

# 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](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)

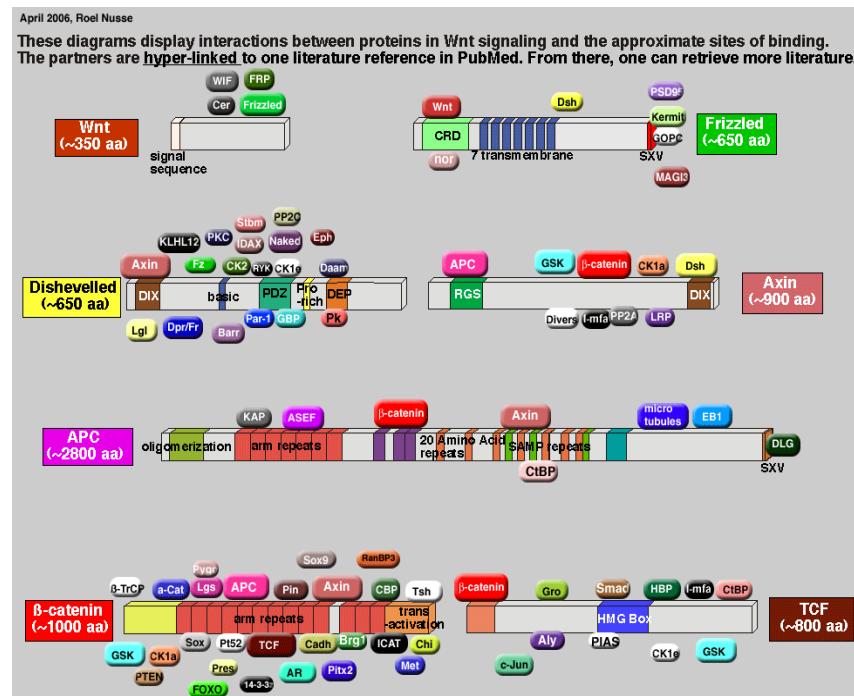


# Unstructured Sources

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

[www.stanford.edu/~ernusse/wntwindow.html](http://www.stanford.edu/~ernusse/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:

## Hedging With Options

**Example:** You expect to receive 100,000 CAD in 3 months and want to lock in a minimum

rate at which to sell CAD against USD. You buy a CAD put:

Current Spot Rate USD/CAD: 1.3700

Strike Price: 1.3761

Maturity: 3 months

Style: European

Premium: 1.22%

This option gives you the right, but not the obligation, to sell CAD at 1.3761 at maturity.

Your cost for this option is USD \$886.56

Scenarios at Maturity with an option hedge:

**CAD appreciates:**  $USD/CAD = 1.2500$

You choose not to exercise your option because you can sell your USD/CAD at the prevailing market rate. Net of the premium you receive is \$79,113.44.

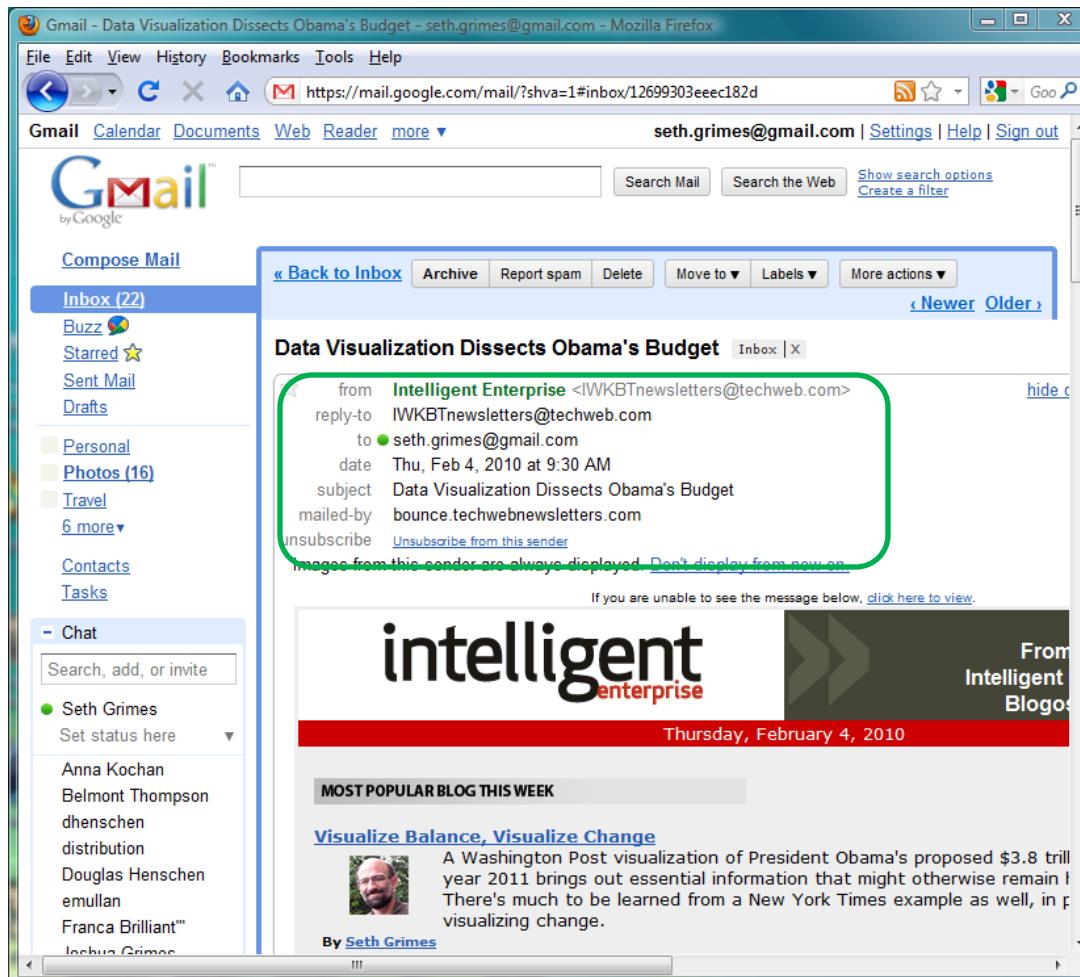
**CAD depreciates:**  $USD/CAD = 1.4900$

You choose to exercise your option and sell your CAD at 1.3761, receiving \$72,669.14 versus the prevailing market rate where you would only receive \$67,114.09. Net of the premium you receive is \$71,782.58.



# Unstructured Sources

Sources contain/provide metadata:



# Unstructured Sources

Sources may intermix content and “noise”:

The screenshot shows a Mozilla Firefox browser window displaying the Yahoo! Finance page for Oracle Corporation (ORCL). The page includes a top navigation bar with links like File, Edit, View, History, Bookmarks, Tools, Help, New User? Register, Sign In, and Help. Below the bar is a search field and a "Web Search" button. A red circle highlights the top navigation area. The main content area features a stock quote for Oracle Corp. (ORCL) at \$22.13, down 0.71 (3.11%). Another red circle highlights this quote. To the left, a sidebar menu is circled in green, containing links for Quotes Summary, Real-Time Options, Historical Prices, Charts Interactive, Basic Chart, Basic Tech. Analysis, News & Info Headlines, Financial Blogs, Company Events, Message Boards, Company Profile, Key Statistics, SEC Filings, Competitors, Industry Components, and Analyst Coverage. A red circle highlights the SEC Filings link in the sidebar. The central content area has a green border around the "SEC Filings" section, which lists recent filings with details like date, form type, and title. A red circle highlights this section. At the bottom right, there is an advertisement for the 2010 Taurus car, featuring a dark background and the text "THE NEW 2010 TAURUS CATCH-ME IF-YOU-CAN 2010 TAURUS SHO vs 2009 AUDI A6 ROLL OVER FOR VIDEO". A red circle highlights this advertisement.

Dow **↓ 3.15%** Nasdaq **↓ 3.64%**

**YAHOO! FINANCE**

**Oracle Corp. (ORCL)** On Jun 4: **22.13 ↓ 0.71 (3.11%)**

**SEC Filings**

The ORCL Annual Report is now available. [Free Annual Report](#)

**RECENT FILINGS**

Date	Form	Title
4-Jun-10	8-K	Costs Associated with Exit or Disposal Activities <a href="#">Summary</a> - <a href="#">Full Filing at EDGAR Online (16kb)</a>
14-May-10	8-K	Creation of a Direct Financial Obligation or an Obligation under an Off-Balance Sheet <a href="#">Summary</a> - <a href="#">Full Filing at EDGAR Online (15kb)</a>
7-Apr-10	8-K/A	Financial Statements and Exhibits <a href="#">Summary</a> - <a href="#">Full Filing at EDGAR Online (2mb)</a>
29-Mar-10	10-Q	Quarterly Report <a href="#">Summary</a> - <a href="#">Full Filing at EDGAR Online (3mb)</a>
25-Mar-10	8-K	Results of Operations and Financial Condition, Other Events, Financial Statements and <a href="#">Summary</a> - <a href="#">Full Filing at EDGAR Online (924kb)</a>
12-Feb-10	8-K	Change in Directors or Principal Officers <a href="#">Summary</a> - <a href="#">Full Filing at EDGAR Online (13kb)</a>
28-Jan-10	8-K	Completion of Acquisition or Disposition of Assets, Other Events, Financial Statement

ADVERTISEMENT

**THE NEW 2010 TAURUS**  
CATCH-ME IF-YOU-CAN  
2010 TAURUS SHO vs 2009 AUDI A6  
ROLL OVER FOR VIDEO

# Unstructured Sources

“Unstructured” materials likely contain structure:

**Table of Contents**

**ORACLE CORPORATION**  
NOTES TO CONDENSED CONSOLIDATED FINANCIAL STATEMENTS  
February 28, 2010  
(Unaudited)

three-month time period from the date we have purchased goods or services from that same customer are revised and disclosure. When we acquire goods or services from a customer, we negotiate the purchase separately from consideration to be at arm's length, and settle the purchase in cash. We recognize new software license revenues from Concurrent Transactions if all of our revenue recognition criteria are met and the goods and services are delivered.

**Inventories**  
Inventories are stated at the lower of cost (first in, first out) or market value. We evaluate our ending inventories for obsolescence. This evaluation includes analysis of sales levels by product and projections of future demand (less than one month or less). Inventories in excess of future demand are written down. In addition, we assess the impact of and write-off inventories that are considered obsolete.

**Shipping Costs**  
Our shipping and handling costs for hardware systems products sales are included in expenses for hardware presented.

**Acquisition Related and Other Expenses**  
Acquisition related and other expenses consist of personnel related costs for transitional and certain other employee expenses, integration related professional services, certain business combination adjustments after the measurement period has ended, and certain other operating expenses, net. Stock-based compensation included in acquisition from unvested options or restricted stock-based awards (primarily consisting of restricted stock units) assumed accelerated upon termination of the employees pursuant to the original terms of those options or restricted stock adoption of the FASB's revised accounting guidance for business combinations as of the beginning of fiscal 2009. Expenses are now recorded as expenses in our statements of operations that would previously have been included and capitalized as a part of the accounting for our acquisitions pursuant to previous accounting rules as professional services fees.

(in millions)	Three Months Ended February 28,		
	2010	2009	2008
Transitional and other employee related costs	\$ 5	\$ 3	\$ 10
Stock-based compensation	10	3	14
Professional fees and other, net	18	16	25
Business combination adjustments, net	1	—	5
<b>Total acquisition related and other expenses</b>	<b>\$ 34</b>	<b>\$ 27</b>	<b>\$ 50</b>
	<b>\$ 98</b>		

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<sup>1</sup>, Huidong Yuan, Carole D. Weaver<sup>2</sup>, Junhao Mao, Gist H.Farr III<sup>2</sup>, Daniel J.Sussman<sup>3</sup>, Jos Jonkers<sup>4</sup>, David Kimelman<sup>2</sup> and Dianqing Wu<sup>5</sup>**

<sup>1</sup>Department of Pharmacology and Physiology, University of Rochester, NY 14642, <sup>2</sup>Department of Biochemistry, University of Washington, Seattle, WA 98195-7350, <sup>3</sup>Department of Obstetrics, Gynecology and Reproductive Sciences, University of Maryland, Baltimore, MD, USA, <sup>4</sup>Shanghai Institute of Biochemistry, The Chinese Academy of Sciences, Shanghai, Peoples Republic of China and <sup>5</sup>Division of Molecular Genetics, The Netherlands Cancer Institute, Amsterdam, The Netherlands

<sup>3</sup>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 Frat1 may be important for this signaling pathway. Furthermore, *Xenopus* Axin and Frat1 interact with Dvl and GSK in a similar manner, suggesting that the Wnt signaling pathway may be conserved in vertebrates.

**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 (Drl), 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

Alta Plana

2010 Special Libraries Association

# *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.”*

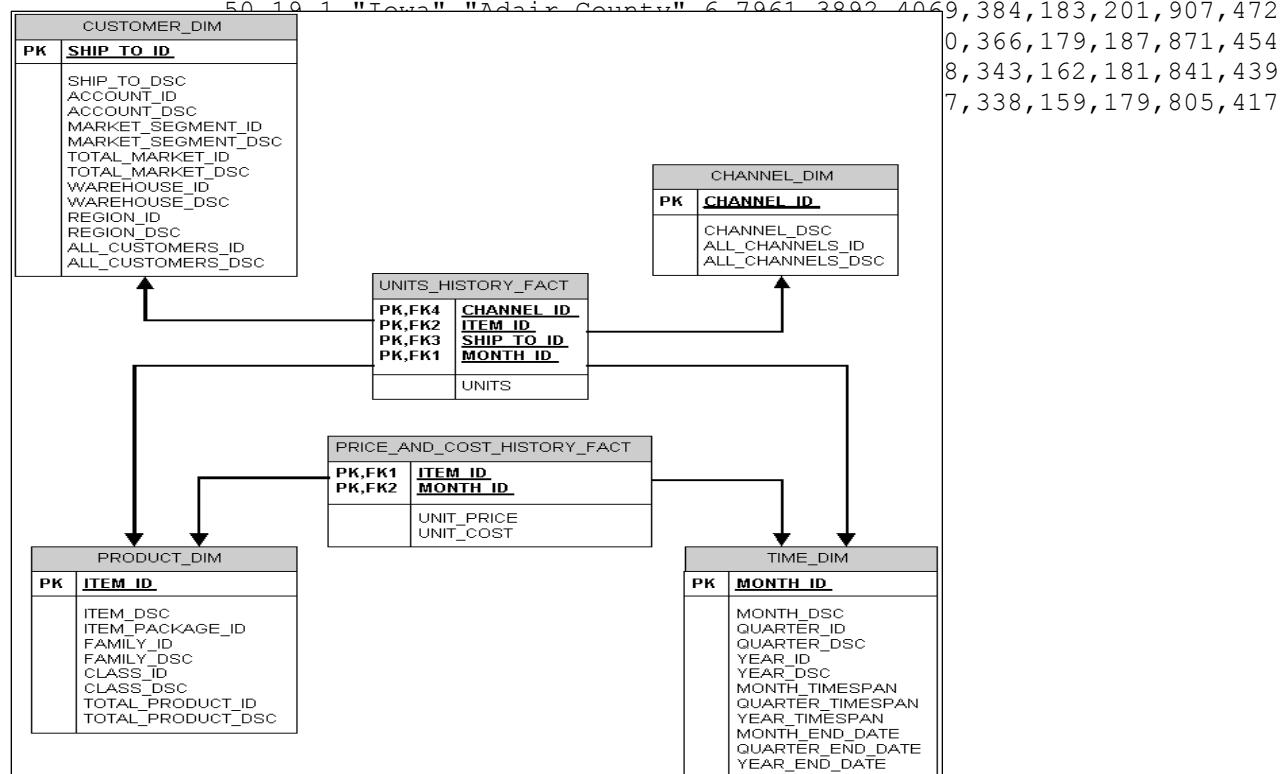
-- Prabhakar Raghavan, Yahoo Research

# Business Intelligence

Conventional BI feeds off:

```
"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",6,7961,3902,4069,384,183,201,907,472
0,366,179,187,871,454
8,343,162,181,841,439
7,338,159,179,805,417
```

It runs off:



# Business Intelligence

Conventional BI produces:

The screenshot displays the Pentaho Business Intelligence Platform Portal Demo interface. At the top, there's a navigation bar with links like Home, Getting Started, Reporting, Business Rules, Printing, Bursting, Widgets, DataSource, Secure, and Advanced. Below the navigation bar, there are several sections:

- Filters:** A section where users can apply filters to other controls on the page. It includes dropdowns for REGION (Central) and DEPARTMENT (Executive Management), and a button for Update.
- Headcount Data:** A table showing headcount details by position. The data is as follows:

Position	Actual	Budget	Variance
SVP Strategic Development	\$383,242	\$403,405	\$20,163
SVP Partnerships	\$367,415	\$392,100	\$24,685
CEO	\$549,625	\$522,250	-\$27,375
SVP WW Operations	\$476,000	\$725,887	\$249,887
Total	\$1,776,282	\$2,043,642	\$267,360

- Headcount Costs:** A pie chart showing headcount costs across four categories: CEO (\$46,025), SVP WW Operations (\$476,000), SVP Strategic Development (\$383,242), and SVP Partnerships (\$367,415).
- Actual Headcount - % Variance from Budget:** Four donut charts showing the percentage variance from budget for CEO (-3.1%), SVP Partnerships (-2.9%), SVP Strategic Development (-1.13%), and SVP WW Operations (-0.68%).
- Actual Headcount - % Variance from Budget:** A detailed table showing actual headcount, budget, variance, and variance percent for various regions and departments. The data is as follows:

Region	Department	Positions	Measures			
			Actual	Budget	Variance	Variance Percent
<b>-All Regions</b>	<b>All Departments</b>	+ All Positions	143,639,982.00	143,199,389.00	-440,593.00	- .31%
	<b>Executive Management</b>	+ All Positions	8,299,022.00	6,494,168.00	195,144.00	3.00%
	<b>Finance</b>	+ All Positions	12,224,220.00	12,087,406.00	-136,814.00	-1.13%
	<b>Human Resource</b>	+ All Positions	13,075,463.00	12,989,341.00	-86,122.00	- .68%
	<b>Marketing &amp; Communication</b>	+ All Positions	13,910,753.00	13,770,267.00	-140,486.00	-1.02%
	<b>Product Development</b>	+ All Positions	10,644,102.00	10,786,611.00	142,509.00	1.32%
	<b>Professional Services</b>	+ All Positions	76,317,649.00	76,098,206.00	-219,443.00	- .29%
	<b>Sales</b>	+ All Positions	11,168,773.00	10,973,392.00	-195,381.00	-1.78%
<b>Central</b>	+ All Departments	+ All Positions	37,893,162.00	38,397,600.00	504,438.00	1.31%
<b>Eastern</b>	+ All Departments	+ All Positions	35,248,940.00	35,487,861.00	238,921.00	.67%
<b>Southern</b>	+ All Departments	+ All Positions	35,248,940.00	34,803,861.00	-445,079.00	-1.28%
<b>Western</b>	+ All Departments	+ All Positions	35,248,940.00	34,510,067.00	-738,873.00	-2.14%

## Text-Bl: 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

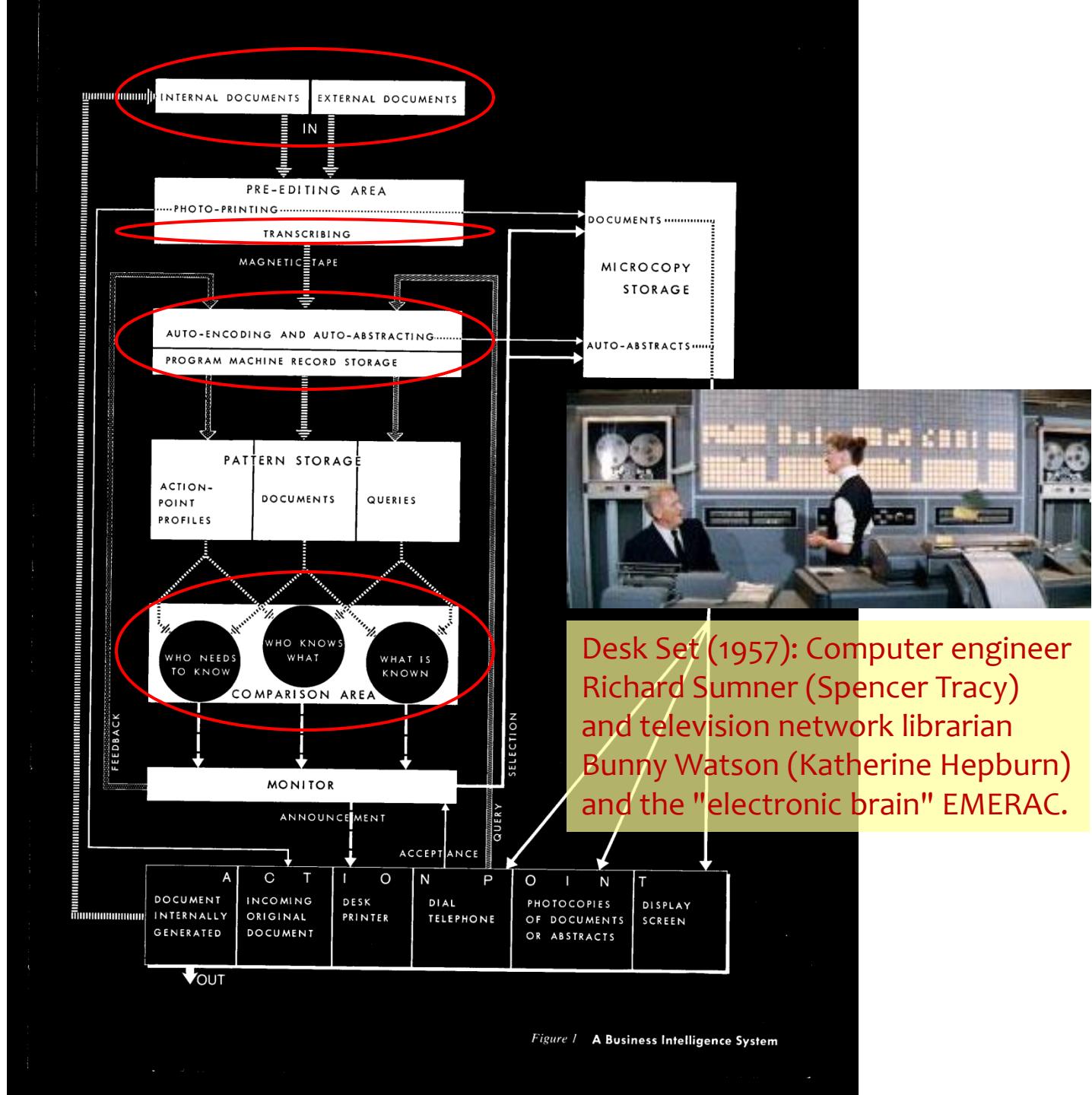
“A Business Intelligence System”

IBM Journal, October 1958

What was IT like in the ‘50s?

# Document input and processing

Knowledge handling is key



## *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



The screenshot shows a Google search results page in Mozilla Firefox. The search query is "how did the dow perform yesterday". The results page displays 10 results out of approximately 159,000. The first result is from "Channel NewsAsia :: View topic - Yesterday dow down 3 digit ...". The second result is from "Dow Drops By 26.36, To 1959.05 - New York Times". The third result is from "DOW GREETS '87 WITH 31.36 SPURT - New York Times". On the right side of the results, there are "Sponsored Links" for "Take Stock of the Market" and "Stock". The bottom of the screen shows the Firefox status bar with "Done".

# 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.

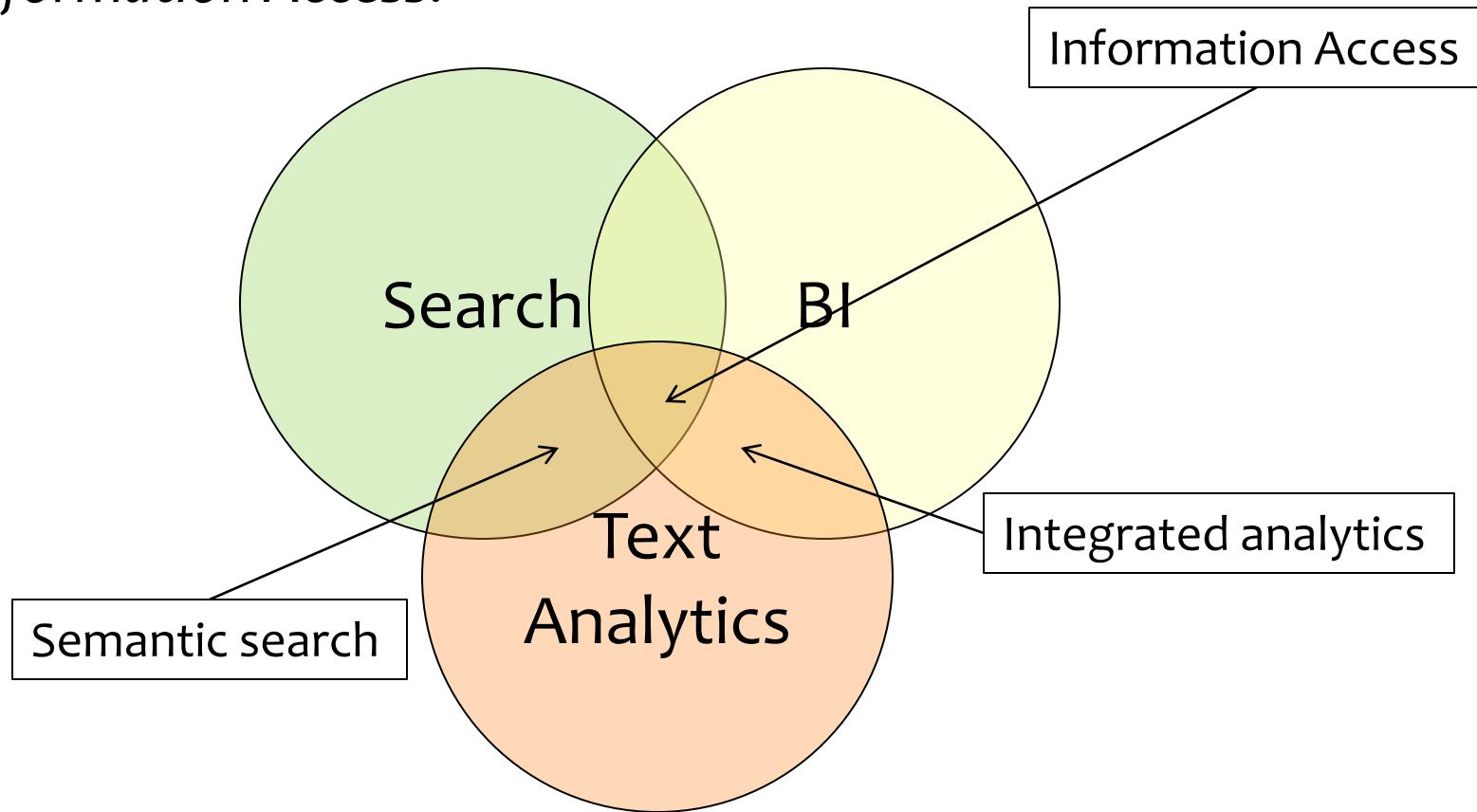
- Concrete and abstract attributes (e.g., 10-year, expensive, comfortable).

- 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 –

**map massachusetts - Google Search - Mozilla Firefox**

File Edit View History Bookmarks Tools Help

Web Images Maps News Shopping Gmail more ▾

Google

map massachusetts

Search Advanced Search Preferences

Web Maps Images

Results 1 - 10 of about 6,500,000 for [map massachusetts](#). (0.12 seconds)

[Massachusetts maps.google.com](#)

Start address   
Get directions

**orc1 - Google Search - Mozilla Firefox**

File Edit View History Bookmarks Tools Help

Web Images Videos Maps News Shopping Gmail more ▾

Google

orc1

Search Advanced search

About 773,000 results (0.10 seconds)

**softitel new york - Bing - Mozilla Firefox**

File Edit View History Bookmarks Tools Help

Web Images Videos Shopping News Maps More

Bing

softitel new york

ALL RESULTS

Local

RELATED SEARCHES

- Sofitel New York City
- Hotel Metro
- Kitano New York
- Affinia 50
- Bryant Park Hotel New York
- New York City Hotels
- The Michelangelo New York
- Grand Hyatt New York

Peru: Population

- 28,674,757 (July 2007)
- 4,500,000 (1908)
- 2,660,881 (1876)

Source: Freebase

**population peru - Bing - Mozilla Firefox**

File Edit View History Bookmarks Tools Help

Web Images Videos Shopping News Maps More

Bing

population peru

ALL RESULTS

Images of Population Peru

RELATED SEARCHES

- Peru Map
- Peru Flag
- Peru Climate
- Peru Language
- Peru Products
- Peru Holidays
- Peru Religion
- Peru Population 2008

Demographics of Peru

Estimated at 22 million in 1990  
50 years (it was slightly more than the ...  
[ddg.com/LIS/aurelia/perpop.htm](#)

Demographics of Peru

Overview · Language · Education  
This article is about the demog...

**ORCL - Oracle Corporation (NASDAQ)**

Google Finance Yahoo Finance MSN Money DailyFinance CNN Money Reuters

21.54 -0.22 (-1.01%) Jun 9 4:00pm ET  
22.01 +0.47 (2.18%) After Hours

Open: 21.77 Volume: 0  
High: 22.09 Avg Vol: 33,824,000  
Low: 21.48 Mkt Cap: 108.11B

Disclaimer

**ORCL: Summary for Oracle Corporation - Yahoo! Finance**

Get detailed information on Oracle Corporation (ORCL) including quote performance, Real-Time ECN, technical chart analysis, key stats, insider transactions, ...  
[finance.yahoo.com/q?s=orcl](#) Cached - Similar

**Oracle Corp. (ORCL) Stock -- Seeking Alpha**

Website  
45 W 44th St - New York  
(212) 354-8844  
★★★★★ • 43 reviews

Get directions from  get directions

**Hotel Sofitel New York - Luxury hotel NEW YORK - Official Web Site**

The ultimate in comfort and convenience, this thirty-story building in midtown Manhattan—a contemporary statement in limestone and glass—is just a stone's throw from Fifth Avenue.

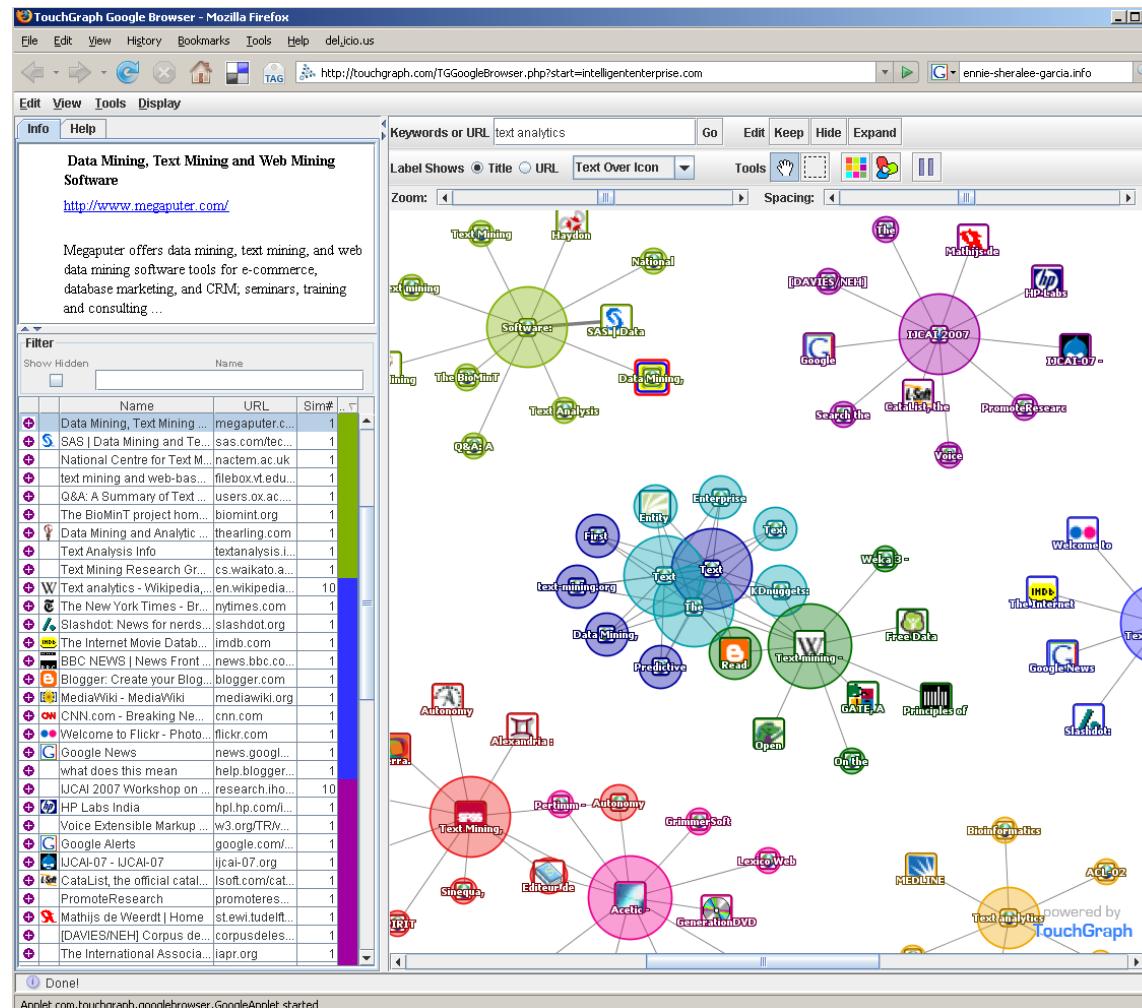
[www.sofitel.com/gb/hotel-2185-sofitel-new-york/index.shtml](#) Cached page

**Hotel Sofitel New York (New York City, NY) - Hotel Reviews ...**

45 West 44th Street New York City NY 10036  
User rating: 5/5 - 4 Star hotel • 1 review  
[www.tripadvisor.com/Hotel\\_Review-g60763-d208454-Reviews-Sofitel\\_New\\_York-New\\_York\\_City...](#) Cached page

# Intelligent Results Presentation

E.g., results clustering on extracted topics:



# Text Data Mining

Data Mining = Knowledge Discovery in Data.

Text Mining = Data Mining of textual sources.

Clustering and Classification.

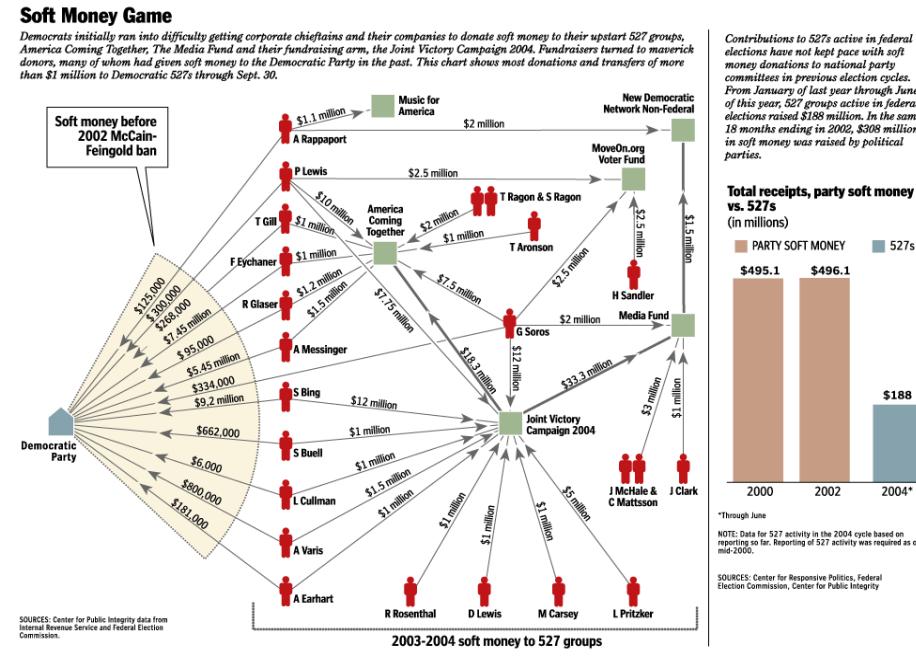
Link Analysis.

Association Rules.

Predictive Modelling.

Regression.

Forecasting.



# Text Data Mining Enables Content Exploration

**Newssift.com - Mozilla Firefox**

File Edit View History Bookmarks Tools Help

http://www.newssift.com/results.jsp?tv=sap&sm=8&n=8&po=0&ed=8&d=8&sort=8&sequence=8&freeTextRe

**FirstRain**

Welcome Penny Log Out My Account Feedback Help

C:MicrosoftCorporation Run My Folders History

Last 31 days change

**SELECTIVITY**  
See all results >

Narrow down to results that mention:

**BUSINESS LINES**  
clear all + more Business Lines

- Operating Systems (1180)
- Search Engines (1472)
- Handheld Operating Systems (4408)

**COMPANIES**  
clear all + more Companies

- Google Inc. (4866)
- Apple Inc. (2889)
- Yahoo Inc. (1412)

**TOPICS**  
clear all + more Topics

- Windows Operating Systems (2168)
- Search Engine - Bing (2152)

Display: Titles | Details Results 1 - 30 of 2387 (from 29092 including similar results) 1 2 3 4 5 6 7 8 9 10 next

**12-MAR-2010**

To Increase Your Social Media Marketing, Think of the Search Engines as Your Customers Article Alley

Kevin Grandia and Tim Kolke: The biggest threat to Google might be your friends The Huffington Post

Microsoft Stumbles with Office Updates Maximum PC

Yahoo Improves Search Results, Reminds World That It Still Has A Search Engine (YHOO) Silicon Alley Insider 1 similar result(s)

Analysts' review of SharePoint 2010 MSON Blogs

Motorola + Android + Bing: Is Microsoft jumping ship? GoMo News

**FURTHER SUGGESTED RESEARCH**

Competitors of Microsoft Corporation

- Google Inc.
- Yahoo Inc.
- Apple Inc.
- Hewlett-Packard Company
- Research In Motion Limited

Industry Topics

- Mobile Applications Stores
- Internet Marketing & Advertising
- Windows Operating System
- Video Gaming Console War
- Software As A Service (SaaS) Trends

**EVENT HIGHLIGHTS**

**Alloy (Notes/SAP integration device)**

IBM - Ed Brill, May 5, 2009, By Ed Brill

Alloy by IBM and SAP to run on a mid-Today, we're demonstrating technolog access SAP-related workflow on a BlackBerry... I am going to be at SAP on Monday and Tuesday in Orlando

**SAP announces Business availability**

ITWorld Canada, May 5, 2009

The mass shipment represents the customers... SAP announced that available Tuesday, three months after software vendor unveiled the library to grant customers better insight int

**1-10 of 3566 Articles:**

Sort by: **Relevance** | Date

Timeframe: May 7, 2009 to May 6, 2009

**Page Filter:** SAP AG

Companies Organization

0 1 16 0

**Trademarks**

Invention AT&T Inc. Jim Shepherd Novell Jim Kagermann Henning Kagermann Jim Shepherd Bill McDermott Show More

Supply Chain Software as a Service Vale Inco Novell Jim Kagermann Henning Kagermann Jim Shepherd Bill McDermott Show More

Acquisition Wipro Ltd. AT&T Inc. Jim Shepherd Novell Jim Kagermann Henning Kagermann Jim Shepherd Bill McDermott Show More

Radio-Frequency Identification Cisco Systems Inc. Vale Inco Novell Jim Kagermann Henning Kagermann Jim Shepherd Bill McDermott Show More

Environment BlackBerry Oracle Corporation Peoplesoft Inc. Gartner Inc. IDC International Data Corp. Inviscid Media Limited Show More

Ecosystem Novell Jim Kagermann Henning Kagermann Jim Shepherd Bill McDermott Show More

Business Objects SA International Business Machines... Microsoft Corporation Hewlett-Packard Company Show More

Incubate Media Limited Show More

**Quotes**

can be taken into SAP seamlessly. Pending added. "We will also be working on capturing SAP data using hand-held terminals so that construction personnel will find it more convenient to interact with SAP. Besides, we will be automating all the employee-employer transactions using an employee portal. The same will also serve as

Transferring data from www.silobreaker.com...

**Related Documents**

Documents | Stories All (1425) | News (139) | Reports (0) | Blogs (32) | Audio/Video (0) | Fact Sheets (2) sort by: Date | Relevance

NETSUITE ENTERS SAP'S CORE MARKET WITH NEW-ON-DEMAND VERTICAL SUITE FOR MANUFACTURERS - Smart News Network [1 hour ago]

Business Objects launches Predictive workbench - FSN [4 hours ago]

## *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:

9 17 1 20

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:

$$f = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

**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

[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](#)

[Login](#) | [Become a member!](#)

[search](#)

**RANKS.NL** | **KEYWORD DENSITY & PROMINENCE v1.5b**

Url tested : <http://altaplana.com/SentimentAnalysis.html>

[— More Domain / URL info —](#)

word	repeats	density	Prominence	word	repeats	density	Prominence
sentiment	18 L,I	<b>1.26%</b>	46.93	for	17 L	<b>1.19%</b>	34.44
that	15	<b>1.05%</b>	55.22	text	15 L	<b>1.05%</b>	58.77
analytics	12 L	<b>0.84%</b>	52.83	from	10	<b>0.70%</b>	71.16
management	9 H	<b>0.63%</b>	50.37	analysis	9 L,I	<b>0.63%</b>	50.61
our	8	<b>0.56%</b>	20.36	are	8	<b>0.56%</b>	56.38
influence	7 H	<b>0.49%</b>	78.46	customer	7 H	<b>0.49%</b>	33.75
which	6	<b>0.42%</b>	63.18	understanding	6	<b>0.42%</b>	47.34
she	6	<b>0.42%</b>	68.22	notes	6	<b>0.42%</b>	51.18
have	6	<b>0.42%</b>	35.14	can	6	<b>0.42%</b>	55.43
been	6	<b>0.42%</b>	28.93	understand	5	<b>0.35%</b>	57.77
they	5	<b>0.35%</b>	54.28	sources	5	<b>0.35%</b>	87.31
not	5	<b>0.35%</b>	37.68	more	5	<b>0.35%</b>	42.90
mining	5	<b>0.35%</b>	55.84	mail	5	<b>0.35%</b>	63.50
extraction	5	<b>0.35%</b>	40.15	enterprise	5 H	<b>0.35%</b>	40.59
way	4	<b>0.28%</b>	23.61	time	4	<b>0.28%</b>	20.59
take	4	<b>0.28%</b>	14.78	surveys	4 L	<b>0.28%</b>	50.39
support	4	<b>0.28%</b>	21.75	results	4	<b>0.28%</b>	38.58
potential	4	<b>0.28%</b>	39.97	positive	4	<b>0.28%</b>	56.36
opinion	4	<b>0.28%</b>	71.71	networks	4 U	<b>0.28%</b>	75.02

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:  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 : Opportunities for follow-on focus on applications . It 's the breadth of opportunity to extract and analyze attitudinal information from sources as varied as notes and survey responses and the difficulty of the technical applications so interesting .

We will explore three applications influence networks , assess customer 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 [ 5 ] to mapping studies . She seeks to understand the correlation of include event and interview transcripts , presentations , media abstracts , with her clients ' scientific and promotional message concentrated on sources where large volumes of readily mine adding blogs to the mix .

Jeff Catlin , CEO of text-analytics vendor [ Lexalytics ] [ 6 ] , describes his company as his company s best success story . Cisco us executives have the highest correlation to positively moving the found that certain executives had a positive influence on the 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).**

<http://wordle.net>

Total word occurrences in the document:

Different words in document:

### Total of different words

### Less different common words

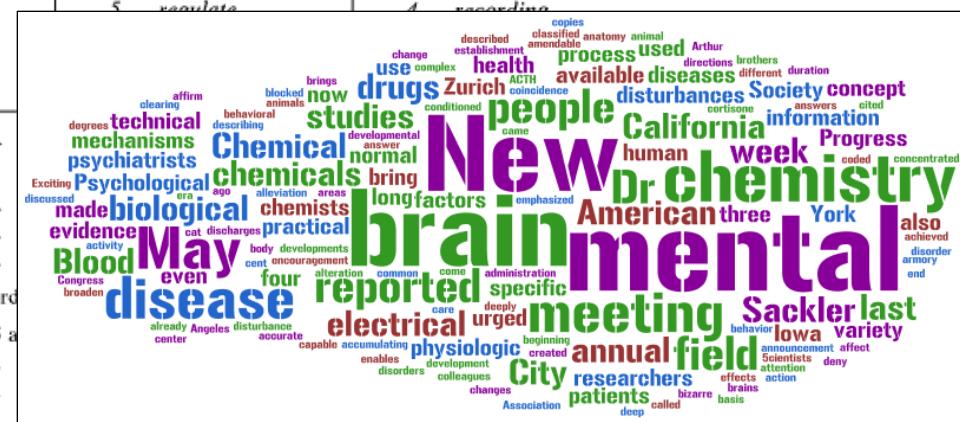
#### Different non-common words

Ratio of all word occurrences to different non-common words

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

### Total occurrences

### Different words



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.”*

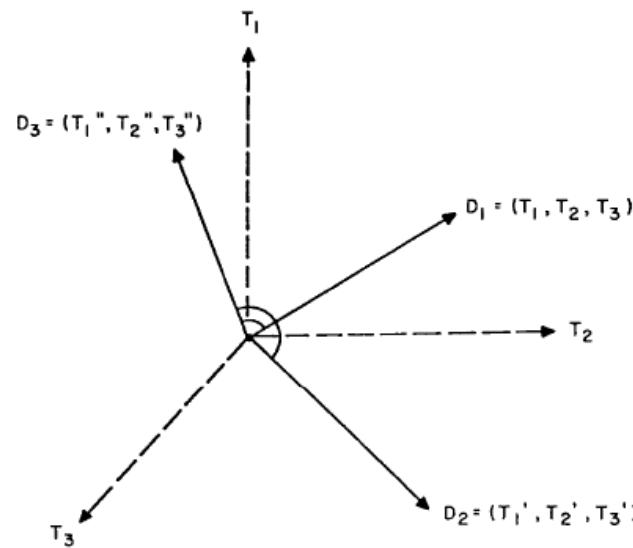
-- H.P. Luhn, *The Automatic Creation of Literature Abstracts*, IBM Journal, 1958.

## 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.

Fig. 1. Vector representation of document space.



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

# 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](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.

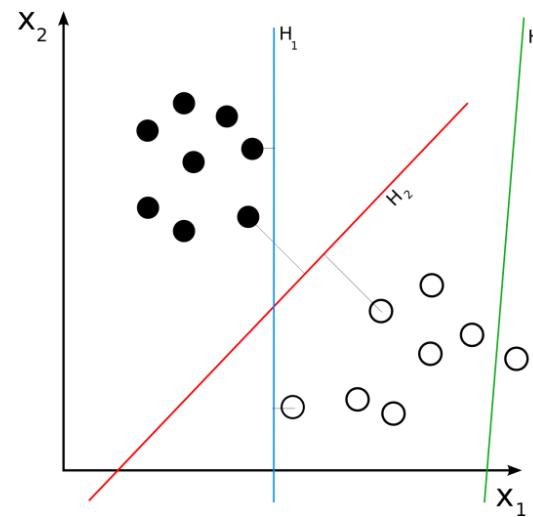
Takes care of synonymy, polysemy, stemming, etc.

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...**

102 - Total			Total 3 word phrases : 45 - Total Repeats : 93		
phrase	repeats	density	phrase	repeats	density
customer experience management	3 H	0.63 %	customer experience management	3 H	0.63 %
enterprise feedback management	3 H	0.63 %	enterprise feedback management	3 H	0.63 %
of text analytics	3	0.63 %	of text analytics	3	0.63 %
analytics can be	2	0.42 %	analytics can be	2	0.42 %
analyze attitudinal information	2	0.42 %	analyze attitudinal information	2	0.42 %
and analyze attitudinal	2	0.42 %	and analyze attitudinal	2	0.42 %
and survey responses	2	0.42 %	and survey responses	2	0.42 %
applied to extract	2	0.42 %	applied to extract	2	0.42 %
articles blog postings	2	0.42 %	articles blog postings	2	0.42 %
as articles blog	2	0.42 %	as articles blog	2	0.42 %
as varied as	2	0.42 %	as varied as	2	0.42 %
attitudinal information from	2	0.42 %	attitudinal information from	2	0.42 %
be applied to	2	0.42 %	be applied to	2	0.42 %
blog postings e	2	0.42 %	blog postings e	2	0.42 %
call center notes	2	0.42 %	call center notes	2	0.42 %
can be applied	2	0.42 %	can be applied	2	0.42 %
center notes and	2	0.42 %	center notes and	2	0.42 %
ceo of text	2	0.42 %	ceo of text	2	0.42 %
experience for help	2	0.42 %	experience for help	2	0.42 %

**“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

analyze attitudinal	2	0.28 %	96.66	online consumer forums	2	0.42 %	55.90
and analyze	2	0.28 %	96.73	postings e mail	2	0.42 %	95.96
and other	2	0.28 %	97.70	real time two	2	0.42 %	10.50

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 beginning 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 (tranquillizers, 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 Martí-Ibañez and three brothers, Dr. Mortimer D. Sackler,

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.

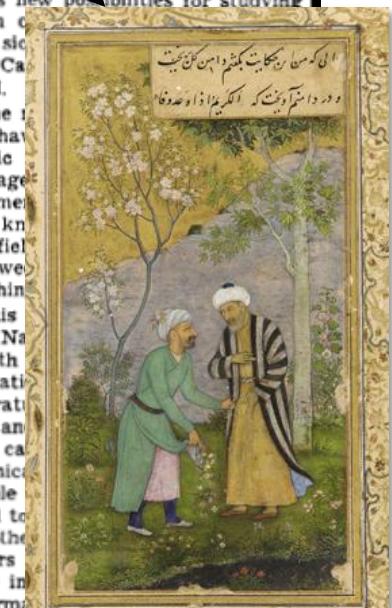
### Chemistry of the Brain

At the psychologist's meeting here, a technique for tracing electrical activity in specific portions of the animal brain was described by researchers of the University of California, who reported in cat electrically which In this reported sequence in various In this the brain may be located. Furthermore, the electrical pathways so traced out can be blocked temporarily by the use of chemicals. This opens new possibilities for studying brain and side effects of the Caisized.

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.

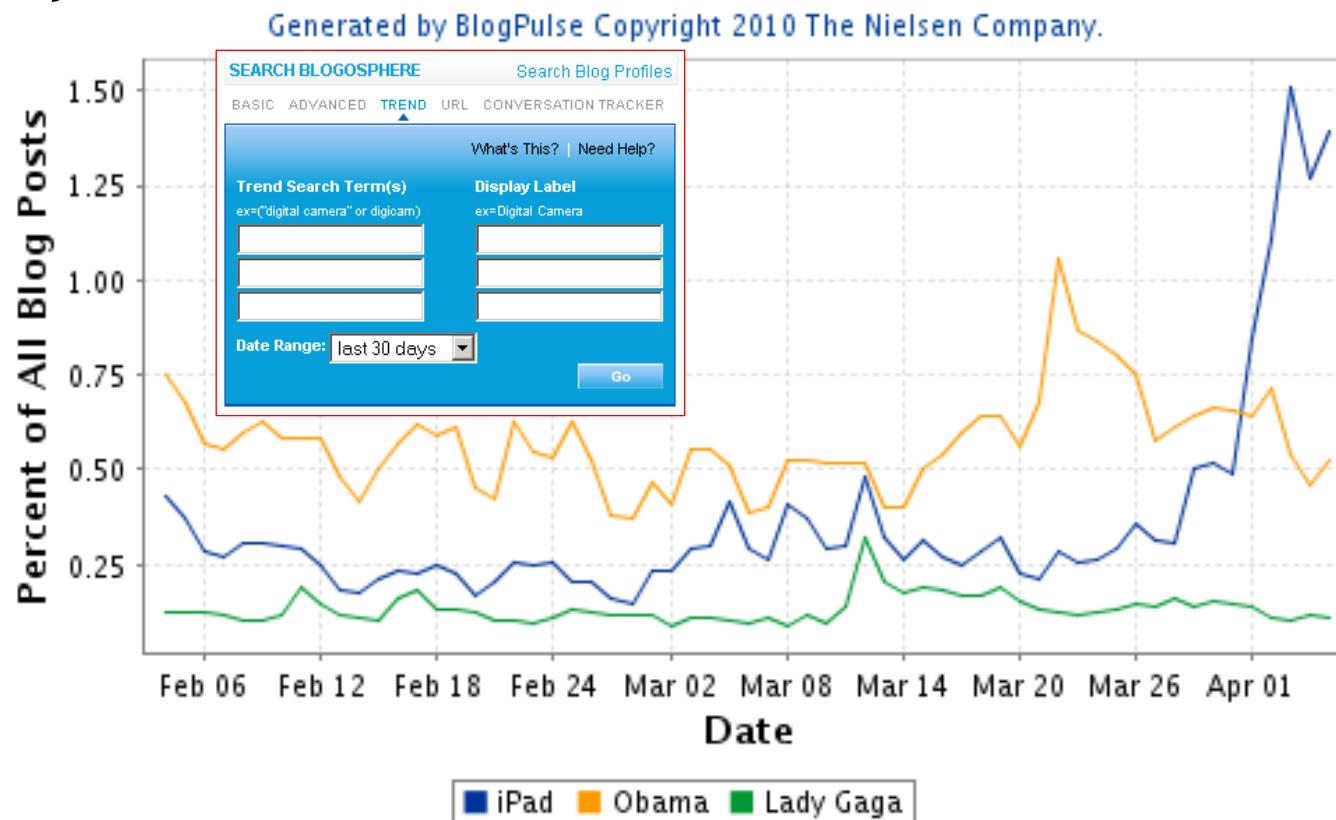
"It led some years ago to the development of a blood test which was 80 per cent accurate in the identification of schizophrenic patients," they said. "It permitted us on physiologic grounds to deny that the psychoneuroses and the psychoses were lesser and greater degrees of the same disease process,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



# 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

The screenshot shows a Mozilla Firefox browser window displaying the Connexor website at <http://www.connexor.eu/technology/machinese/demo/syntax/>. The page title is "Connexor - Technology - Machinese - Demo - Machinese Syntax - demo - Mozilla Firefox". The Connexor logo is visible in the top left. A navigation bar includes links for Home, Company, Solutions, Technology (which is selected), Partners, and Contact. A breadcrumb trail shows the path: Technology > Machinese > Demo > Machinese Syntax - demo. The main content area features a heading "Machinese Syntax" and a paragraph describing it as a syntactic parser that returns base forms and compound structure, produces part-of-speech classes, inflectional tags, noun phrase markers and syntactic dependencies. Below this is a text input field containing the sentence "What's the best price for new laptop that I'll use for business trips and around the office?", followed by a dropdown menu set to "English text" and a "Apply Syntax" button. A note at the bottom states, "This demo is intended for evaluation purposes only."

# Parts of Speech



# Parts of Speech

**Connexor - Technology - Machinese - Demo - Machinese Phrase Tagger - demo - Mozilla Firefox**

File Edit View History Bookmarks Tools Help del.icio.us

http://www.connexor.eu/technology/machinese/demo/tagger/ Google

**connexor**  
natural knowledge

Sitemap

Home Company Solutions Technology Partners Contact

Technology > Machinese > Demo > Machinese Phrase Tagger - demo

**English Machinese Phrase Tagger 4.6 analysis:**

Text	Baseform	Phrase syntax and part-of-speech
What	what	nominal head, pro-nominal
's	be	main verb, indicative present
the	the	premodifier, determiner
best	good	premodifier, superlative adjective, noun phrase begins
price	price	nominal head, noun, noun phrase continues
for	for	postmodifier, preposition, noun phrase continues
new	new	premodifier, adjective, noun phrase continues
laptop	lap top	nominal head, noun, noun phrase ends
that	that	nominal head, pro-nominal
I	I	nominal head, pro-nominal
'll	will	auxiliary verb, indicative
use	use	main verb, infinitive
for	for	preposed marker
business	business	premodifier, noun phrase continues
trips	trip	nominal head, plural
... 1 ...	...	...

Connexor Oy, Helsinki Business and Science Park  
© Connexor Oy, Power

Done

## From POS to Relationships

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

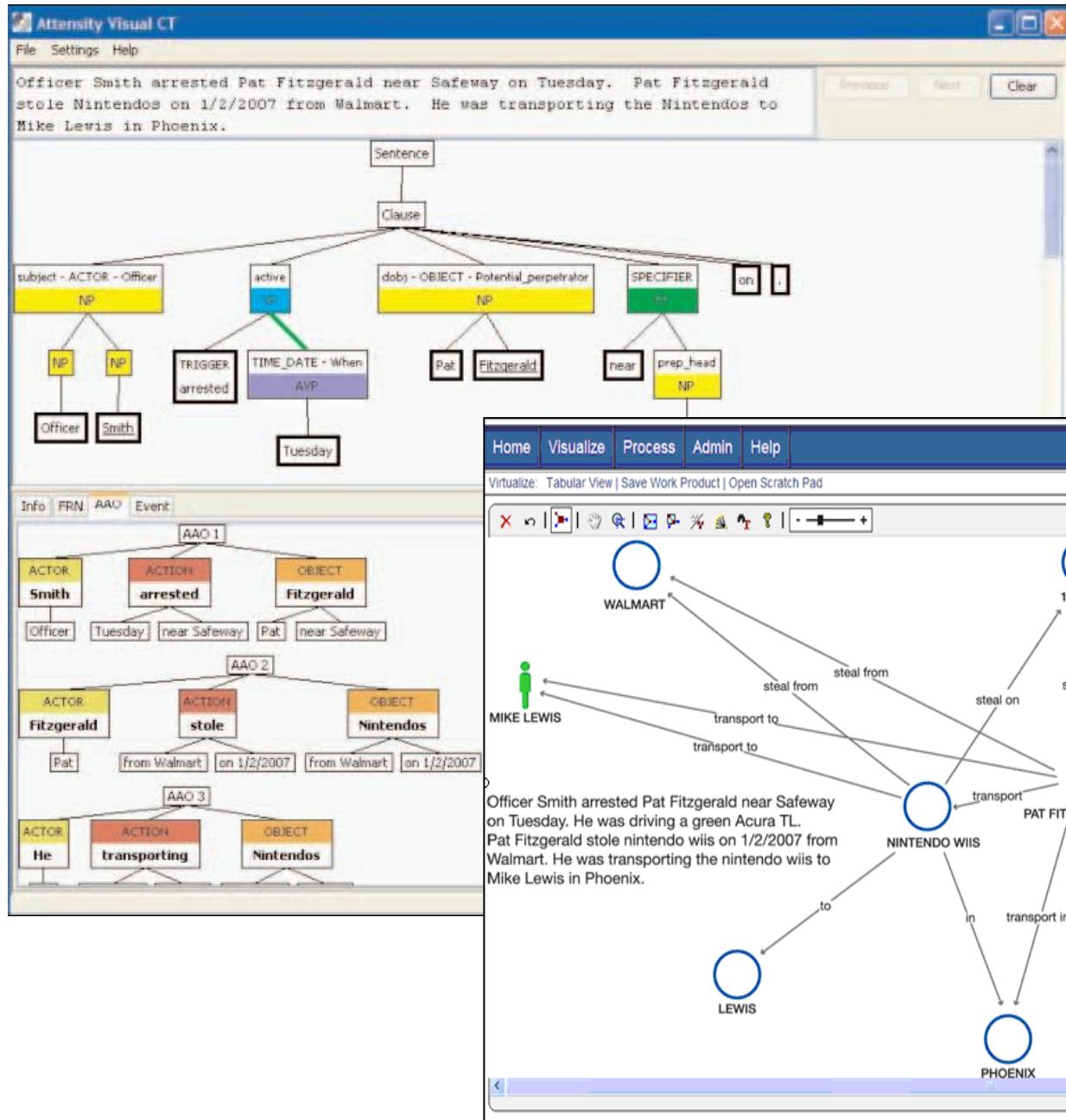
<subject> <verb> <object>

– 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/>



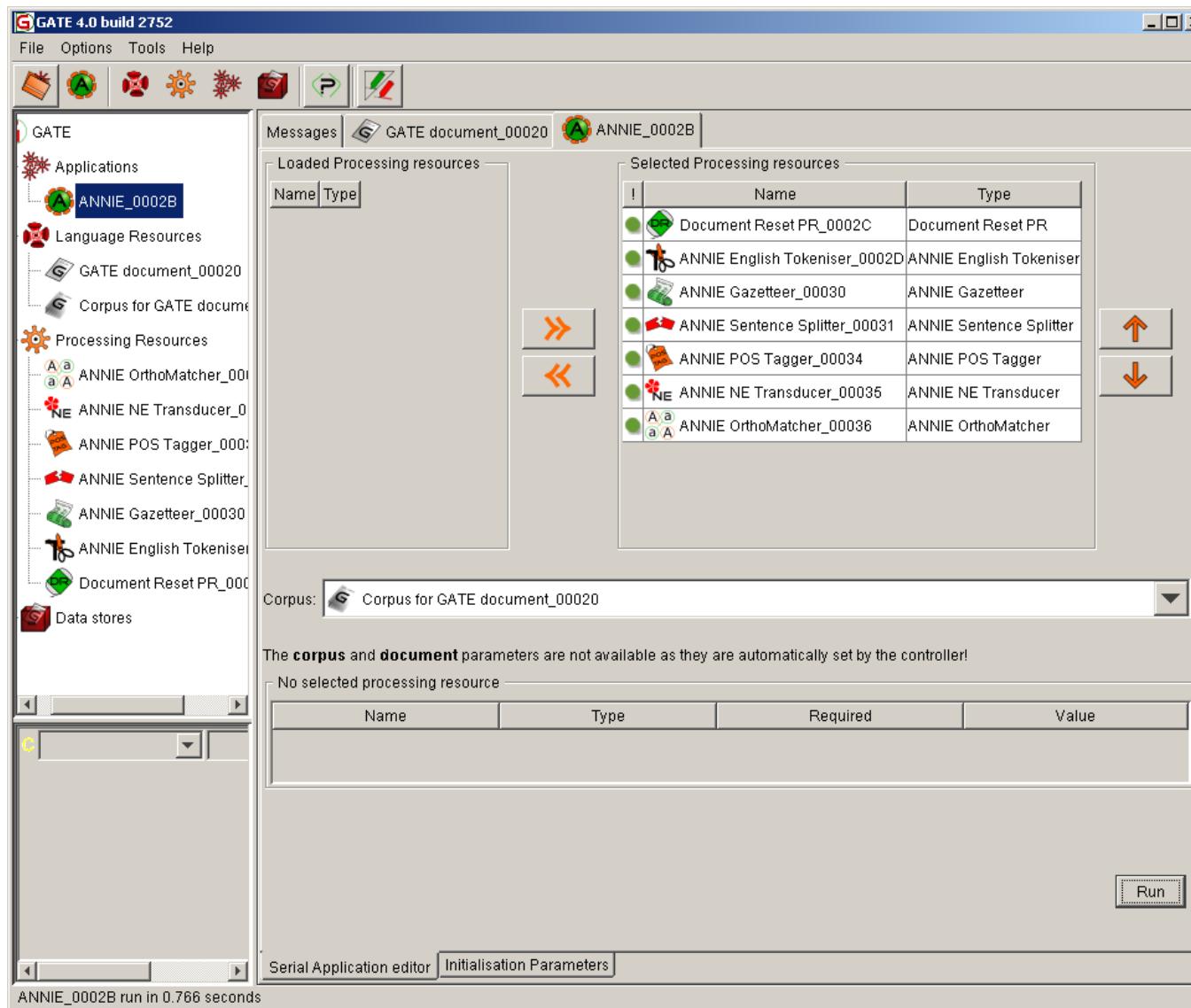
# Tagging with GATE

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

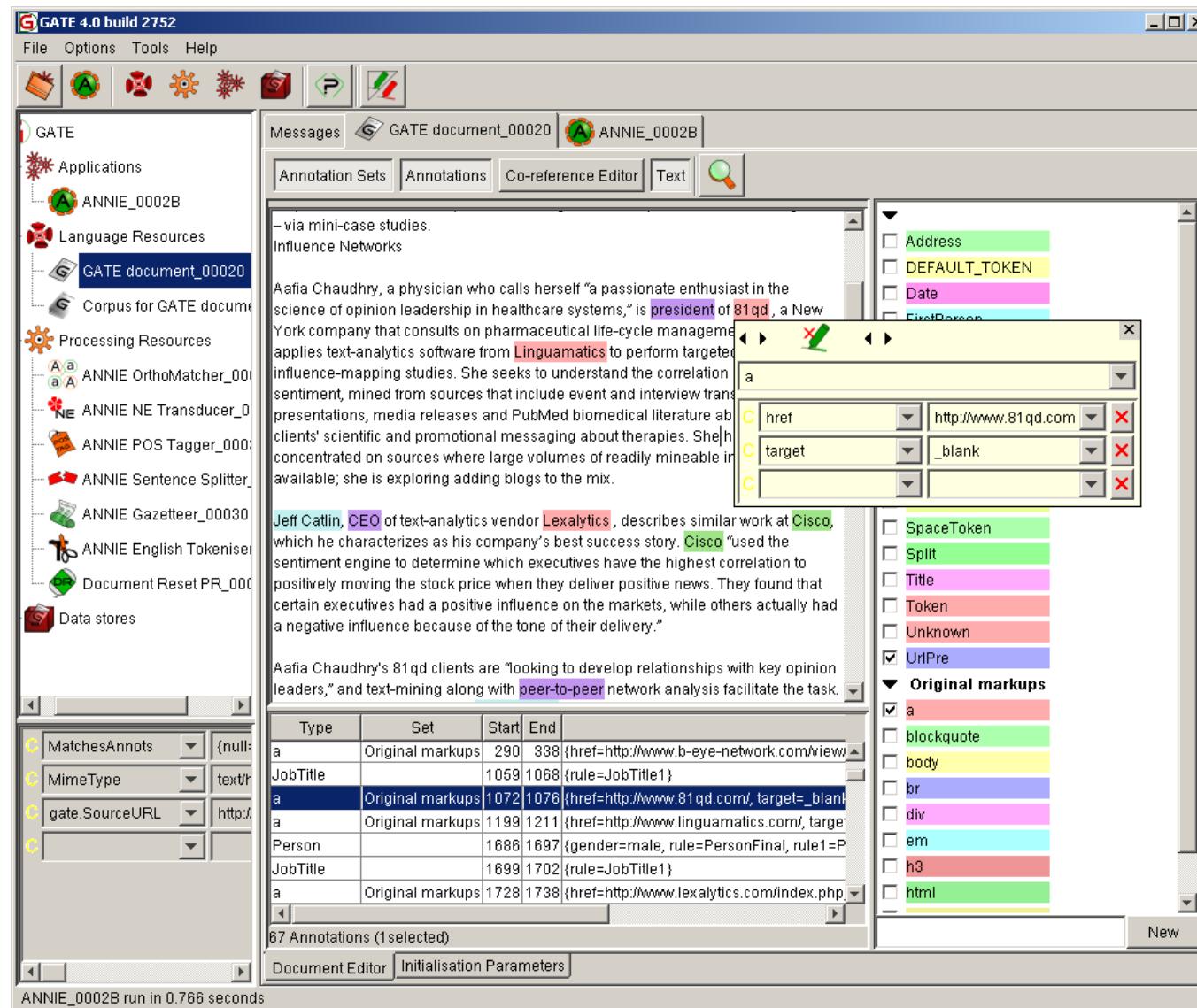
The screenshot illustrates the GATE 4.0 build 2752 interface for text mining and visualization. On the left, a browser window displays the BeyeNETWORK website, specifically a blog post titled "Sentiment Analysis: A Focus on Applications" by Seth Grimes. The post discusses the breadth of opportunities for sentiment analysis across various sources like articles, blog posts, and survey responses. On the right, the GATE interface shows the annotated version of the same text. Annotations are highlighted with red boxes, indicating specific entities or concepts identified by the system. The interface includes a toolbar at the top, a sidebar with various resources, and a main panel for viewing and managing annotations. A detailed table of annotations is shown in the bottom right, listing each annotation's type, start and end positions, and associated features.

Type	Set	Start	End	Features
a	Original markups	48	59	(href=/channels/index.php?filter_channel=1394, t)
a	Original markups	266	266	(href=http://www.clarabridge.com/, isEmptyAndSt)
a	Original markups	290	338	(href=http://www.b-eye-network.com/view/6744, t)
a	Original markups	1072	1076	(href=http://www.81qd.com/, target=_blank)
a	Original markups	1199	1211	(href=http://www.linguamatics.com/, target=_blan)
a	Original markups	1728	1738	(href=http://www.lexalytics.com/index.php, target=
a	Original markups	3919	3937	(href=http://www.andersonanalytics.com/, target=

# 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" />
</TextWithNodes>
<!-- The default annotation set -->
<AnnotationSet>
  <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>
</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.

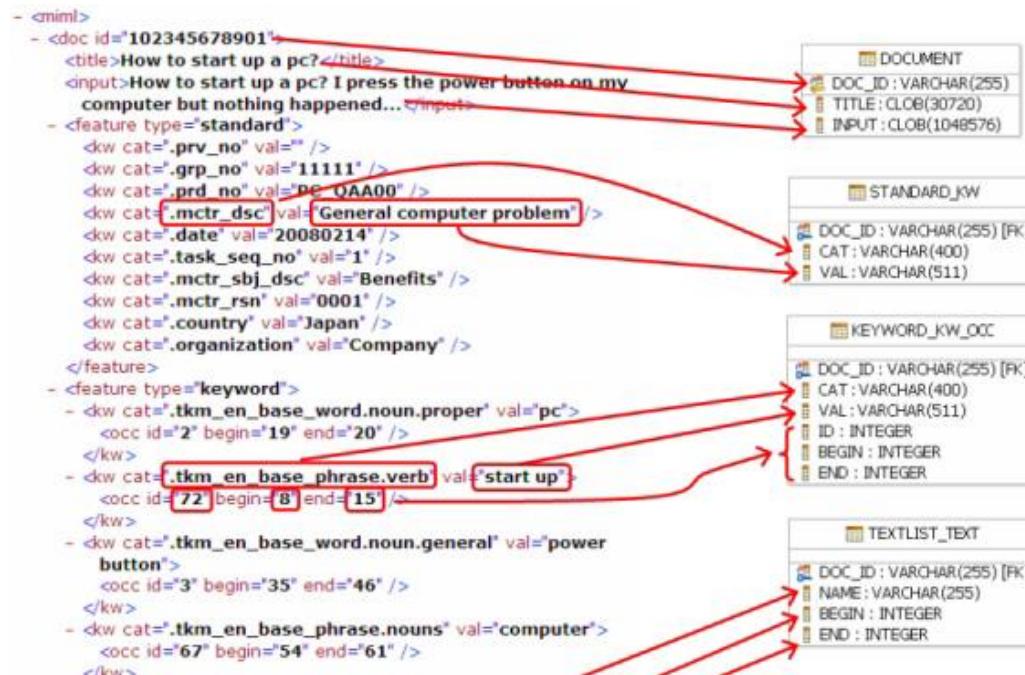
We can have both discovered and predetermined classifications (taxonomies) of text features.

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.”*



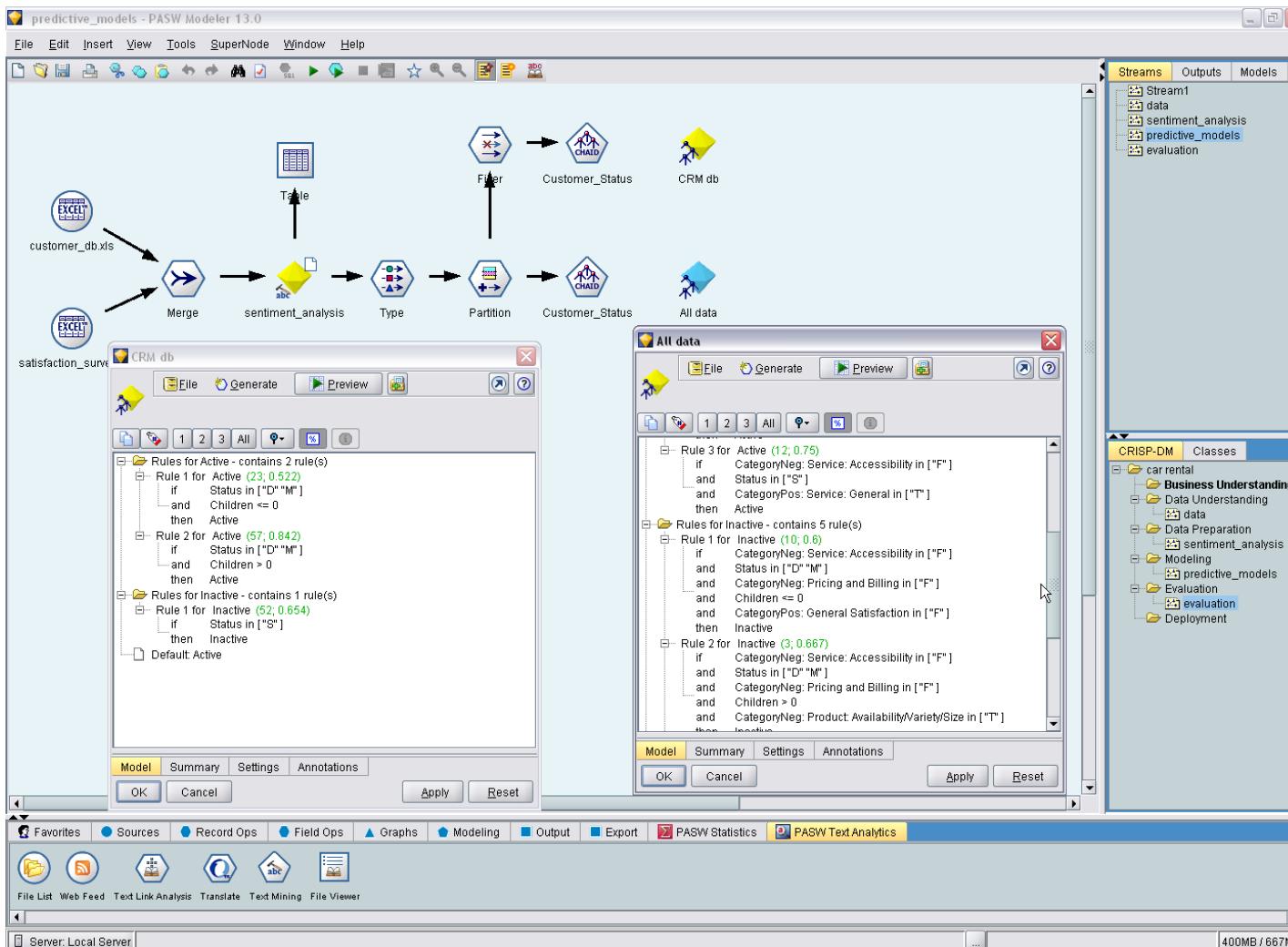
# Sophisticate Pattern Matching

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

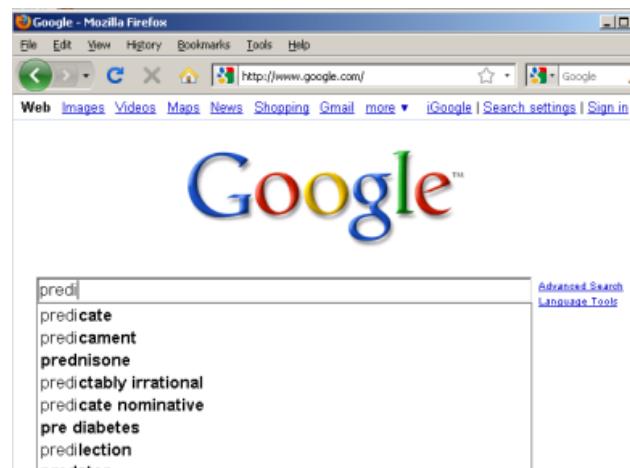
[http://en.wikipedia.org/wiki/File:  
ITap\\_on\\_Motorola\\_C350.jpg](http://en.wikipedia.org/wiki/File:ITap_on_Motorola_C350.jpg)



Disambiguation

: use dictionaries, context

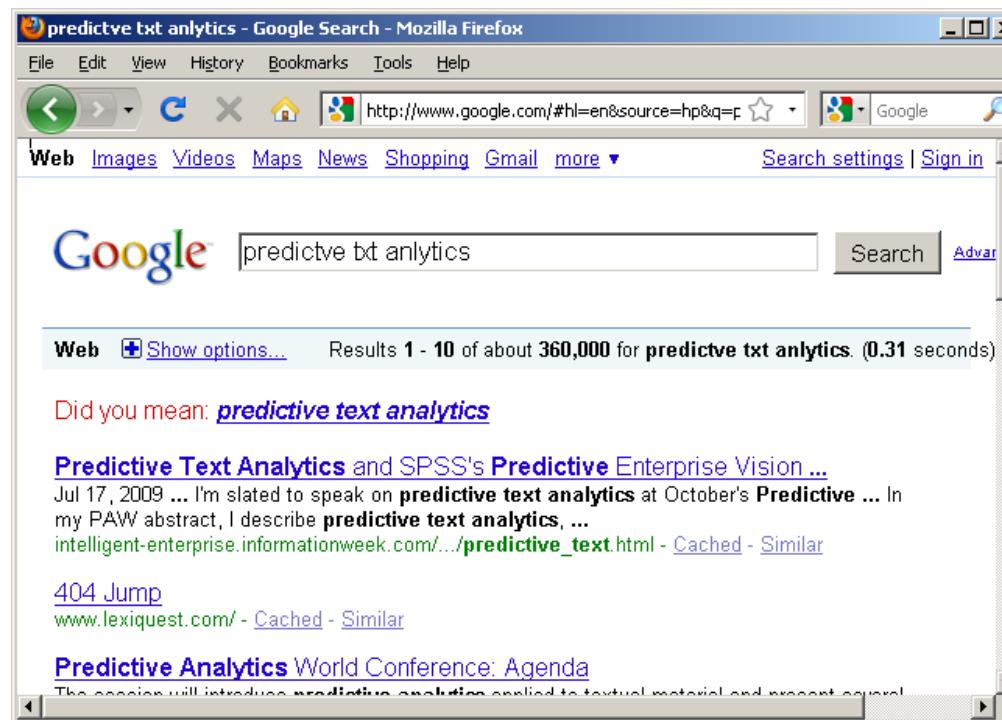
Error correction



## 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...

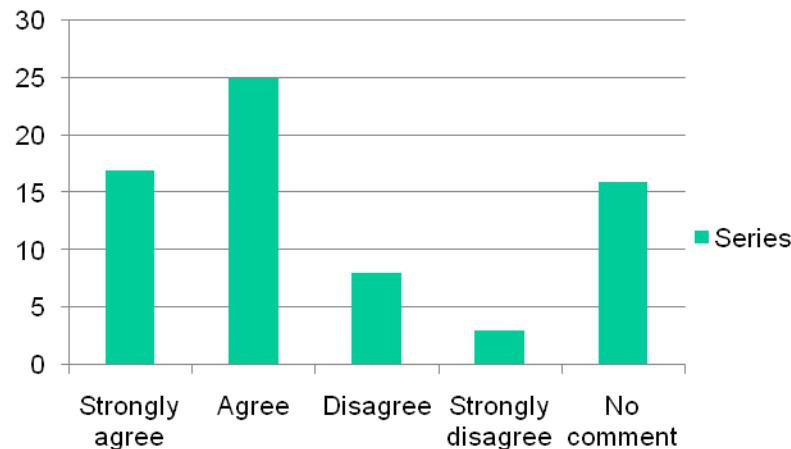
# Structured & ‘Unstructured’

The respondent is invited to explain his/her attitude:

My overall experience was positive.	<input type="radio"/>				
<b>Please complete the section below if your contact with us involved permitting/licensing/registration assistance.</b>					
The regulations were understandable.	<input type="radio"/>				
The application instructions were understandable.	<input type="radio"/>				
The terms and conditions of the permit, license, or registration were understandable.	<input type="radio"/>				
<b>Please indicate the name(s) of any staff person you would like to commend:</b>					
<input type="text"/>					
<b>Comments:</b>					
<input type="text"/>					
<b>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:</b>					
<input type="text"/>					

# Structured & ‘Unstructured’

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



The guest reviews are submitted by our customers after their stay at [Hilton New York](#). They are opinions of guests that reflect their experiences when staying at this hotel.

**Total 8**

Based on 60 reviews

Staff: 7.8 Services: 7.8 Clean: 8.6 Comfort: 8.4 Value: 7.6 Show scores of: All reviewers 60 Young couples 5 Families with young children 2 Families with older children 6 With friends 9

**Individual guest reviews**

Reviews are ordered by language and date with a maximum of 25 reviews

[Previous page](#) Showing 1 - 25 (Total 60) [Next page](#)

- Errol** Mature couple NOWRA, Australia April 5, 2009
 

Overall stay was excellent and typical of what I would expect from any Hilton Hotel.

Being constantly harassed by staff at a counter set up between the entrance and the lifts, to be given a welcome gift which is really only a waste of our time as it is a promotion for selling timeshare units. We have experienced this before in places like Las Vegas, but did not expect it at the New York Hilton.
- Jane** WITH FRIENDS DRONFIELD, United Kingdom April 1, 2009
 

We had the most fabulous stay at the Hilton. we were given a free upgrade on the room which was lovely. The bed and bedding were divine loved getting in after a very busy day. The hotel was in an excellent location for everything. We were very well looked after and I highly recommend this hotel. The bridges cocktails were fab.
- Jim** Family with older children FORT MCMURRAY, Alberta, Canada March 30, 2009
 

Room size, staff & cleanliness of hotel excellent. Great location for many attractions such as times sq. & central park. Subway only 2 blocks so access to everywhere.

In hotel dining somewhat expensive.
- Anonymous** Family with older children UNITED STATES OF AMERICA
 

auto clean was asked every day to use the clean facilities, you can...

## 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

The image displays three separate browser windows side-by-side:

- Newssift.com - Mozilla Firefox:** A search interface for news topics. It shows a sidebar with a pie chart for sentiment analysis (Positive 326, Neutral 1205, Negative 1000) and another pie chart for article sources. The main area lists search terms like "Health Care Reform" and "Barack Obama".
- softitel new york - Bing - Mozilla Firefox:** A search results page for "softitel new york" on Bing. It includes a sidebar with related searches such as "Sofitel Luxury Hotels", "Sofitel New York City", "Hotel Metro", "Kitano New York", "Affinia 50", and "Bryant Park Hotel New". Below the search bar, there's a link to "Listings for softitel near New York, NY".
- RankSpeed - Canon PowerShot G11 - Mozilla Firefox:** A sentiment analysis tool for the Canon PowerShot G11. It shows a list of tweets from users like @uwphotostore, @mrv\_pwp, @canoneosrebel, @Colin\_Kea, @Macols, @xLighty, and @shawtyfied. To the right, a bar chart visualizes the distribution of sentiments: excellent (24.4%), good image (6.7%), fragile (0%), good zoom (0%), cool (0%), poor quality (0%), easy (24.4%), powerful (6.7%), problem (0%), fast (6.7%), solid (0%), and slow (6.7%).

**WE twenz pro service: influence analytics for Twitter - Mozilla Firefox**

File Edit View History Bookmarks Tools Help

waggeneredstrom.com https://wexview.waggeneredstrom.com/twendzpro/trial.aspx?d=last7&project\_id=55

FEATURES ABOUT TRY IT NOW

LOGIN

**we|twenz pro** SIGN UP

Search GO

**Facebook**

TWEET THIS

1,500  
1,000  
500  
0

Mar 29 Mar 30 Mar 31 Apr 01 Apr 02 Apr 03 Apr 04 Apr 05

Time Period Last 7 Days Frequency Reach Influence Positive Emotion Negative Emotion

**IMPACT**

Overall Impact:	8.3
Engagement:	10.0
Resonance:	8.0
Exposure:	7.0

**KEY PERFORMANCE INDICATORS**

Velocity (VpH):	33.2
Total Potential Reach:	12949046
Avg Potential Reach:	1871
Avg Potential Influence:	2.2

**SUBTOPICS**

- facebook
- google
- mail
- media
- trouble

**EMOTION**

- 0% POS
- 0% NEU
- 100% NEG

**WORD CLOUD**

– against banks battle facebook google  
impact integrates mail marketing media  
minute new not rt say social trouble twitter  
yahoo

**EMOTION BY FREQUENCY: NEGATIVE EMOTION**

Show page 1 out of 10 Next Page >

jasonyormark: 6 days ago.. followers: 42243

Social Media Minute: Yahoo Mail Integrates Facebook, Google in Trouble with ... By Jason Harris

delicious50: 3 days ago.. followers: 15606

Ted Facebook Marketing Campaign — Social Media Optimization <http://bit.ly/9O9kiR>

bexdeep: 8 days ago.. followers: 24390

Facebook versus Twitter – Battle for Social Media Space | Globalthoughtz Books <http://mxy.in/5nv2y>

ScottyMore: 4 days ago.. followers: 33838

Facebook icons that are surprisingly not utter crap | social media ... <http://bit.ly/bHIStc>

**CONVERSATION**

all tweets:	6109
retweets:	1446
replies:	155
links:	5212
hashtags:	2548
questions:	758

**TOP POS/NEU INFLUENCERS**

**TOP NEGATIVE INFLUENCERS**

**INFLUENCER DENSITY**

**EMOTION BY FREQUENCY**

**EMOTION BY INFLUENCE**

Rated negative?

## 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

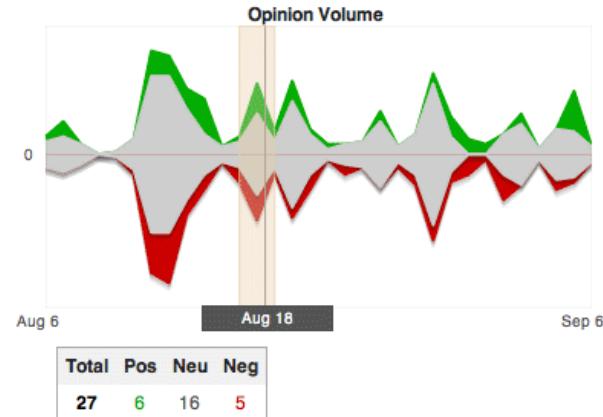
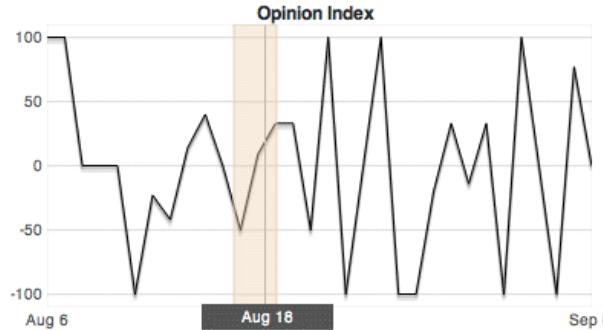
**Opinion Holder:** International Energy Agency, Boone Pickens, Energy Information Administration

[Create an alert](#)[Look Back:](#) 1d 1w 1m 3m 6m**Opinion Holder (4440)**[clear selected](#) | [show all](#)

- Barack Obama (331)
- John McCain (305)
- International Energy Agency (132)
- Nancy Pelosi (78)
- Sarah Palin (76)
- Boone Pickens (65)
- Congress (62)
- 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](#) Alternative Energies (ARAOI)

August 06, 2008 - September 06, 2008



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 US Government, stated that in 2004 the United States produced about 22 percent of global carbon dioxide emissions, primarily because the US economy is the

# ... and Missteps

**twitrratr**

SEARCHED TERM **sentiment analysis**

POSITIVE TWEETS	NEUTRAL TWEETS
5	25

16.13% POSITIVE

"Kind" = type, variety, not a sentiment.

External reference

@aduncanfreeman i'm always kind by potential of melding sentiment and mapping with my kind of analysis. (view)

interesting topics: 1) mining and sentiment 2) statistical language for ir 3) learning to rank

you are right, that's kinder is actually a great engine for market sentiment analysis (view)

@missmci great article, right now i'm (well my tester is) doing scale tests on my sentiment analysis layer the space could use more research (view)

80.65% NEUTRAL

"Why Google Really Wants Twitter: Real Time Sentiment Analysis Scoring" http://tinyurl.com/c35

"Why Google Really Wants Twitter: Real Time Sentiment Analysis Scoring" marketingshift.com/2/google-really-wants (view)

Data Value: Online Sentiment Analysis vs. Face-to-Face Focus Groups http://bit.ly/2LOouQ (view)

Data Value: Online Sentiment Analysis vs. Face-to-Face Focus Groups http://bit.ly/3jDcmI (view)

Data Value: Online Sentiment Analysis vs. Face-to-Face Focus Groups (view)

13.07% POSITIVE

i will honestly miss senator ted kennedy. he was a great american. (view)

wasn't a huge ted kennedy fan, however, can't help but feel our country has lost a great man or a piece of history (view)

years ago i was fishing in hyannis port and ted kennedy sailed by me in his boat. he waved. it's a good memory. #fb (view)

ted kennedy died. what a great senator. (view)

a little teary over the passing of ted kennedy. with his family tragedies, i am glad he was able to live a full successful life. (view)

83.20% NEUTRAL

R.I.P. Senator Ted Kennedy. "The Lion of the Senate." (view)

Rest in peace Ted Kennedy...you will be missed. (view)

Ted Kennedy open thread...share your thoughts...http://tinyurl.com/hgnwyc (view)

RT @berryflavor RT @Juana4ev: Another one dies on the 25th...RIP TED KENNEDY---damn that's creepy (view)

RIP Ted Kennedy. I'll pour a little liquor out for you...on your grave. (view)

Regardless of your beliefs, Ted K (view)

3.73% NEGATIVE

just woke up this morning with sad news being my first listen. ted kennedy has died. (view)

r.i.p. ted kennedy, i'm getting tired of tweeting all these famous ppl dying, but ted gets a pass... (view)

we all realize the hcplan cannot possibly move on without ted kennedy. may he rest in peace (after confessing to god for his bad acts). (view)

good morning awaken to sad news rip senator ted kennedy (view)

ted kennedy died. can say a lot of negative things about him and a lot of positive. either god for his beliefs

Complete misclassification

Unfiltered duplicates

# 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”



hntrpyanfar: \*minor happydance\* - Mozilla Firefox

File Edit View History Bookmarks Tools Help

<http://hntrpyanfar.livejournal.com/225723.html>

LiveJOURNAL™

Explore LJ Life Entertainment Music Culture News & Politics

hntrpyanfar ( hntrpyanfar) wrote,  
@ 2008-08-27 13:55:00

Current mood:  contemplative

Current music: Should I Stay or Should I Go - The Clash

Entry tags: [life](#), [science](#)

**\*minor happydance\***

Our collaborator got back to us on my 1st author paper! We're addre

This is good, because I recently met A, another grad student on the weeds. Her opinion as to why she has no interviews yet is that s credit.

I really, really hope that's not it. Even if our paper goes in Sept 1, w Oct 1... and I don't wanna wait that long for people to call me back!

Now I'm wondering if I should even send out resumes without that... Collaborator, for taking so long!

Ach vell.

<u>Happy</u>	<u>Sad</u>	<u>Angry</u>
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 advertising.
- 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 –

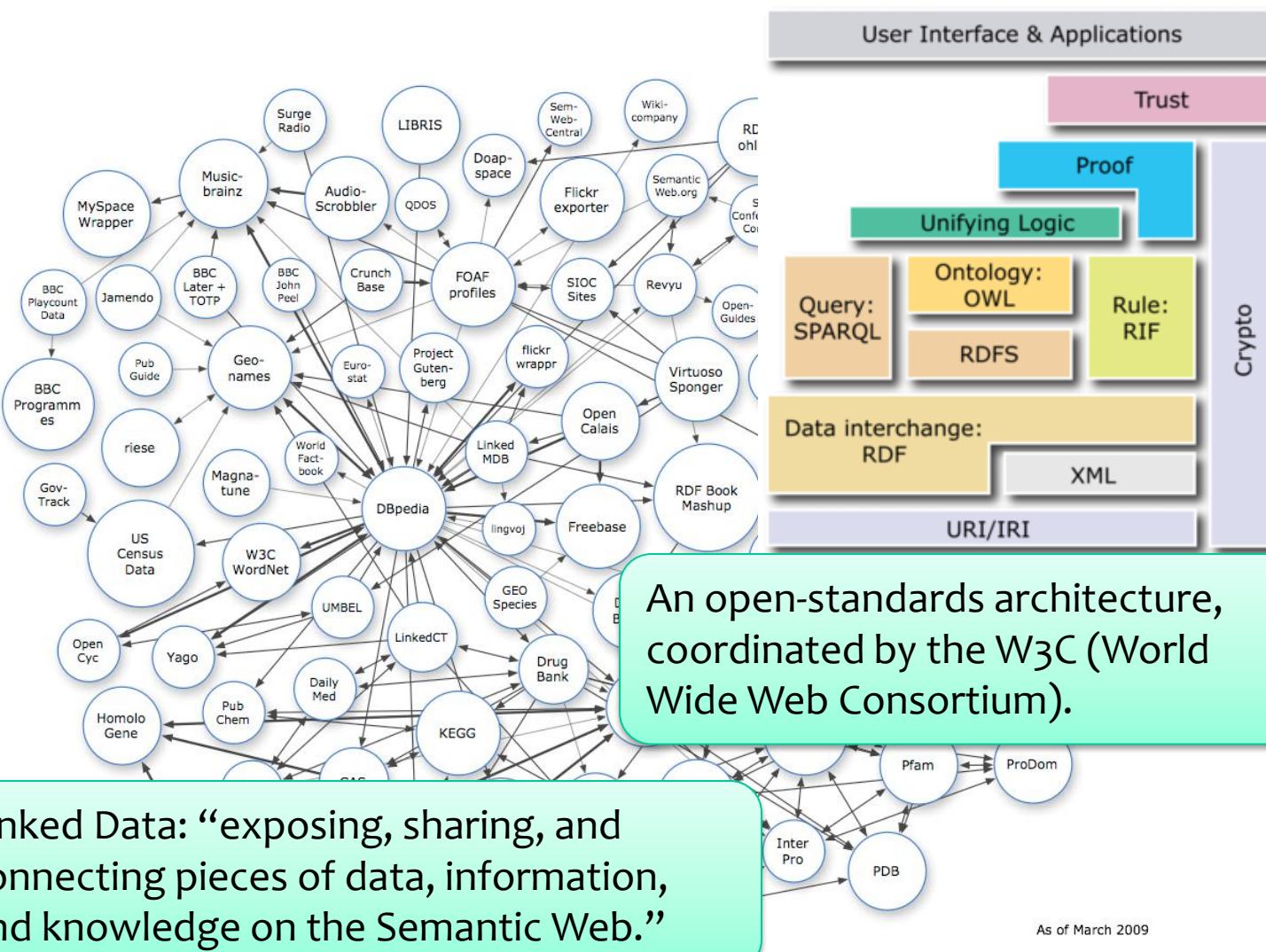
- Monitoring corporate communications
- Managing electronic stored information for production in event of litigation

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



# 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.

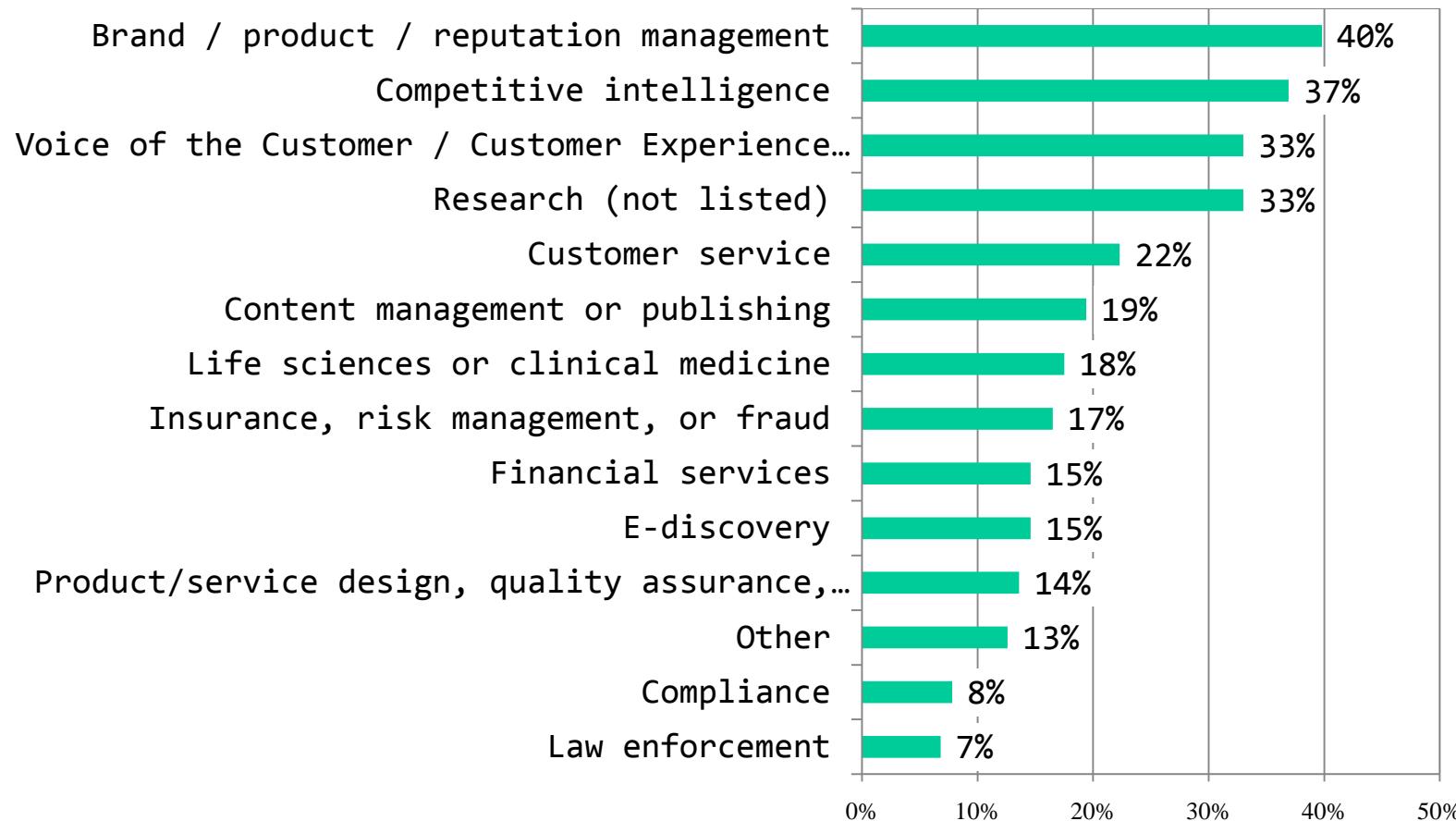
Last year, I published a study report, “Text Analytics 2009: User Perspectives on Solutions and Providers.”

<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%

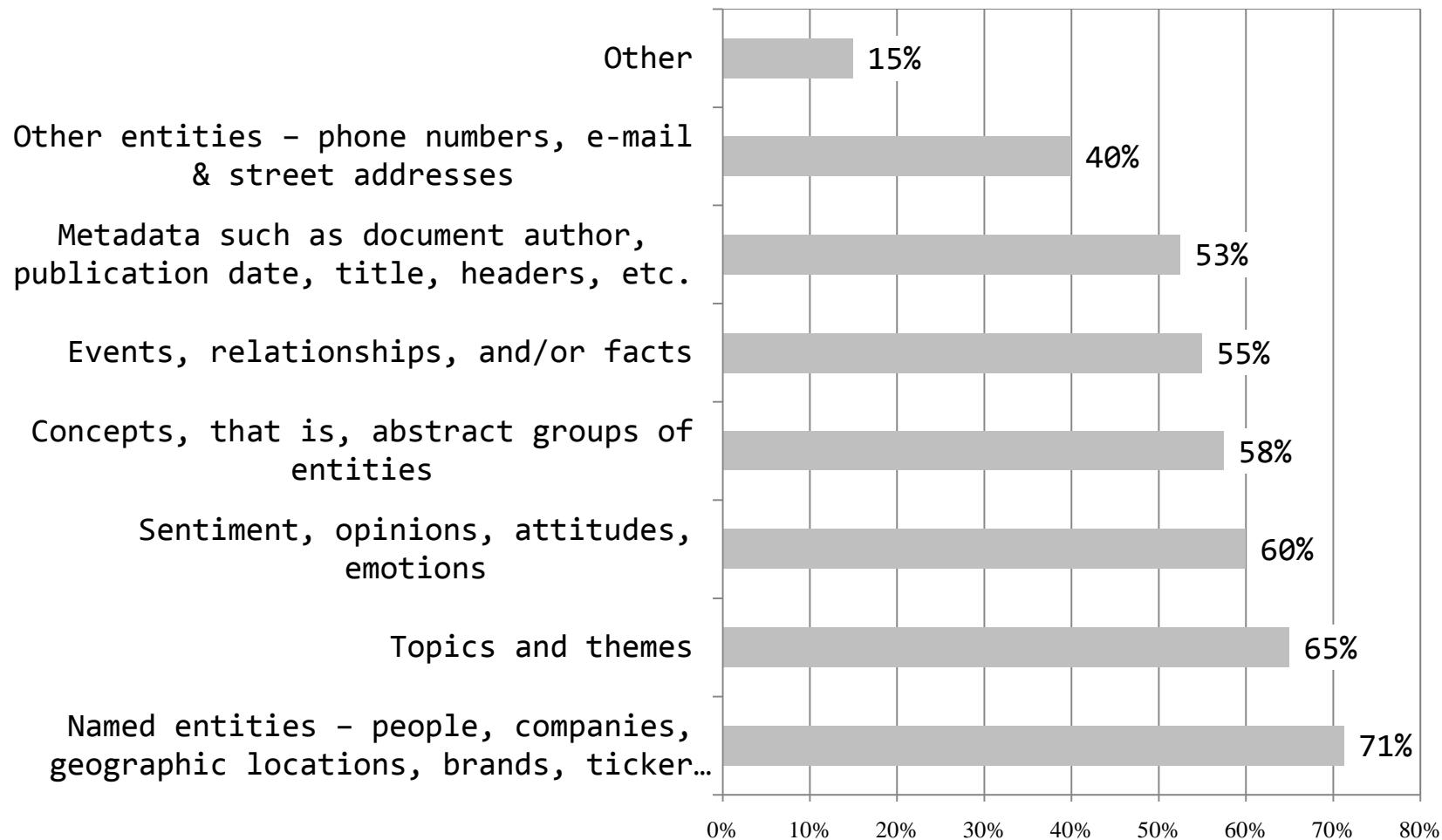
e-mail and correspondence 38%

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

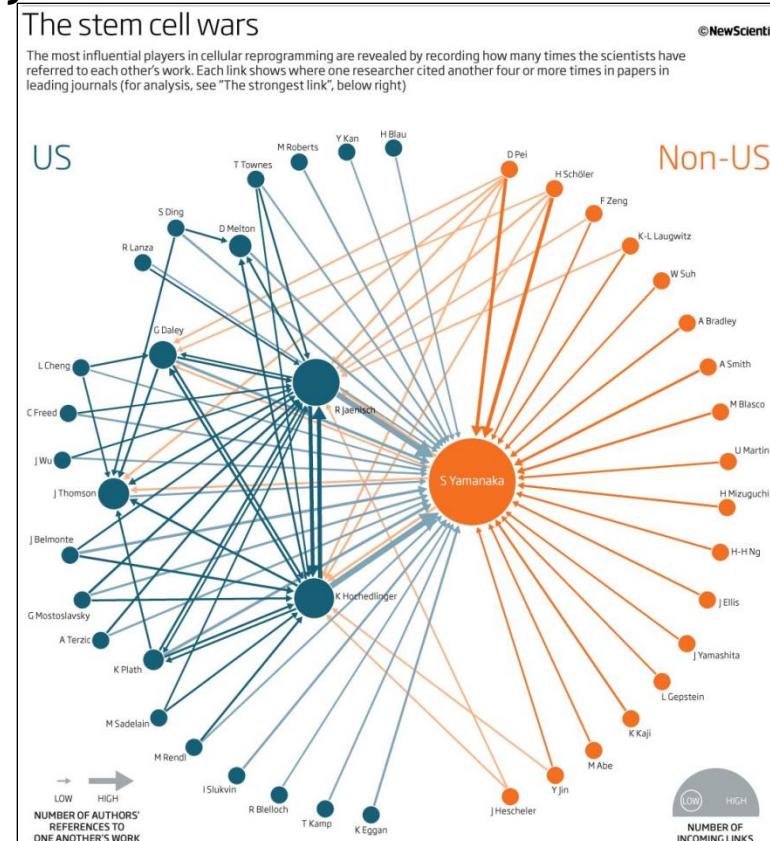
The illustrations that follow complement the ones we've seen...

# 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.”

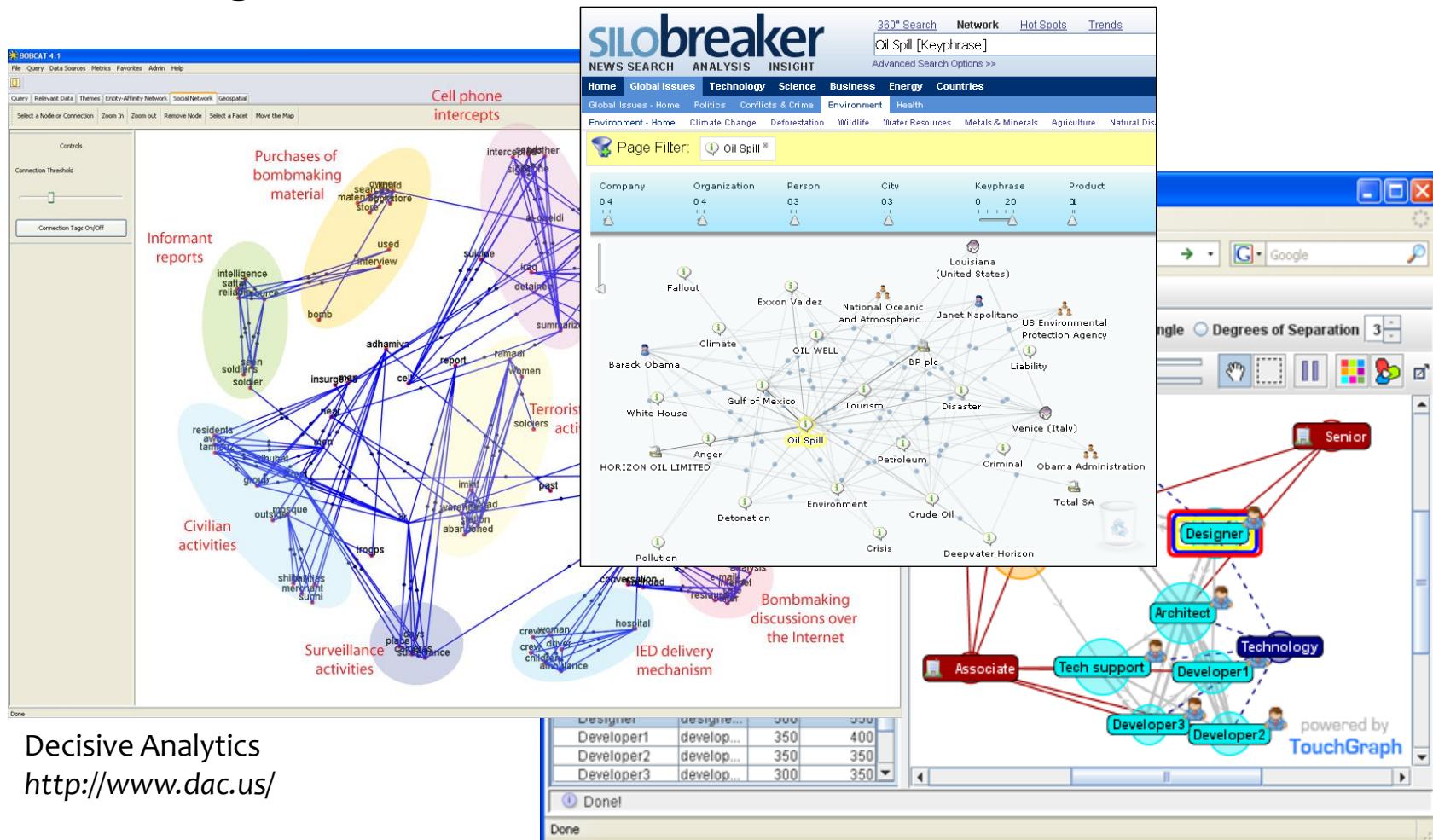
-- Peter Aldhous, “The stem cell wars: data, methods and results,” New Scientist, 2 June 2010.

<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

**IBM Many Bills: A Visual Bill Explorer - Enrolled Bills (51 to 100 of 114 bills) - Mozilla Firefox**

File Edit View History Bookmarks Tools Help

IBM Many Bills

Add bills to view them later or save them as a new collection! [Add all](#)

Enrolled Bills (51 to 100 of 114 bills) BY ANONYMOUS on Mar 17 2010  
Bills passed by both house and senate.

Share This Collection

Previous Page 1 2 3 Next Page

Category	Bill ID	Title	Author	Date	Description
International	S. 1707	ENR	H. R. 3614	ENR	Enhanced Partnership with Pakistan Act
Commerce	1. Short title; table of contents	Comm	2. Additional temporary extension	Tran	3. Extension of Airport and
Transportation	2. Definitions of terms	Inte	3. Findings of Congress	Budg	4. Extension of airport improvement
Armed Forces	3. Statement of principles	Inte	4. Statement of principles	Inte	5. Extension of expiring authorities
Labor And Industry	4. Priority in contracts and orders	Defe	5. Design of the Defense	Defe	6. Strengthening domestic Act of 1950
1. Short title; table of contents	1. Short title; table of contents	1. Short title; table of contents			
2. Reauthorization of Defense Production Act	2. Revisions to second-tier	2. Revisions to second-tier			
3. Declaration of policy	3. Third-Tier emergency unemployment	3. Third-Tier emergency unemployment			
4. Federal Aviation Administration Act of 1950 (50 U.S.C. § 402 of the Contract Work Act)	4. Federal Aviation Administration Act of 1950 (50 U.S.C. § 402 of the Contract Work Act)	4. Federal Aviation Administration Act of 1950 (50 U.S.C. § 402 of the Contract Work Act)	4. Federal Aviation Administration Act of 1950 (50 U.S.C. § 402 of the Contract Work Act)	4. Fourth-tier emergency unemployment	4. Fourth-tier emergency unemployment
5. Transfer of funds	5. Transfer of funds	5. Transfer of funds			
6. Expansion of modernization	6. Expansion of modernization	6. Expansion of modernization			
7. Treatment of	7. Expansion of productive capacity	7. Expansion of productive capacity	7. Expansion of productive capacity	7. Expansion of productive capacity	7. Expansion of productive capacity

[Feedback](http://manybills.researchlabs.ibm.com/collections/114)

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

# Visualization: Word Tree

many eyes

CHOOSE A DATA SET CHOOSE A VISUALIZATION CUSTOMIZE & PUBLISH

Welcome, SethGrimes | Logout

## Customizing Word Tree

Data set: Sentiment Analysis article by Seth Grimes (Version 1)

Your visualization will look like this:

Search **sentiment** Back Forward  Start  End Occurrence Order Clicks Will Zoom

18 hits

**sentiment**

**analysis**: Click: Zoom to "sentiment analysis" applications by seth grimes published : fel  
Shift-Click: Switch to "analysis" on survey data .  
mined from sources that include event and interview transcripts , presentation  
but also to understand influence networks , per aafia chaudhry's work , and  
not just general sentiment.

**extraction**: engine to determine which e  
from  
to  
to f  
measuring mark  
according to co  
[ understanding  
the trend , cle

**good**, **bad**, **good or bad**, **healthy**, **good for you**, **bad for you**, **is chocolate**, **are diets**, **poisonous to dogs**, **gluten free**, **bad for cats**, **vegan**, **bad for dogs**, **milk good for you**, **dangerous to dogs**

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

explore  
visualizations  
data sets  
comments  
topic centers  
my stuff  
my topic centers  
my watchlist  
my contributions  
messages to me

participate  
create visualization  
upload data set  
create topic center

learn more  
quick start  
visualization types  
data format & style  
about Many Eyes  
FAQ  
blog

contact Us  
contact  
report a bug

legal  
terms of use  
privacy

Popular Tags:  
Visualizations Data Sets

2008 2009 Obama

# Visualization: Phrase Net

Welcome, SethGrimes | Logout

# many eyes

explore  
visualizations  
data sets  
comments  
topic centers  
my stuff  
my topic centers  
my watchlist  
my contributions  
messages to me

participate  
create visualization  
upload data set  
create topic center

learn more  
quick start  
visualization types  
data format & style  
about Many Eyes  
FAQ  
blog

contact Us  
contact  
report a bug

legal  
terms of use  
privacy

Popular Tags:  
Visualizations Data Sets  
2008 2009 Obama

CHOOSE A DATA SET CHOOSE A VISUALIZATION CUSTOMIZE & PUBLISH

## Customizing Phrase Net

Data set: Sentiment Analysis article by Seth Grimes (Version 1)

Your visualization will look like this:

Select a phrase

- `word1 and word2`
- `word1 's word2`
- `word1 of the word2`
- `word1 the word2`
- `word1 a word2`
- `word1 at word2`
- `word1 is word2`
- `word1 [space] word2`

or enter your own

Filters

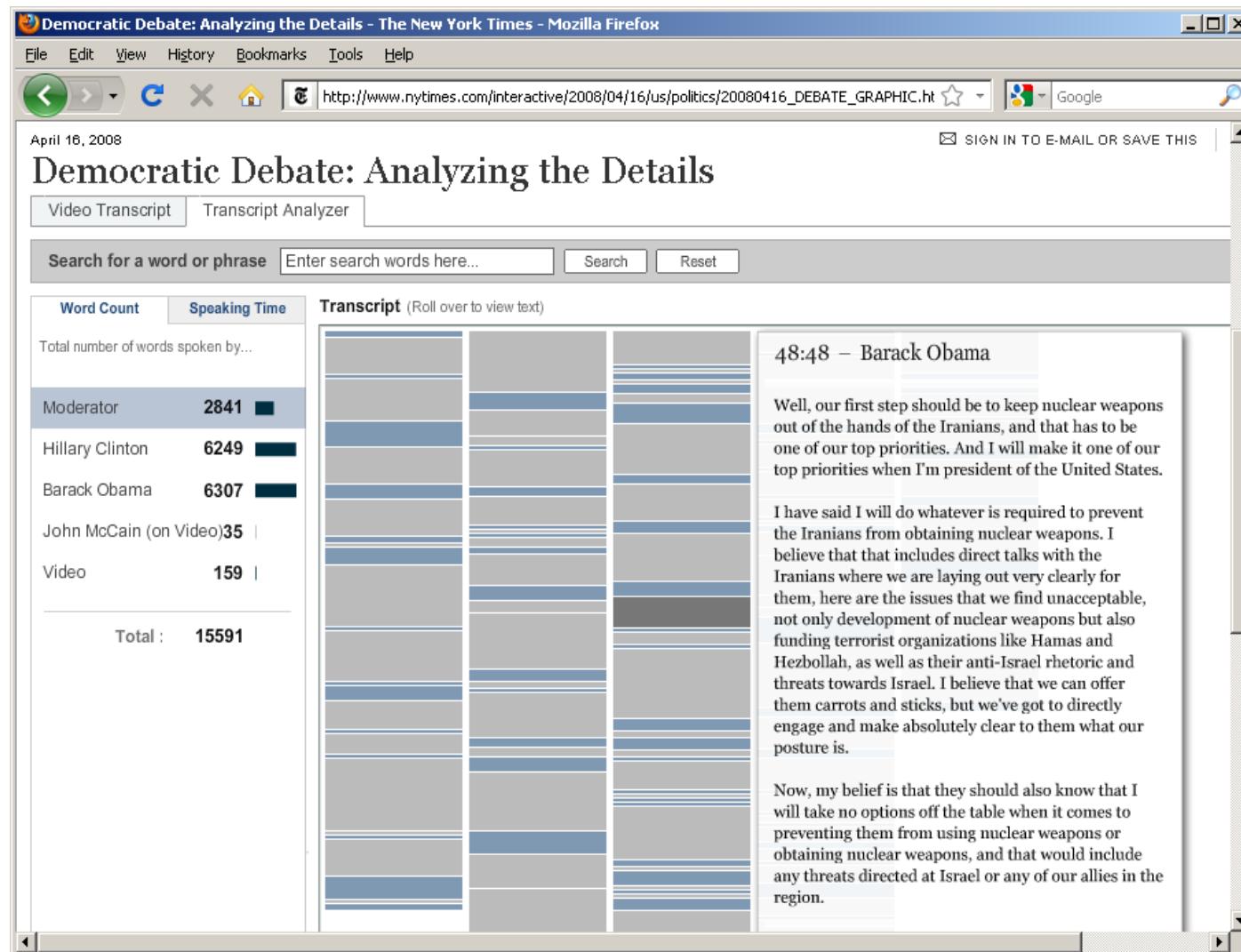
Show top:

Hide common words

Zoom

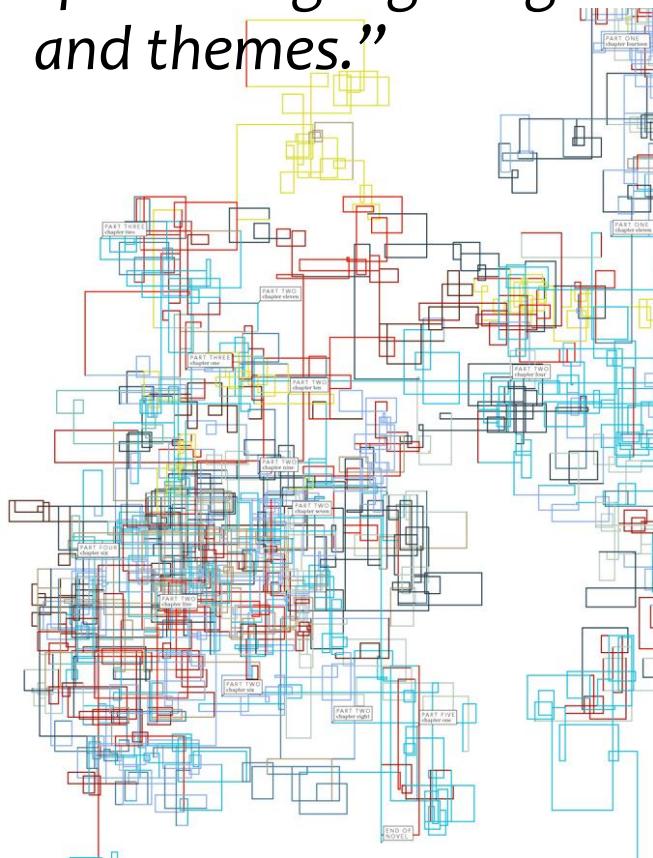
In Out Reset

# Visualization: Discourse Analysis

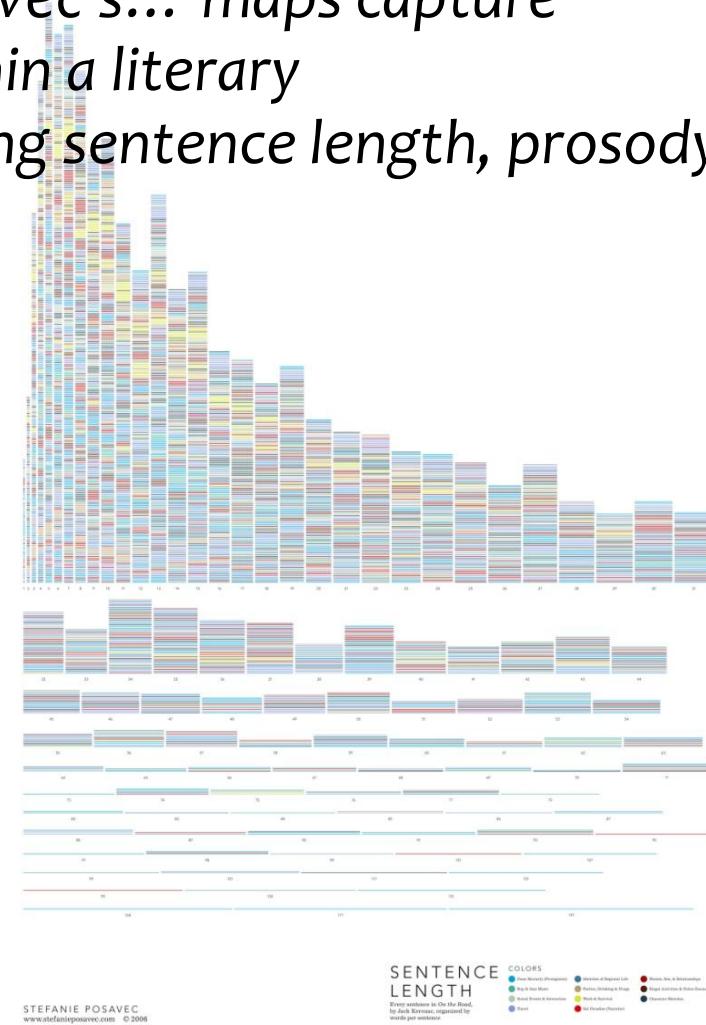


## Visualization: Text Statistics

Literary analysis: “Stefanie Posavec’s... maps capture regularities and patterns within a literary space... highlighting and noting sentence length, prosody and themes.”

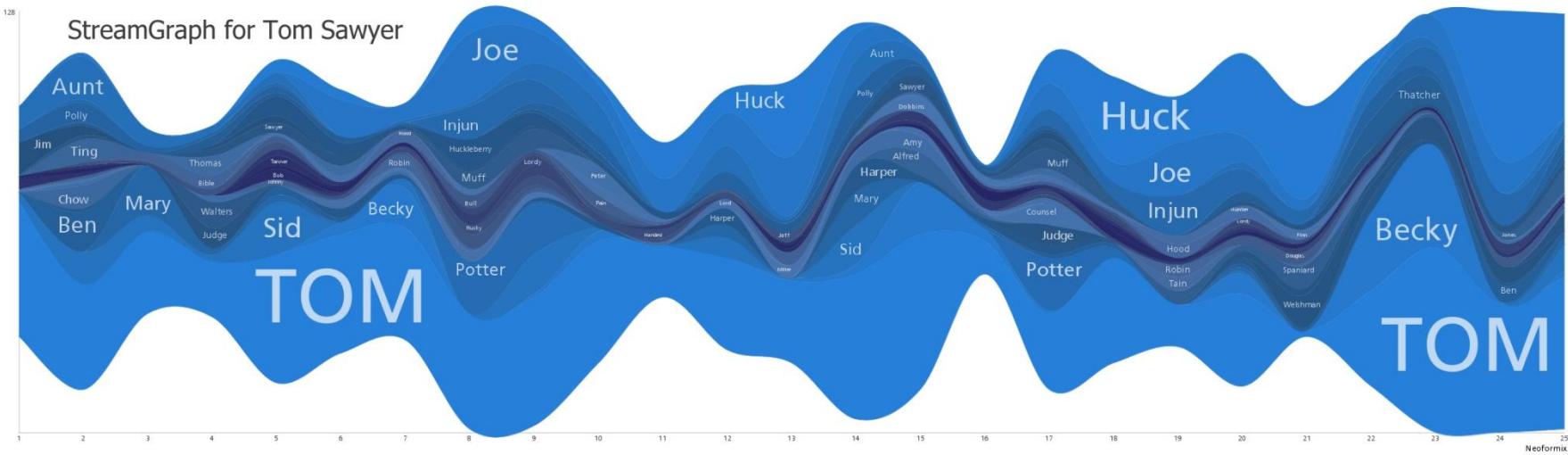


<http://www.notcot.com/archives/2008/04/stefanie-posave.php>



# 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