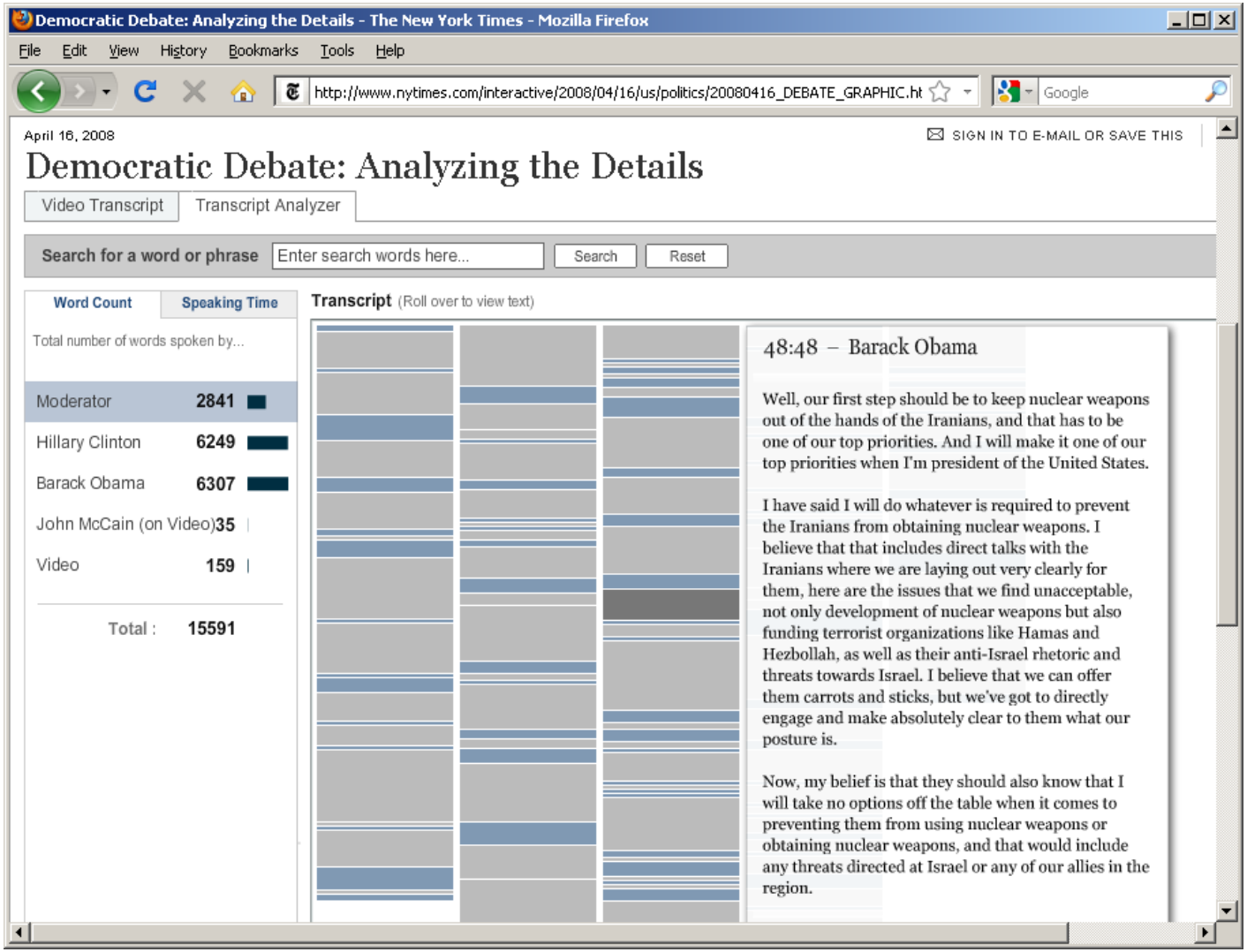


http://www.nytimes.com/2009/12/22/opinion/22viegas.ready.html

Visualization: Phrase Net

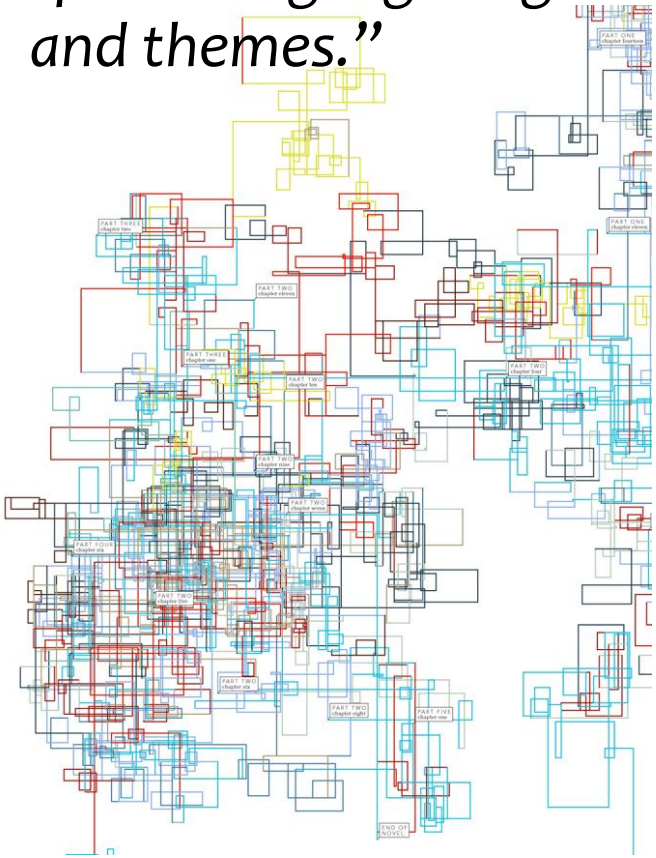


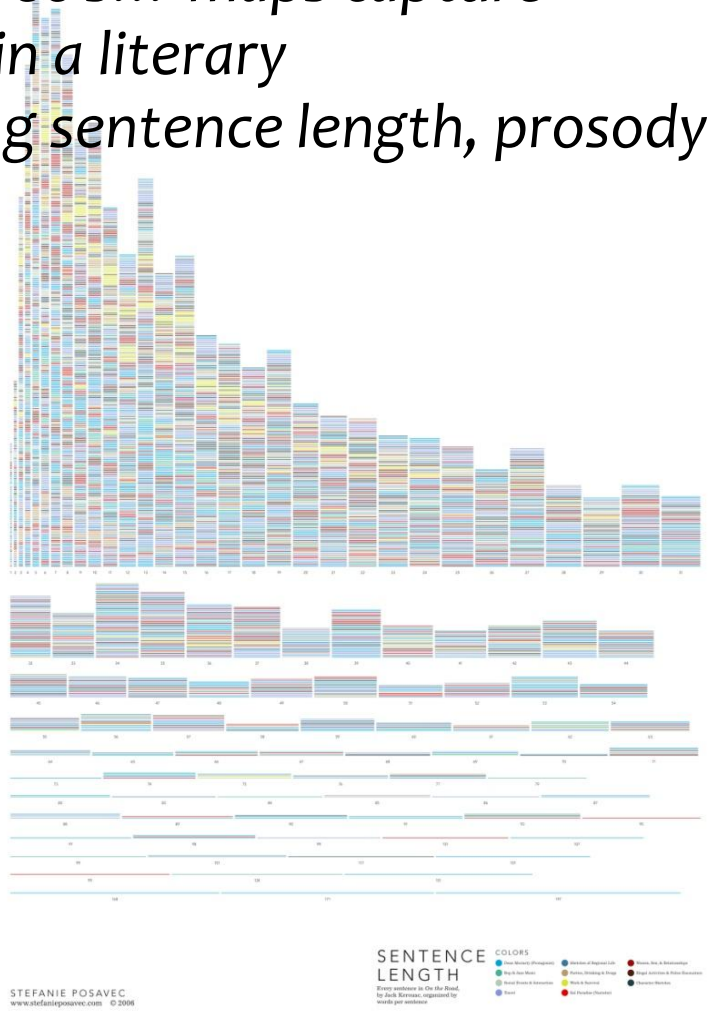
Visualization: Discourse Analysis



Visualization: Text Statistics

Literary analysis: "Stefanie Posavec's… maps capture regularities and patterns within a literary space… highlighting and noting sentence length, prosody and themes."

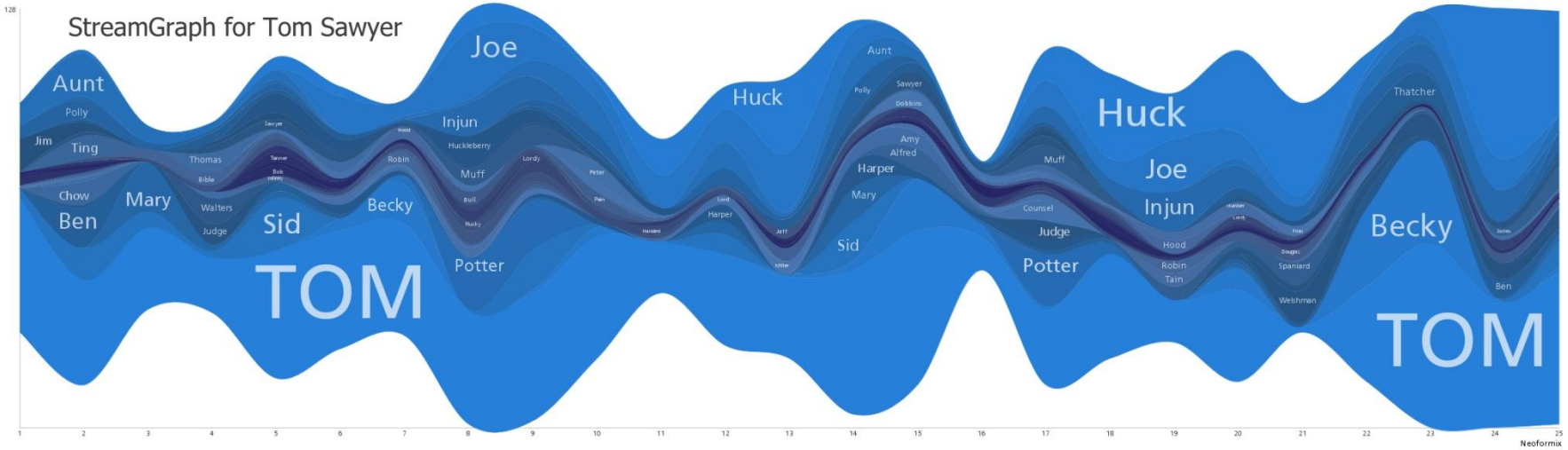




http://www.notcot.com/archives/2008/04/stefanie-posave.php

Visualization: Stream Graph

Acknowledges narrative structure.



http://www.neoformix.com/2008/TomSawyer.html

Selected Resources

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, 2008.

http://nlp.stanford.edu/IR-book/information-retrieval-book.html

Bo Pang and Lillian Lee, "Opinion mining and sentiment analysis," 2008.

http://www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html

Tim Showers, "Visualization Strategies: Text & Documents," 2008.

http://www.timshowers.com/2008/08/visualization-strategies-text-documents/

IBM Many Eyes visualization site.

http://manyeyes.alphaworks.ibm.com/manyeyes/

Neoformix: discovering & illustrating patterns in data.

http://neoformix.com/

Seth Grimes, various material.

http://www.slideshare.net/SethGrimes/, http://sethgrimes.com

From Pattern Matching to Knowledge Discovery Using ***Text Mining and Visualization*** Techniques

Seth Grimes

Alta Plana Corporation

@sethgrimes – 301-270-0795 -- *http://altaplana.com*

Special Libraries Association 2010

June 13, 2010