

Introduction to Text Mining and Semantics

Seth Grimes

--

President, Alta Plana

Alta Plana

NEHRU PROMISES BORDER DEFENSE

Warns Peiping India Would
Resist Further Advances
Into Her Territory

By PAUL GRIBES
Special to The New York Times.

NEW DELHI, India, Oct. 8.—Prime Minister Jawaharlal Nehru warned today that any further aggression by Communist China against India "will certainly be fully resisted."

"I hope it will not be necessary," he said at a crowded news conference. But he warned against "any kind of advance" by the Chinese from positions they now hold inside India's northern borders.

The Prime Minister asserted, "We do not intend to start military operations against any of these places at this stage, when we are dealing with this matter on a political level."

He said the Indian Government did not function "in an excited way with a club in hand" but instead sought to act "with determination, not with anger."

Mr. Nehru recalled that he had sent a letter to Premier Chou En-lai of Communist China Sept. 26, appealing to the Chinese to withdraw their forces so that the border disputes could be "amicably and peacefully settled."

The Prime Minister said he welcomed the "friendly message" sent by Mr. Chou in response to India's greetings on the tenth anniversary of the Communist regime Oct. 1. In this message, made public here last night, Mr. Chou stressed the "age-old friendship" between China and India and described the present difficulties as "merely an episode."

"The tone is friendly and, therefore, an improvement," Mr.

Nehru said. "How far that represents any basic change in the situation, I cannot say. I hope it does."

The Prime Minister, dressed in white homespun with his usual red rose in a buttonhole, answered questions for an hour and a quarter.

Prime Minister Nehru belittled a report that Premier Khrushchev, when visiting Los Angeles, told of the Soviet interception of a highly confidential message from President Eisenhower to Mr. Nehru. The Prime Minister said the only messages he had received from the United States President in the last two or three months were "friendly and formal letters—nothing secret."

Mr. Nehru also expressed the following views:

¶That he is opposed to disbanding the armistice control commissions in Laos, Cambodia, and Vietnam as long as it is possible to carry out the Geneva agreements of 1954 that ended the Indochinese war.

¶That President Charles de Gaulle's latest proposal to end the Algerian strife was "certainly a marked advance" from previous French suggestions, particularly since it "acknowledged the right of self-determination, which was the basic thing."

¶That he disapproved thoroughly of a recent "extraordinary" resolution of the Indian Communist party that would have India weaken considerably her stand on the border issue.

¶That the Soviet Union should be "warmly congratulated" for its rocket around the moon.

Mr. Nehru, who will be 70 years old Nov. 14, was asked whether "you feel a sense of fulfillment in your life or a sense of frustration."

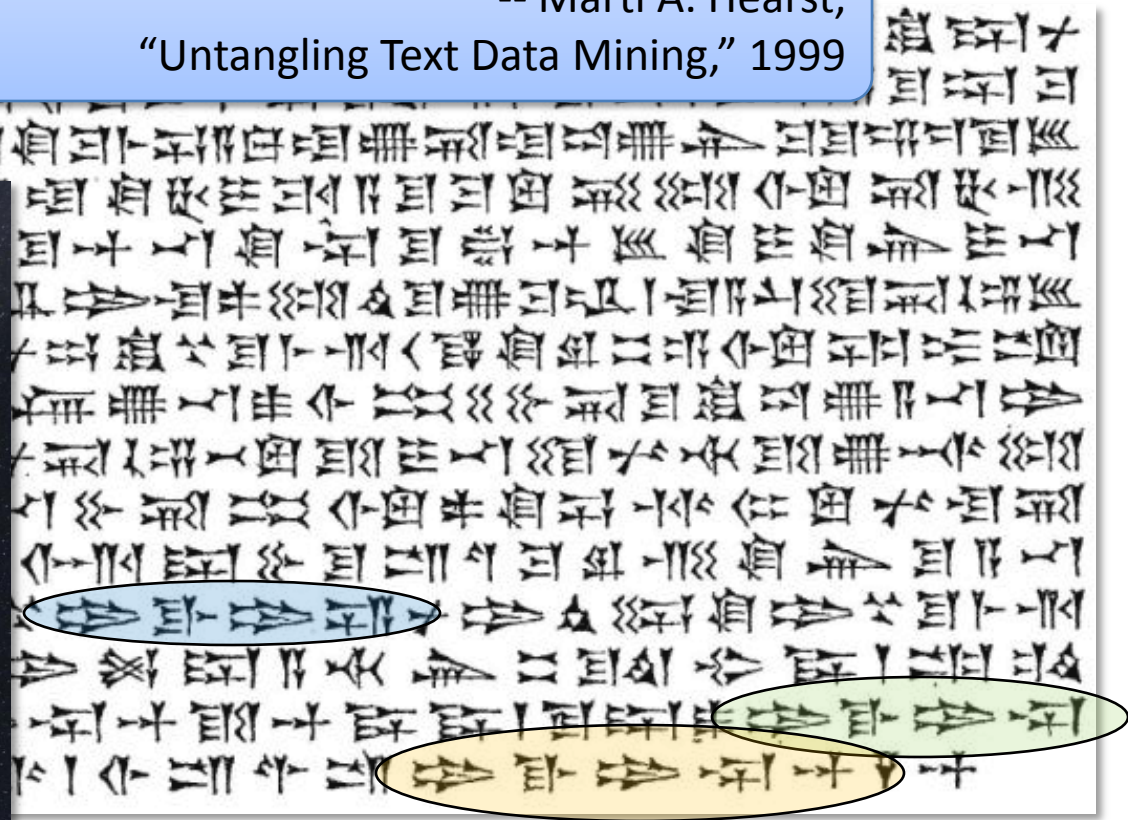
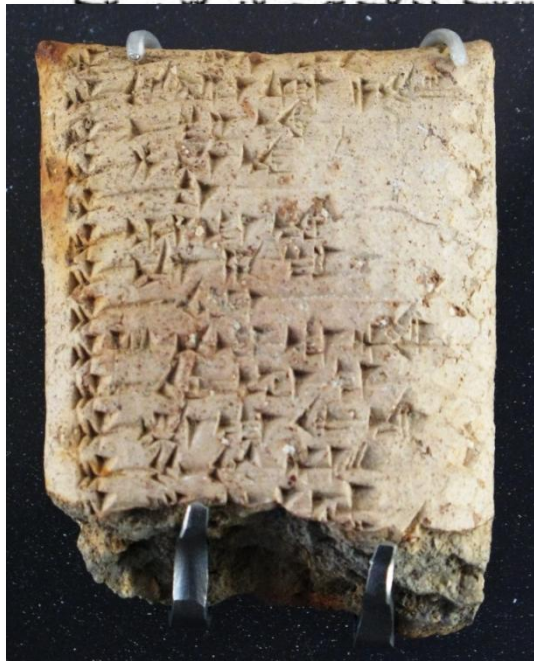
He replied:

"I have absolutely no sense of frustration in my life. I hope my face shows that. If the question is, have I achieved all I want to, my answer is no. But partial achievement comes from time to time."

From text to information

“Text expresses a vast, rich range of information, but encodes this information in a form that is difficult to decipher automatically.”

-- Marti A. Hearst,
“Untangling Text Data Mining,” 1999

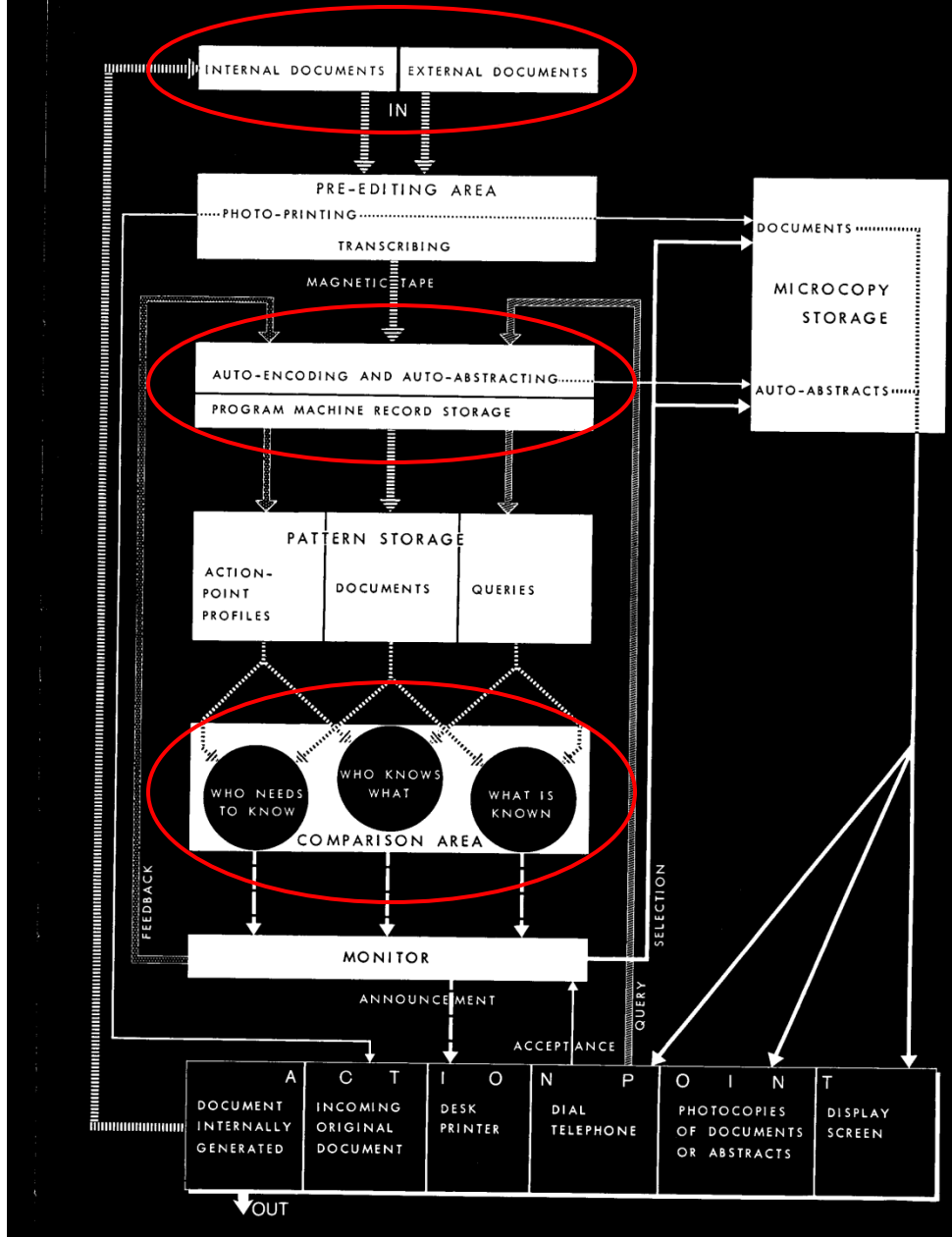


Alta Plana

Document input and processing

Knowledge management

Information extraction



Hans Peter Luhn, "A Business Intelligence System," IBM Journal, October 1958

Figure 1 A Business Intelligence System

Statistical analysis of content

“Statistical information derived from word frequency and distribution is used by the machine to compute a relative measure of significance.”

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

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

Total word occurrences in the document: 2326

Different words in document:

Total of different words 741

Less different common words 170

Different non-common words 571

Ratio of all word occurrences to different non-common words ~4:1

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

Total occurrences 478

Different words 39

Hans Peter Luhn,

“The Automatic Creation of Literature Abstracts,”

IBM Journal, April 1958

Statistical analysis limitations

“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

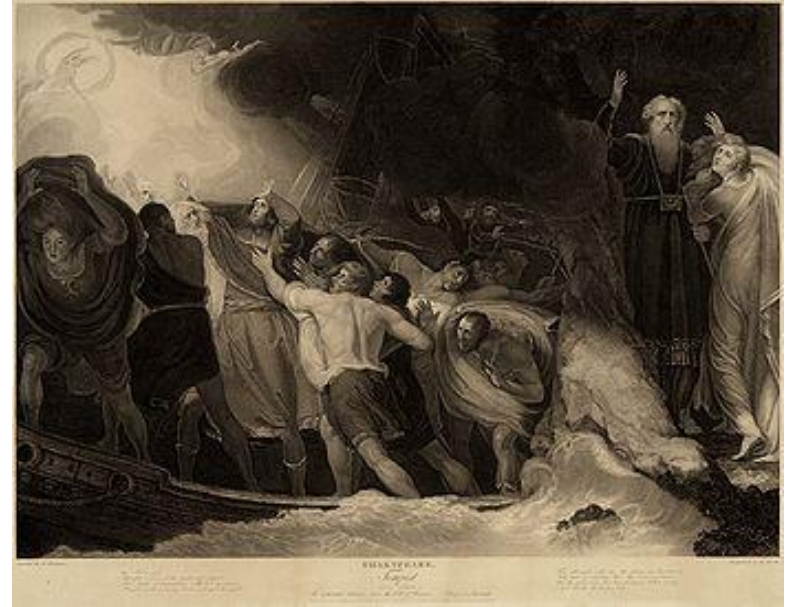
Unknown = new

Miranda:

O wonder!
How many goodly
creatures are there here!
How beauteous mankind
is! O brave new world, That
has such people in't!

Prospero:

'Tis new to thee.



The shipwreck in *The Tempest*, Act I, Scene 1, in an 1797 engraving based on a painting by George Romney.

Semantic links

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 (tranquilizers, for instance) have made long-institutionalized people amenable to therapy.

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

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

Blood May Tell

Dr. Felix Marti-Ibanez and three brothers, Dr. Mortimer D. Sackler,

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

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

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

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

Chemicals now available should be used not only to bring relief to the mentally sick but also to uncover

the biological mechanisms of the disease processes themselves. "Only then will the metabolic era mature and bring to fruition man's long hoped for salvation from the ravages of mental disease," they reported.

Chemistry of the Brain

At the psychologist's meeting here, a technique for tracing electrical activity in specific portions of the animal brain was described by researchers of California reported in cat electric animals which

In this reported sequence its various the brain may be located. Furthermore, the electrical pathways so traced out can be blocked temporarily by the use of chemicals. This poses new possibilities for studying brain and side the California sized.

The country have peptic courage for men that kn ary field last we Washin

This the Na Health informati Literatu and an bers ca technic People vited to or other letters have in Informa Silver



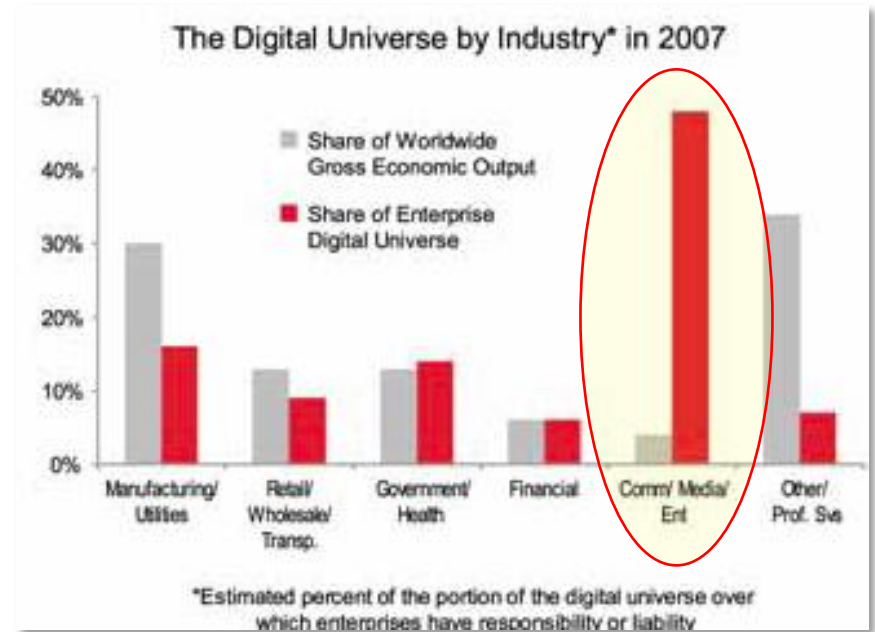
Sentiment analysis



Digital content universe

“The broadcast, media, and entertainment industries garner about 4% of the world’s revenues but already generate, manage, or otherwise oversee 50% of the digital universe.”

Approximately 70% of the digital universe is created by individuals.



“The Diverse and Exploding Digital Universe,”
(IDC, 2008)

The “Unstructured Data” Challenge

- Web sites, news & journal articles, images, video.
- Blogs, forum postings, and social media.
- E-mail, Contact-center notes and transcripts; recorded conversation.
- Surveys, feedback forms, warranty & insurance claims.
- Office documents, regulatory filings, reports, scientific papers.
- And every other sort of document imaginable.

Is Search up to the job?

Search results

How are the quality, value & authority of search results?

The screenshot shows a Google search interface with the query "Is the Hilton New York a good hotel?". The search results are categorized as "Web" and show "Results 1 - 10 of about 1,590,000 for Is the Hilton New York a good hotel?".

Sponsored Links:

- Hilton Hotels New York** (Sponsored Link): www.Hilton.com Book direct on Hilton.com For Our Best Rates, Guaranteed!
- New York Best Hotels** (Sponsored Link): Wholesale Rates at all of the Best Luxury Hotels in New York City. NewYorkCityLuxuryHotels.com

Organic Results:

- New York Hotels - Hilton New York Hotel NYC**: Hilton New York City Hotel - Rockefeller Center Hotel - NYC Hotels ... The Hilton New York Hotel is a sophisticated Manhattan hotel conveniently located in ... www1.hilton.com/en_US/hilton/hotel/NYCNHHH-Hilton-New-York-New-York/index.do-77k - Cached - Similar pages
 - Accommodations
 - Directions
 - Reservations
 - Services & Amenities
 - Video Tour
 - Photos
 - Dining
 - Employment[More results from hilton.com »](#)
- Hotels by Hilton - Hotel Reservations, Deals, and Room Rates**: Find and book hotel rooms online at the official site of Hilton brand hotels. ... Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina ... www1.hilton.com/ - 87k - Cached - Similar pages
- Hilton New York (New York City, NY) - Hotel Reviews - TripAdvisor**: Hilton New York: Save up to 50% off Hotels. Everyday Smart Deal! ... I stayed in this hotel over a weekend and got a good deal on Priceline for \$110 a night ... www.tripadvisor.com/Hotel_Review-g60763-d611947-Reviews-Hilton_New_York-New_York_City-New_York.html - 116k - Cached - Similar pages

HotelTs
opinion

GuestTs
opinionE
about
Priceline

Who profits
from search?

From Web 1.0 to Web 2.0

How can we do better?

“We have many of the tools in place -- from Web 2.0 technologies...”

“The Diverse and Exploding Digital Universe,”
(IDC, 2008)

Web 2.0

“Web 2.0 is the business revolution in the computer industry caused by the move to the Internet as a platform.”

-- Tim O'Reilly, 2004

“[A] move from personal websites to blogs and blog site aggregation, from publishing to participation,... an ongoing and interactive process... to links based on tagging.”

-- Terry Flew, “New Media: An Introduction,” 2008

Web 2.0 is dynamic, personalized, interactive, collaborative.

Economically speaking

How can we do better?

“We have many of the tools in place -- from Web 2.0 technologies... to unstructured data search software and the Semantic Web -- to tame the digital universe. Done right, we can turn information growth into economic growth.”

“The Diverse and Exploding Digital Universe,”
(IDC, 2008)

Text mining: from information to intelligence

Text mining enables smarter search that better responds to user goals, e.g., answers –

The image displays several overlapping search engine results pages (SERPs) for different queries, demonstrating how text mining can provide more relevant and intelligent answers.

- Live Search (Top):** Query: "population peru". Results include "Peru Population, total: 29,041,593" and "2008 estimate · United States C".
- Google (Middle-Left):** Query: "map massachusetts". Shows a map of Massachusetts and surrounding areas.
- Google (Middle-Right):** Query: "illinois unemployment rate". Results include "Unemployment rate, Illinois" with a line graph showing the rate from 2004 to 2009, peaking at 9.4% in March 2009. Source: U.S. Bureau of Labor Statistics. Includes a sponsored link for "Illinois Unemployment Benefits In Illinois".
- Search (Bottom-Left):** Query: "20*30". Results include "20*30 = 600" (Yahoo! Shortcut), "Renault 20/30 - Wikipedia, the free encyclopedia" (describing executive cars), and "Active 20-30 USA & Canada" (for young adults).
- Search (Bottom-Right):** Query: "30 20". Results include "Save on 30 20 and More! Buy, Bid, or Make an Offer now." (eBayMotors.com) and "Find Bargain Prices On 20 30." (BizRate.com).

From Web 2.0 to Web 3.0

For even better findability:

“The Semantic Web is a web of data, in some ways like a global database.”

-- Tim Berners-Lee, 1998


Web 3.0 is Web 2.0 + the Semantic Web + semantic tools.

Recurring themes:

- Semantically enriched content.
- Linked Data.
- Context sensitive.
- Location aware.

Examples


Web [Images](#) [Video](#) [Maps](#) [News](#) [Shopping](#) [Gmail](#) [more](#) [Sign in](#)

 [Advanced Search](#)
[Preferences](#)

Web [Show options...](#) Results 1 - 10 of about 21,900,000 for [Gordon Brown](#). (0.20 seconds)

[Gordon Brown](#) - Wikipedia, the free encyclopedia
Gordon Brown was born in Govan, Glasgow, Scotland. His father was John (1914–1998), a minister of the Church of Scotland and a strong ...
en.wikipedia.org/wiki/Gordon_Brown - [Cached](#) - [Similar](#)


[Daniel Hannan MEP: The devalued Prime Minister of a devalued](#)
 3 min 29 sec - Mar 24, 2009 - ★★★★★
European Parliament speech of 26/03/09. Daniel Hannan is a MEP for the South East of England and author of The Plan: T...
www.youtube.com/watch?v=94IW6Y4tBXs

[Gordon Brown](#)
 1 min 49 sec - Oct 8, 2008 - ★★★★★
Same old jokes, now in spectacular Jerk-O-Vision.
www.youtube.com/watch?v=6QapZi2cLQQ

[News results for Gordon Brown](#)
 [David Miliband revives speculation over Gordon Brown's leadership](#)
David Miliband has described Alan Johnson as "the leading contende...
[Brown](#), fuelling speculation that Mr [Brown](#) will face a challenge for...
[Telegraph.co.uk](#) - [2754 related articles](#) >
[Sarah Brown poised to ride to husband Gordon's rescue again -](#)
[Telegraph.co.uk](#) - [4045 related articles](#) >
[Support For Gordon Brown Waning](#) - CBS News - [711 related artic](#)

[gordonbrown.net](#)
Gordon Brown was shooting IMAX for MacGillivray Freeman Films Producti...
Into Amazing Caves" when Michael [Brown](#) shot this picture from a platform ...
www.gordonbrown.net/ - [Cached](#) - [Similar](#)

[Gordon Brown News | Full Coverage - Yahoo! News UK](#)
Complete [Gordon Brown](#) news coverage from Yahoo! News UK. Find videos in-depth [Gordon Brown](#) commentary in our full coverage news section.
uk.news.yahoo.com/fc/gordon-brown.html - [Cached](#) - [Similar](#)

 **Newsift**
FROM THE FINANCIAL TIMES GROUP

New Search Saved Searches Search History Help Feedback: Take our survey

Search Term: Gordon Brown

Close Suggestions

Enter term	Business Topic	Organization	Place	Person	Theme
OR	Monetary Volatility	Labour Party	London	Gordon Brown	Expenses Scandal
Select linked term to refine search	Labor Force And Unions	Chambre Des Communes	United Kingdom	David Cameron	European Elections
	Litigation And Settlement	Browns	Europe	Alistair Darling	Expenses Claims
	Central Banks And Monetary Policy	European Union	England	Tony Blair	Conservative Leader
	Trade Policy	Liberal Democrats	France	Barack Obama	Expense Claims
		European Parliament	Germany	James Purnell	Cabinet Reshuffle
			Iraq	Jacqui Smith	Expenses System

Sentiment:

1-10 of 12702 Articles: Frequent terms above

Sort by: **Relevance** | Date
Timeframe: Apr 13, 2009 to Jun 12, 2009

[Brown's ratings fall again, Labour gains](#) @
Daily Mail & Guardian, June 12, 2009
British Prime Minister Gordon Brown's personal ratings have fallen again after he survived attempts to force him out of office, although his ruling Labour Party's standing has improved, an opinion poll said on Friday... With a parliamentary election less than a year away, the centre-right opposition Conservative Party is in...

SPONSOR ADVERTISEMENT

- Get news direct to your inbox
- Choose from over 40 daily updates
- Tailor news alerts by sector, industry or keyword
- Read up to 10 articles per month

[Register now](#)

We live in FINANCIAL TIMES®

Text mining / analytics enables Web 3.0 and the Semantic Web.

- Automated content categorization and classification.
- Text augmentation: metadata generation, content tagging.
- Information extraction to databases.
- Exploratory analysis and visualization.

Technical concepts:

- Linked Data
- Microformats, RDF, SPARQL
- OWL

Text mining: users' perspective

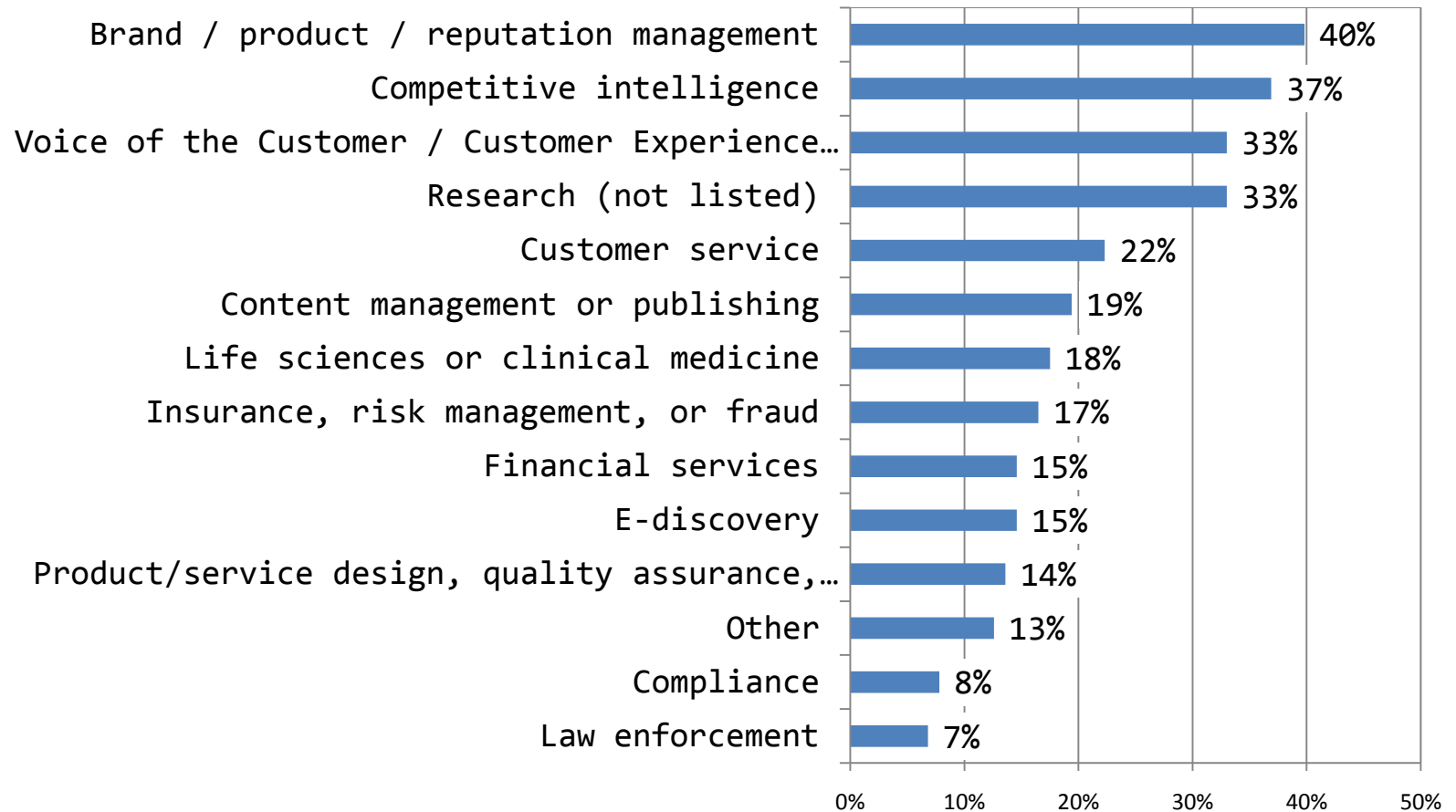
I recently published a study report, “Text Analytics 2009: User Perspectives on Solutions and Providers.”

I estimated a \$350 million global market in 2008, up 40% from 2007.

I relayed findings from a survey that asked...

Primary applications

What are your primary applications where text comes into play?



Analyzed textual information

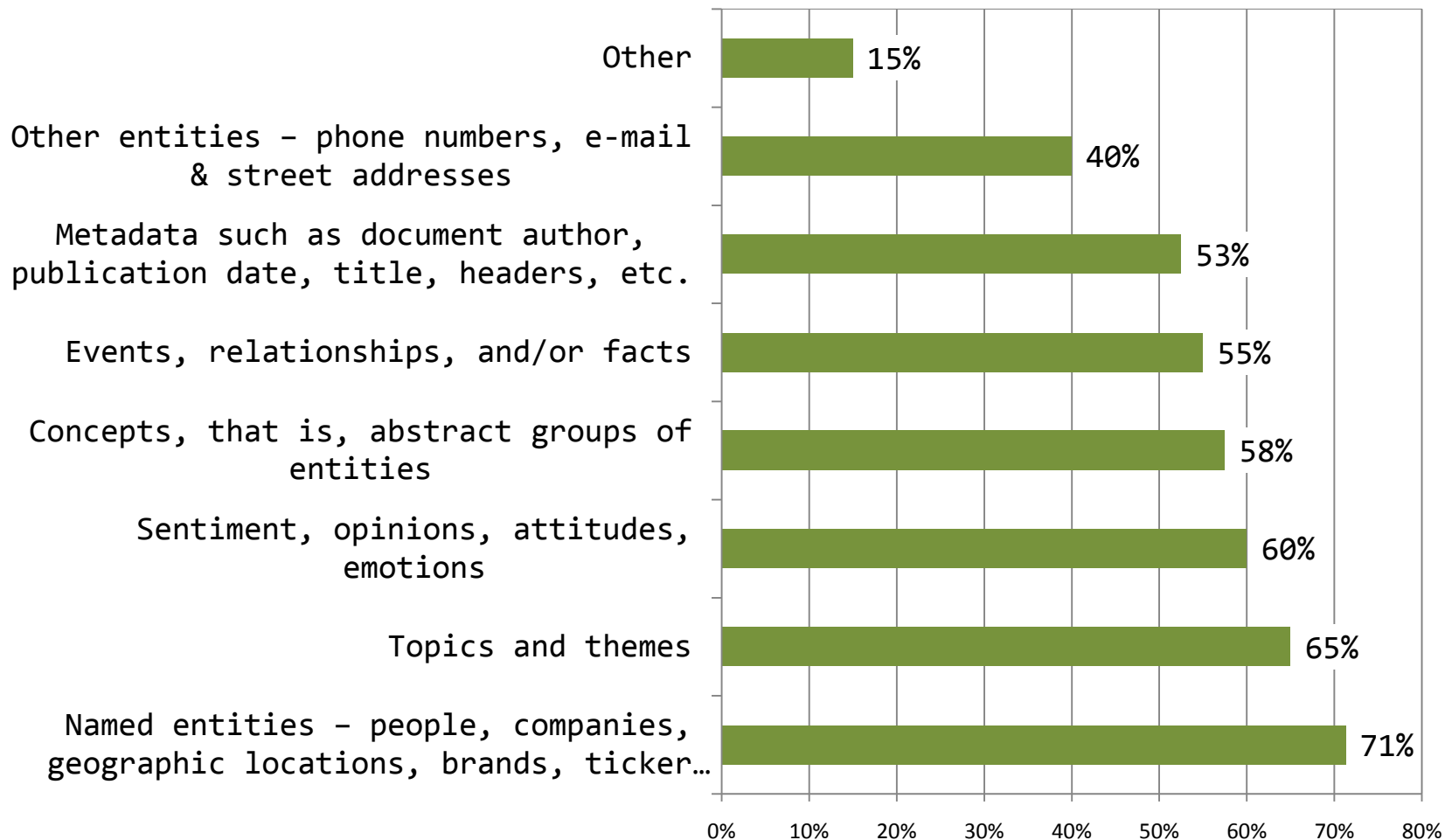
What textual information are you analyzing or do you plan to analyze?

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:



Overall satisfaction

Please rate your overall experience – your satisfaction – with text mining.

