Text Analytics for Dummies

Seth Grimes

Alta Plana Corporation

301-270-0795 -- *http://altaplana.com*

Text Analytics Summit 2008

Workshop

June 15, 2008

Introduction

Seth Grimes –

Principal Consultant with Alta Plana Corporation.

Contributing Editor, *IntelligentEnterprise.com*.

Channel Expert, *B-Eye-Network.com*.

Founding Chair, Text Analytics Summit, *textanalyticsnews.com*.

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

I am not paid to promote any vendor.

Perspectives

Perspective #1: You're a business analyst or other "end user."

- You have lots of text, and you want an automated way to deal with it.

Perspective #2: You work in IT.

- You support end users who have lots of text.

Perspective #3: Other?

- You just want to learn about text analytics.

Perspectives

Perspective #1a, 2a: Extending analysis.

You want to extend an existing business intelligence (BI) / data-mining initiative to encompass information from textual sources.

Perspective #1b, 2b: New to analysis.

You don't do traditional data analysis (yet).

Perspectives

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 #2 and enables #3 & #4.

Text analytics can be automated or interactive.

Key Message -- #1

If you are not analyzing text – if you're analyzing only transactional information – you're missing opportunity or incurring risk…

“Industries such as travel and hospitality and retail live and die on customer experience.” – *Clarabridge CEO Sid Banerjee*

This is the “Unstructured Data” challenge

Key Message -- #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.

Text Analytics is an answer to the “Unstructured Data” challenge

Key Message -- #3

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.

Let’s deal with text-BI first. Here's an image and a quotation from a 1958 paper introducing BI as a method for processing documents and extracting knowledge...

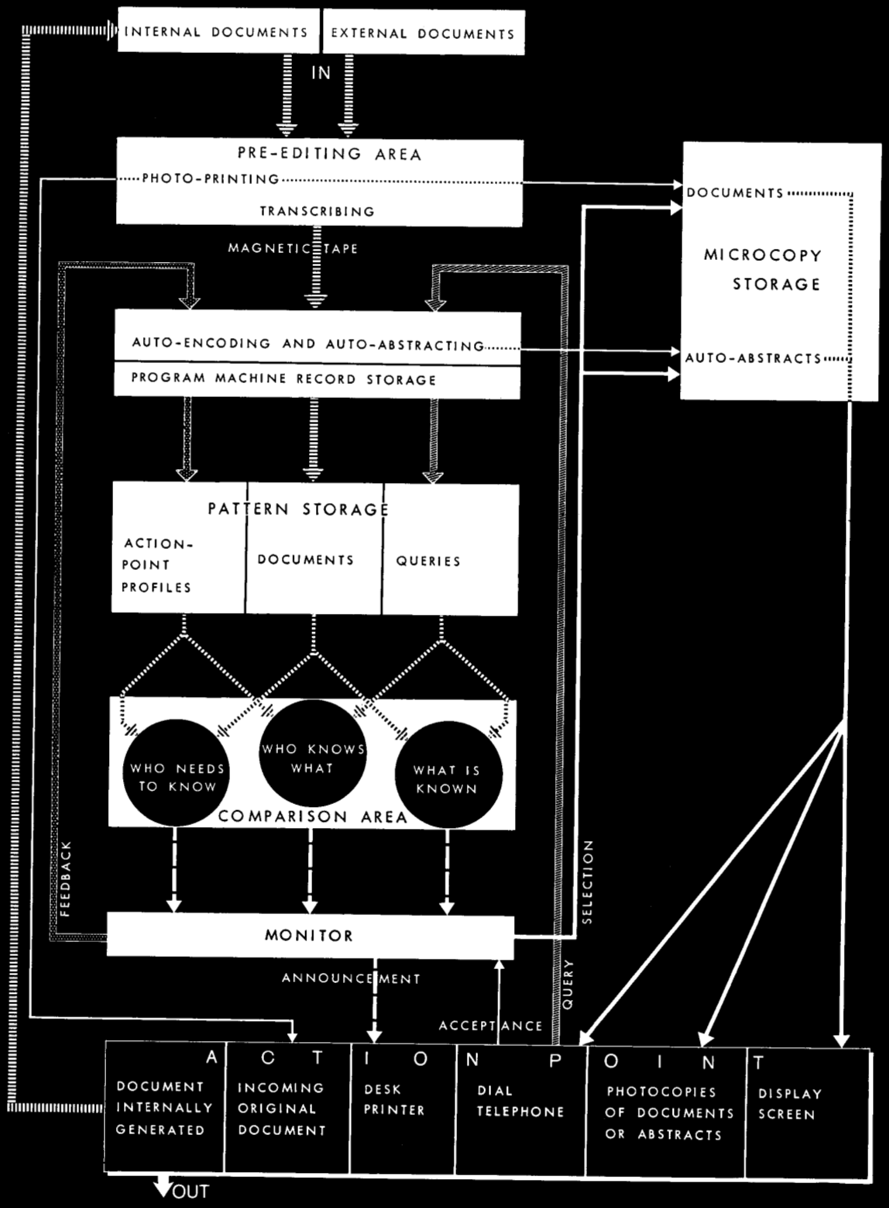


Figure 1 **A Business Intelligence System**

Business Intelligence

What is business intelligence (BI)?

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 communication facility serving the conduct of a business (in the broad sense) may be referred to as an intelligence system. The notion of intelligence is also defined here, in a more general sense, 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*

Why does BI not focus on textual documents?

The "Unstructured Data" Challenge

"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, former CTO of enterprise-search vendor Verity (now part of Autonomy)*

That's where BI operates, on data in a relational table that originated in transactional systems.

Yet it's a truism that 80% of enterprise information is in "unstructured" form.

The "Unstructured Data" Challenge

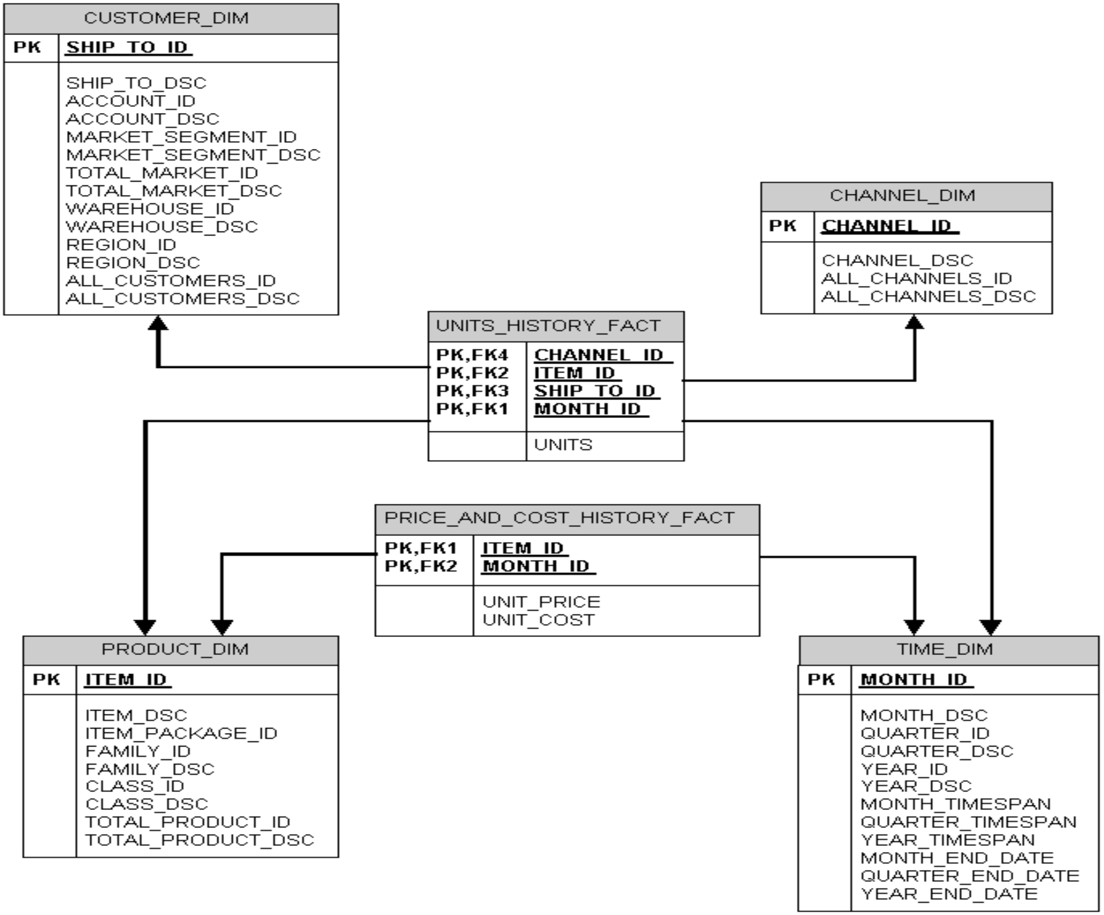
Traditional BI feeds off:

```
"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",6,7961,3892,4069,384,183,201,907,472
50,19,1,"Iowa","Adair County",7,7875,3855,4020,366,179,187,871,454
50,19,1,"Iowa","Adair County",8,7795,3817,3978,343,162,181,841,439
50,19,1,"Iowa","Adair County",9,7714,3777,3937,338,159,179,805,417
```

The "Unstructured Data" Challenge

Traditional BI feeds off:

```
"SUMLEV","STATE","COUNTY","STNAME",
50,19,1,"Iowa","Adair County",1,824
50,19,1,"Iowa","Adair County",2,824
50,19,1,"Iowa","Adair County",3,821
50,19,1,"Iowa","Adair County",4,809
50,19,1,"Iowa","Adair County",5,800
50,19,1,"Iowa","Adair County",6,796
50,19,1,"Iowa","Adair County",7,787
50,19,1,"Iowa","Adair County",8,779
50,19,1,"Iowa","Adair County",9,771
```



CUSTOMER_DIM
PK SHIP_TO_ID
SHIP_TO_DSC
ACCOUNT_ID
ACCOUNT_DSC
MARKET_SEGMENT_ID
MARKET_SEGMENT_DSC
TOTAL_MARKET_ID
TOTAL_MARKET_DSC
WAREHOUSE_ID
WAREHOUSE_DSC
REGION_ID
REGION_DSC
ALL_CUSTOMERS_ID
ALL_CUSTOMERS_DSC

CHANNEL_DIM
PK CHANNEL_ID
CHANNEL_DSC
ALL_CHANNELS_ID
ALL_CHANNELS_DSC

UNITS_HISTORY_FACT
PK,FK4 CHANNEL_ID
PK,FK2 ITEM_ID
PK,FK3 SHIP_TO_ID
PK,FK1 MONTH_ID
UNITS

PRICE_AND_COST_HISTORY_FACT
PK,FK1 ITEM_ID
PK,FK2 MONTH_ID
UNIT_PRICE
UNIT_COST

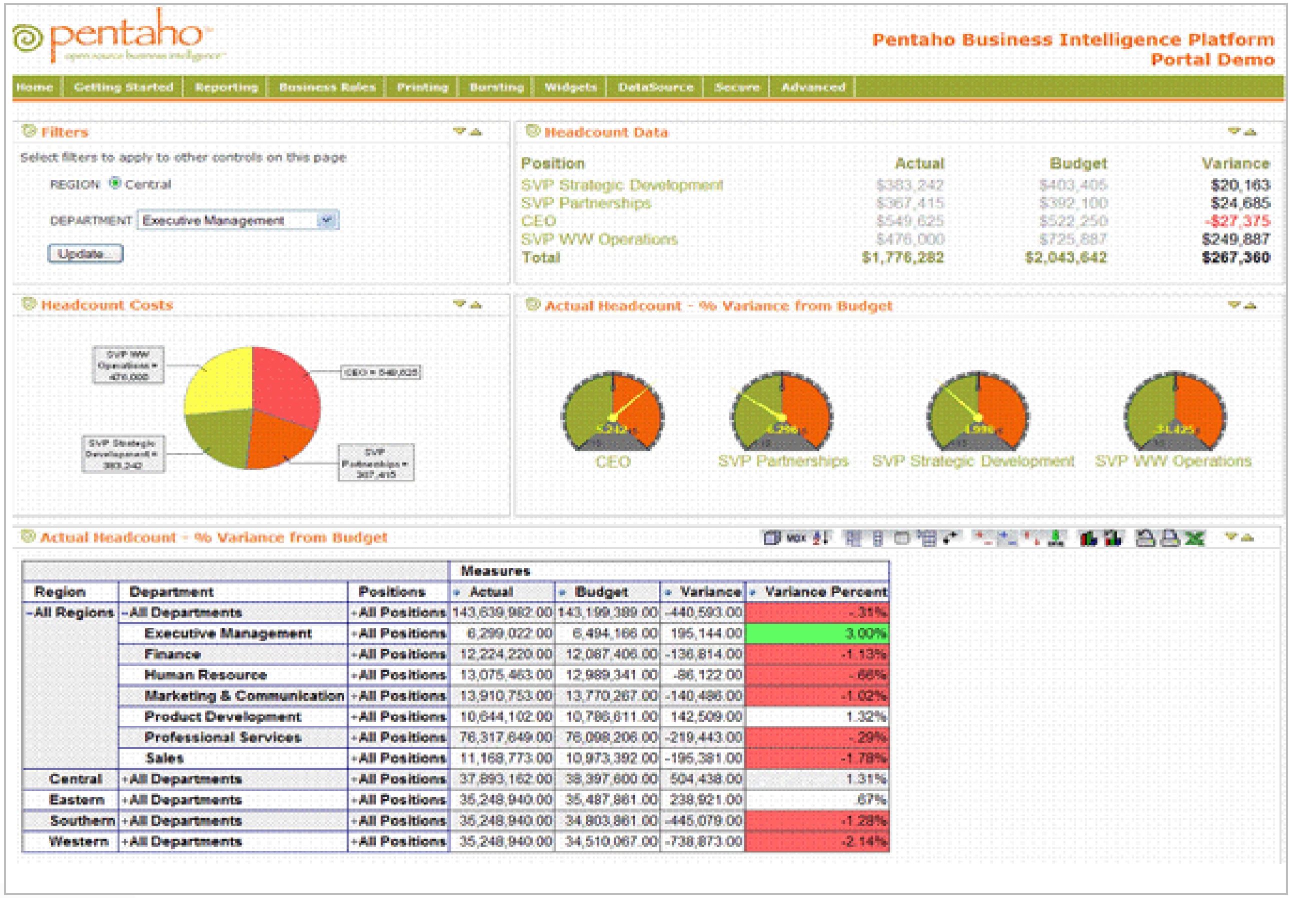
PRODUCT_DIM
PK ITEM_ID
ITEM_DSC
ITEM_PACKAGE_ID
FAMILY_ID
FAMILY_DSC
CLASS_ID
CLASS_DSC
TOTAL_PRODUCT_ID
TOTAL_PRODUCT_DSC

TIME_DIM
PK MONTH_ID
MONTH_DSC
QUARTER_ID
QUARTER_DSC
YEAR_ID
YEAR_DSC
MONTH_TIMESPAN
QUARTER_TIMESPAN
YEAR_TIMESPAN
MONTH_END_DATE
QUARTER_END_DATE
YEAR_END_DATE

It runs off:

The "Unstructured Data" Challenge

Traditional BI produces:


pentaho
open source business intelligence
Pentaho Business Intelligence Platform
Portal Demo
Home
Getting Started
Reporting
Business Rules
Printing
Bursting
Widgets
DataSource
Secure
Advanced
Filters
Headcount Costs
SVP WW Operations = 476,000
CEO = 549,625
SVP Strategic Development = 383,242
SVP Partnerships = 367,415
Actual Headcount - % Variance from Budget
CEO
SVP Partnerships
SVP Strategic Development
SVP WW Operations

Filters

Select filters to apply to other controls on this page

REGION ● Central

DEPARTMENT Executive Management

Update

Headcount Data

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

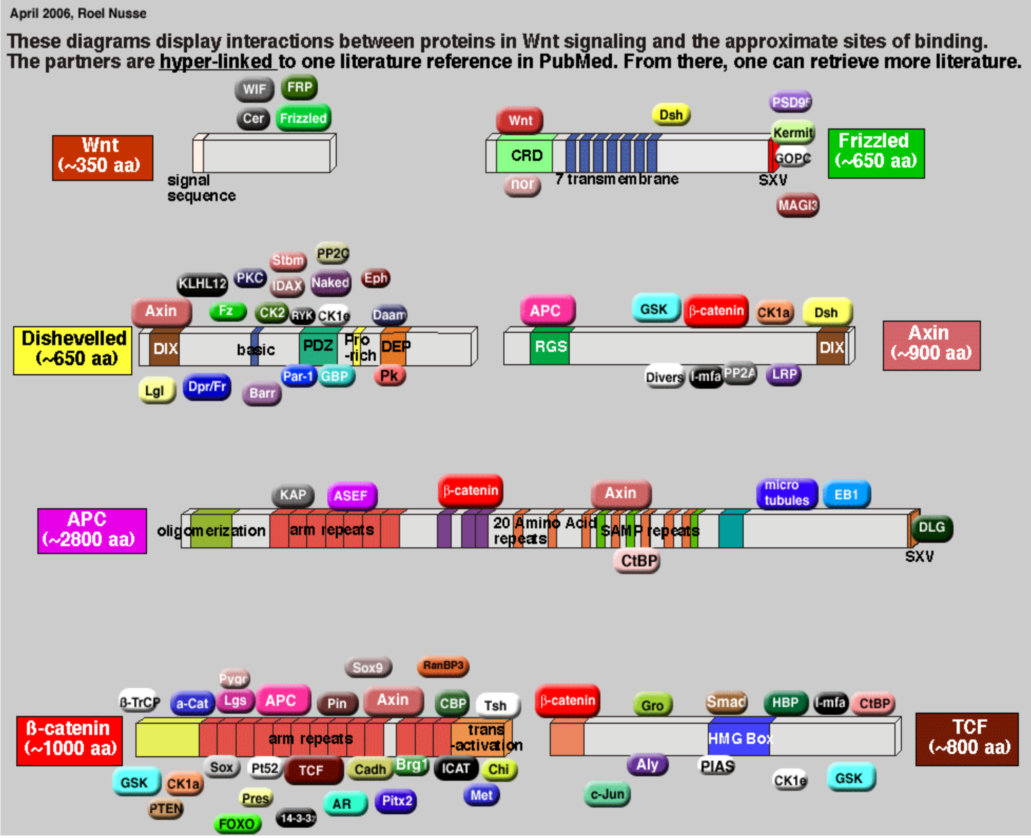
Actual Headcount - % Variance from Budget

			Measures			
Region	Department	Positions	Actual	Budget	Variance	Variance Percent
-All Regions	-All Departments	+All Positions	143,639,982.00	143,199,389.00	-440,593.00	-.31%
	Executive Management	+All Positions	6,299,022.00	6,494,166.00	195,144.00	3.00%
	Finance	+All Positions	12,224,220.00	12,087,406.00	-136,814.00	-1.13%
	Human Resource	+All Positions	13,075,463.00	12,989,341.00	-86,122.00	-.66%
	Marketing & Communication	+All Positions	13,910,753.00	13,770,267.00	-140,486.00	-1.02%
	Product Development	+All Positions	10,644,102.00	10,786,611.00	142,509.00	1.32%
	Professional Services	+All Positions	76,317,649.00	76,098,206.00	-219,443.00	-.29%
	Sales	+All Positions	11,168,773.00	10,973,392.00	-195,381.00	-1.78%
Central	+All Departments	+All Positions	37,893,162.00	38,397,600.00	504,438.00	1.31%
Eastern	+All Departments	+All Positions	35,248,940.00	35,487,861.00	238,921.00	.67%
Southern	+All Departments	+All Positions	35,248,940.00	34,803,861.00	-445,079.00	-1.28%
Western	+All Departments	+All Positions	35,248,940.00	34,510,067.00	-738,873.00	-2.14%

http://www.pentaho.com/products/dashboards/

The "Unstructured Data" Challenge

Some information doesn't come from a data file.



April 2006, Roel Nusse

These diagrams display interactions between proteins in Wnt signaling and the approximate sites of binding. The partners are hyper-linked to one literature reference in PubMed. From there, one can retrieve more literature.

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

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. We also found that Dvl, Axin and GSK can form a ternary complex bridged by Axin, and that Frat1 can be recruited into this complex probably by Dvl. The observation that the Dvl-binding domain of either Frat1 or Axin was able to inhibit Wnt-1-induced LEF-1 activation suggests that the interactions between Dvl and Axin and between Dvl and Frat may be important for this signaling pathway. Furthermore, Wnt-1 appeared to promote the disintegration of the Frat1-Dvl-GSK-Axin complex, resulting in the dissociation of GSK from Axin. Thus, formation of the quaternary complex may be an important step in Wnt signaling, by which Dvl recruits Frat1, leading to Frat1-mediated dissociation of GSK from Axin.

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

The "Unstructured Data" Challenge

Consider:

- E-mail, news & blog articles, forum postings, and other social media.
- Contact-center notes and transcripts.
- Surveys, feedback forms, warranty claims.
- And every kind of corporate documents imaginable.

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 the Nasdaq composite gained 6.84, or 0.32 percent, to 2,162.78.

Search

So there’s data and other interesting information in text. How do we get at it?

Search is not the answer. It returns documents.

Analysts want facts, answers to questions. And what if you're unsure what question to ask?

All the same, let's think about searches and answers…

Search

Search involves –

Words & phrases: search terms & natural language.

Qualifiers: include/exclude, and/or, not, etc.

Answers involve –

Entities: names, e-mail addresses, phone numbers

Concepts: abstractions of entities.

Facts and relationships.

Abstract attributes, e.g., “expensive,” “comfortable”

Opinions, sentiments: attitudinal data.

... and sometimes BI objects.

Search

Q&A may involve hidden knowledge:

What was the population of Paris in 1848?

Concepts and complexity:

What's the best price for new laptop that I'll use for business trips and around the office?

Opinion:

What do people think of the *Iron Man* movie?

Calculation and structuring:

Who were the top 4 sales people for each product line, region, and quarter for the last two years?

Search

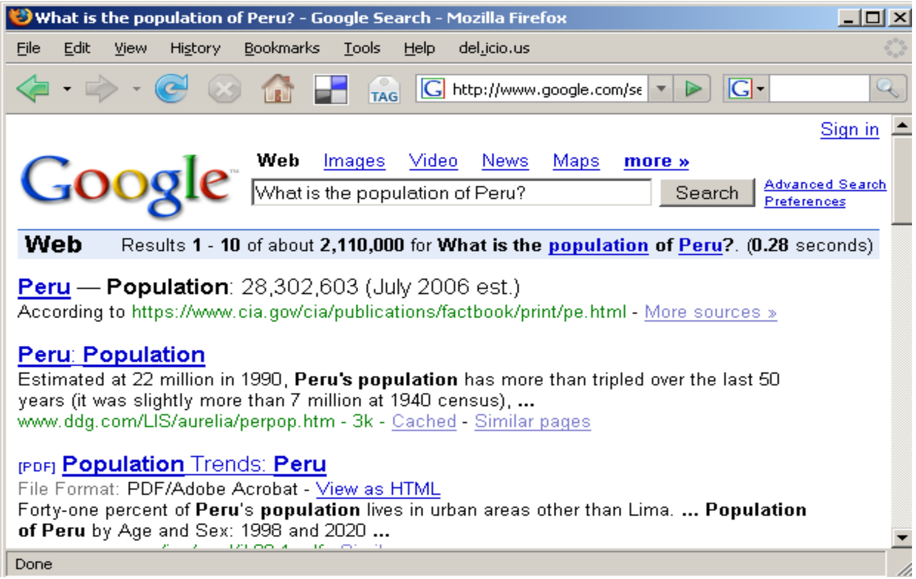
Search is not enough.

- *Search helps you find things you already know about. It doesn't help you* ***discover*** *things you're unaware of.*
- *Search results often lack* ***relevance****.*
- *Search finds documents, not* ***knowledge****.*
- *Search doesn't enable* ***unified analytics*** *that links data from textual and transactional sources.*

Text analytics can make it better...

Beyond Search: Analysis

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


What is the population of Peru? - Google Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
TAG
http://www.google.com/se
Sign in
Google
Web Images Video News Maps more »
Search
Advanced Search
Preferences
Done

What is the population of Peru?

Web Results **1** - **10** of about **2,110,000** for **What is the population of Peru?**. (**0.28** seconds)

Peru — Population: 28,302,603 (July 2006 est.)
According to https://www.cia.gov/cia/publications/factbook/print/pe.html - More sources »

Peru: Population
Estimated at 22 million in 1990, **Peru's population** has more than tripled over the last 50 years (it was slightly more than 7 million at 1940 census), **...**
www.ddg.com/LIS/aurelia/perpop.htm - 3k - Cached - Similar pages

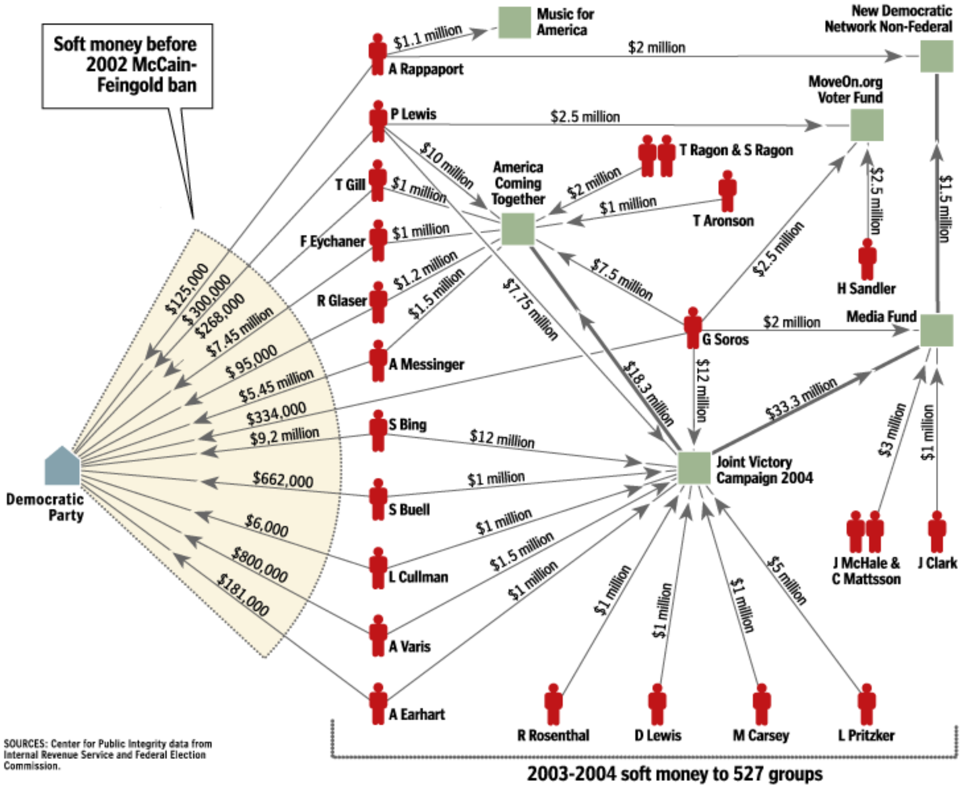
[PDF] **Population** Trends: **Peru**
File Format: PDF/Adobe Acrobat - View as HTML
Forty-one percent of **Peru**'s **population** lives in urban areas other than Lima. **... Population of Peru** by Age and Sex: 1998 and 2020 **...**

Now on to knowledge discovery, to discerning ***interrelationships of presented facts***...

Beyond Search: Analysis

Soft Money Game

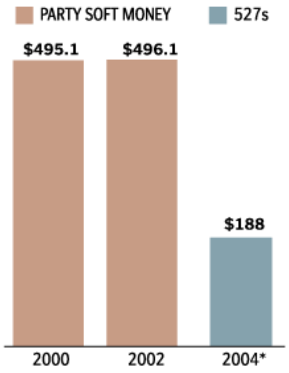
Democrats initially ran into difficulty getting corporate chieftains and their companies to donate soft money to their upstart 527 groups, America Coming Together, The Media Fund and their fundraising arm, the Joint Victory Campaign 2004. Fundraisers turned to maverick donors, many of whom had given soft money to the Democratic Party in the past. This chart shows most donations and transfers of more than $1 million to Democratic 527s through Sept. 30.



SOURCES: Center for Public Integrity data from Internal Revenue Service and Federal Election Commission.

Contributions to 527s active in federal elections have not kept pace with soft money donations to national party committees in previous election cycles. From January of last year through June of this year, 527 groups active in federal elections raised $188 million. In the same 18 months ending in 2002, $308 million in soft money was raised by political parties.

Total receipts, party soft money vs. 527s
(in millions)



*Through June

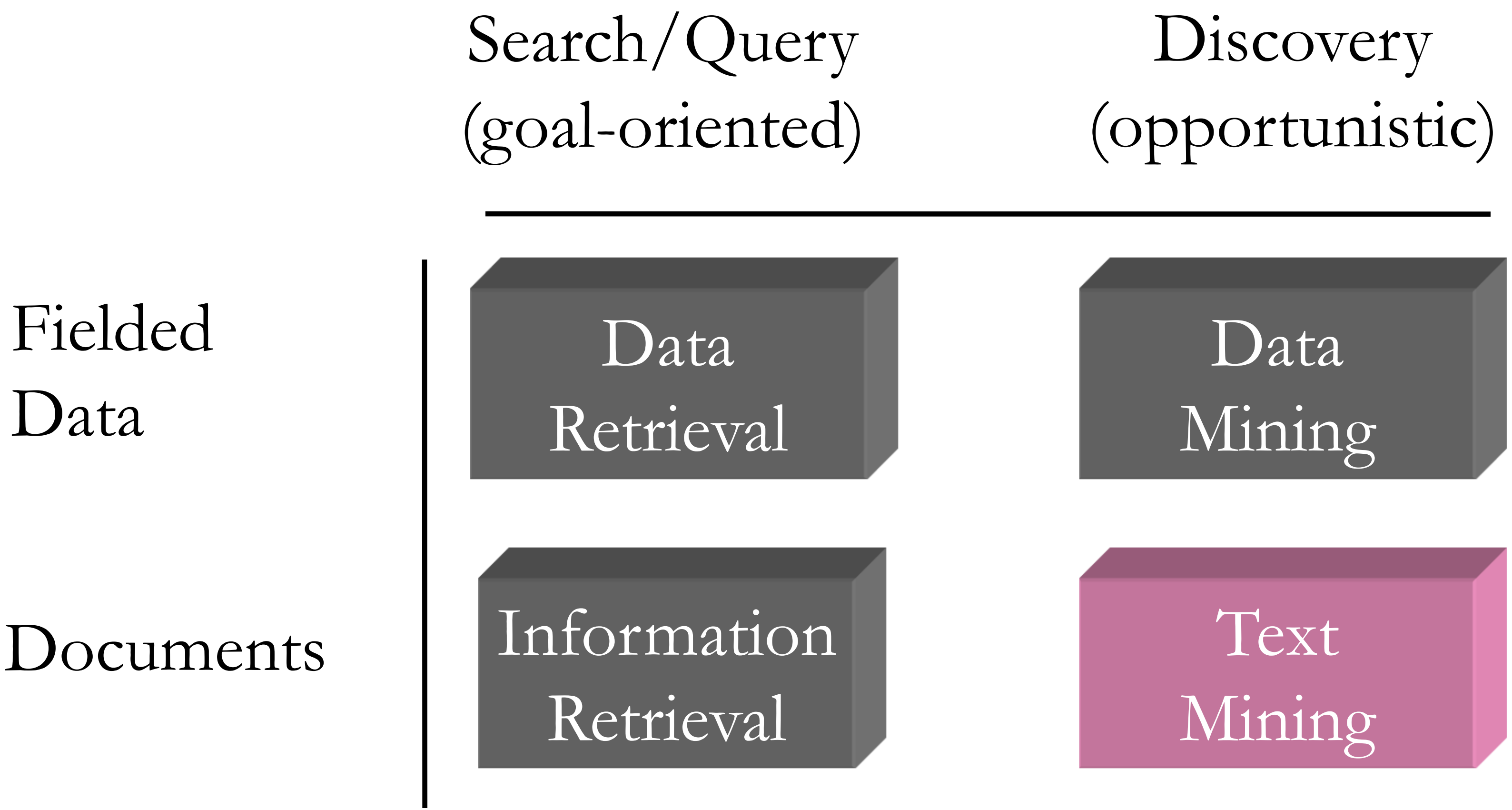
NOTE: Data for 527 activity in the 2004 cycle based on reporting so far. Reporting of 527 activity was required as of mid-2000.

SOURCES: Center for Responsive Politics, Federal Election Commission, Center for Public Integrity

GRAPHICS REPORTING BY SARAH COHEN, JAMES V. GRIMALDI OF THE WASHINGTON POST, AND THE CENTER FOR PUBLIC INTEGRITY. GRAPHIC BY LOUIS SPIRITO—THE WASHINGTON POST

www.washingtonpost.com/wp-srv/politics/daily/graphics/527Diagram_101704.html

Text Mining



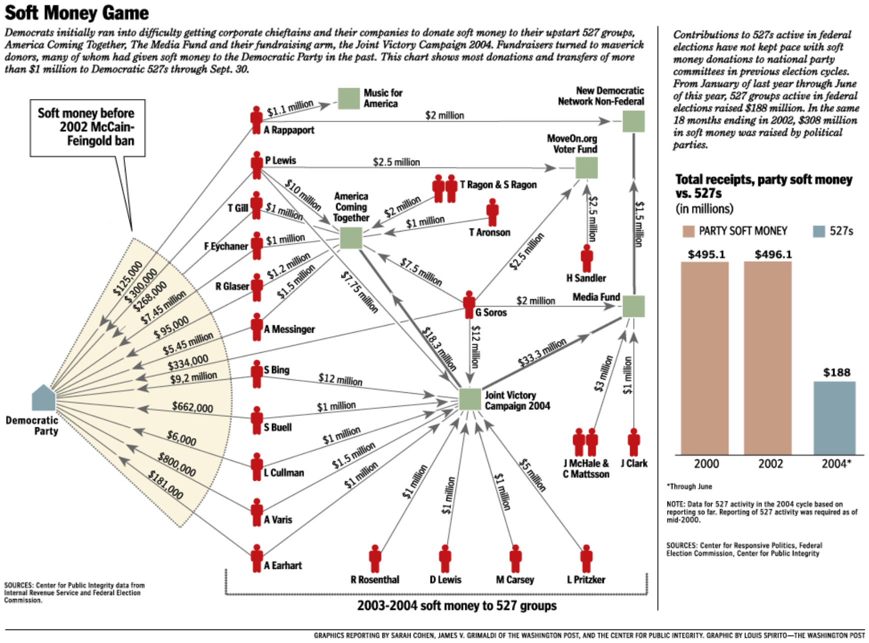
	Search/Query (goal-oriented)	Discovery (opportunistic)
Fielded Data	Data Retrieval	Data Mining
Documents	Information Retrieval	Text Mining

Based on Je Wei Liang, *www.database.cis.nctu.edu.tw/seminars/2003F/TWM/slides/p.ppt*

Text Mining

Text Mining = Data Mining of textual sources.

- Clustering and classification.
- Link Analysis.
- Prediction.
- Association rules.
- ~~Regression.~~
- ~~Forecasting.~~



Soft Money Game

Democrats initially ran into difficulty getting corporate chieftains and their companies to donate soft money to their upstart 527 groups, America Coming Together, The Media Fund and their fundraising arm, the Joint Victory Campaign 2004. Fundraisers turned to maverick donors, many of whom had given soft money to the Democratic Party in the past. This chart shows most donations and transfers of more than $1 million to Democratic 527s through Sept. 30.

SOURCES: Center for Public Integrity data from Internal Revenue Service and Federal Election Commission.

Contributions to 527s active in federal elections have not kept pace with soft money donations to national party committees in previous election cycles. From January of last year through June of this year, 527 groups active in federal elections raised $188 million. In the same 18 months ending in 2002, $308 million in soft money was raised by political parties.

Total receipts, party soft money vs. 527s
(in millions)

*Through June

NOTE: Data for 527 activity in the 2004 cycle based on reporting so far. Reporting of 527 activity was required as of mid-2000.

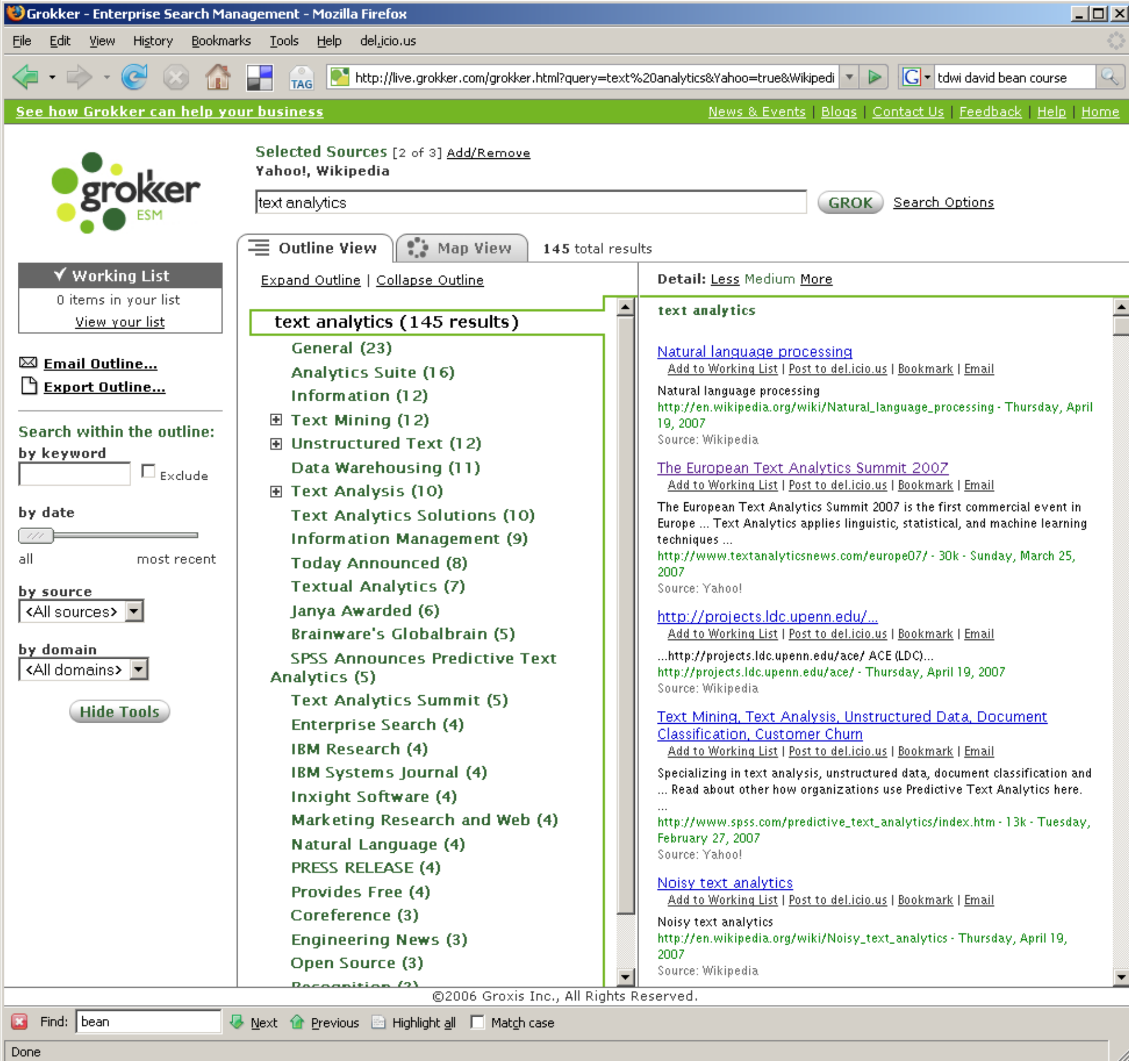
SOURCES: Center for Responsive Politics, Federal Election Commission, Center for Public Integrity

GRAPHICS REPORTING BY SARAH COHEN, JAMES V. GRIMALDI OF THE WASHINGTON POST, AND THE CENTER FOR PUBLIC INTEGRITY. GRAPHIC BY LOUIS SPIRITO—THE WASHINGTON POST

Text Mining = Knowledge Discovery in Text.

Search can be pretty smart.

This slide and the next show dynamic, clustered search results from Grokker…


Grokker - Enterprise Search Management - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
TAG
grokker ESM

http://live.grokker.com/grokker.html?query=text%20analytics&Yahoo=true&Wikipedi

tdwi david bean course

See how Grokker can help your business | News & Events | Blogs | Contact Us | Feedback | Help | Home

Selected Sources [2 of 3] Add/Remove
Yahoo!, Wikipedia

text analytics — GROK — Search Options

Outline View | Map View — 145 total results

✓ Working List
0 items in your list
View your list

Email Outline...
Export Outline...

Search within the outline:
by keyword □ Exclude
by date: all — most recent
by source <All sources>
by domain <All domains>
Hide Tools

Expand Outline | Collapse Outline

- text analytics (145 results)
  - General (23)
  - Analytics Suite (16)
  - Information (12)
  - ⊞ Text Mining (12)
  - ⊞ Unstructured Text (12)
  - Data Warehousing (11)
  - ⊞ Text Analysis (10)
  - Text Analytics Solutions (10)
  - Information Management (9)
  - Today Announced (8)
  - Textual Analytics (7)
  - Janya Awarded (6)
  - Brainware's Globalbrain (5)
  - SPSS Announces Predictive Text Analytics (5)
  - Text Analytics Summit (5)
  - Enterprise Search (4)
  - IBM Research (4)
  - IBM Systems Journal (4)
  - Inxight Software (4)
  - Marketing Research and Web (4)
  - Natural Language (4)
  - PRESS RELEASE (4)
  - Provides Free (4)
  - Coreference (3)
  - Engineering News (3)
  - Open Source (3)
  - Recognition (3)

Detail: Less Medium More

text analytics

Natural language processing
Add to Working List | Post to del.icio.us | Bookmark | Email
Natural language processing
http://en.wikipedia.org/wiki/Natural_language_processing - Thursday, April 19, 2007
Source: Wikipedia

The European Text Analytics Summit 2007
Add to Working List | Post to del.icio.us | Bookmark | Email
The European Text Analytics Summit 2007 is the first commercial event in Europe ... Text Analytics applies linguistic, statistical, and machine learning techniques ...
http://www.textanalyticsnews.com/europe07/ - 30k - Sunday, March 25, 2007
Source: Yahoo!

http://projects.ldc.upenn.edu/...
Add to Working List | Post to del.icio.us | Bookmark | Email
...http://projects.ldc.upenn.edu/ace/ ACE (LDC)...
http://projects.ldc.upenn.edu/ace/ - Thursday, April 19, 2007
Source: Wikipedia

Text Mining, Text Analysis, Unstructured Data, Document Classification, Customer Churn
Add to Working List | Post to del.icio.us | Bookmark | Email
Specializing in text analysis, unstructured data, document classification and ... Read about other how organizations use Predictive Text Analytics here. ...
http://www.spss.com/predictive_text_analytics/index.htm - 13k - Tuesday, February 27, 2007
Source: Yahoo!

Noisy text analytics
Add to Working List | Post to del.icio.us | Bookmark | Email
Noisy text analytics
http://en.wikipedia.org/wiki/Noisy_text_analytics - Thursday, April 19, 2007
Source: Wikipedia

©2006 Groxis Inc., All Rights Reserved.

Find: bean — Next — Previous — Highlight all — □ Match case

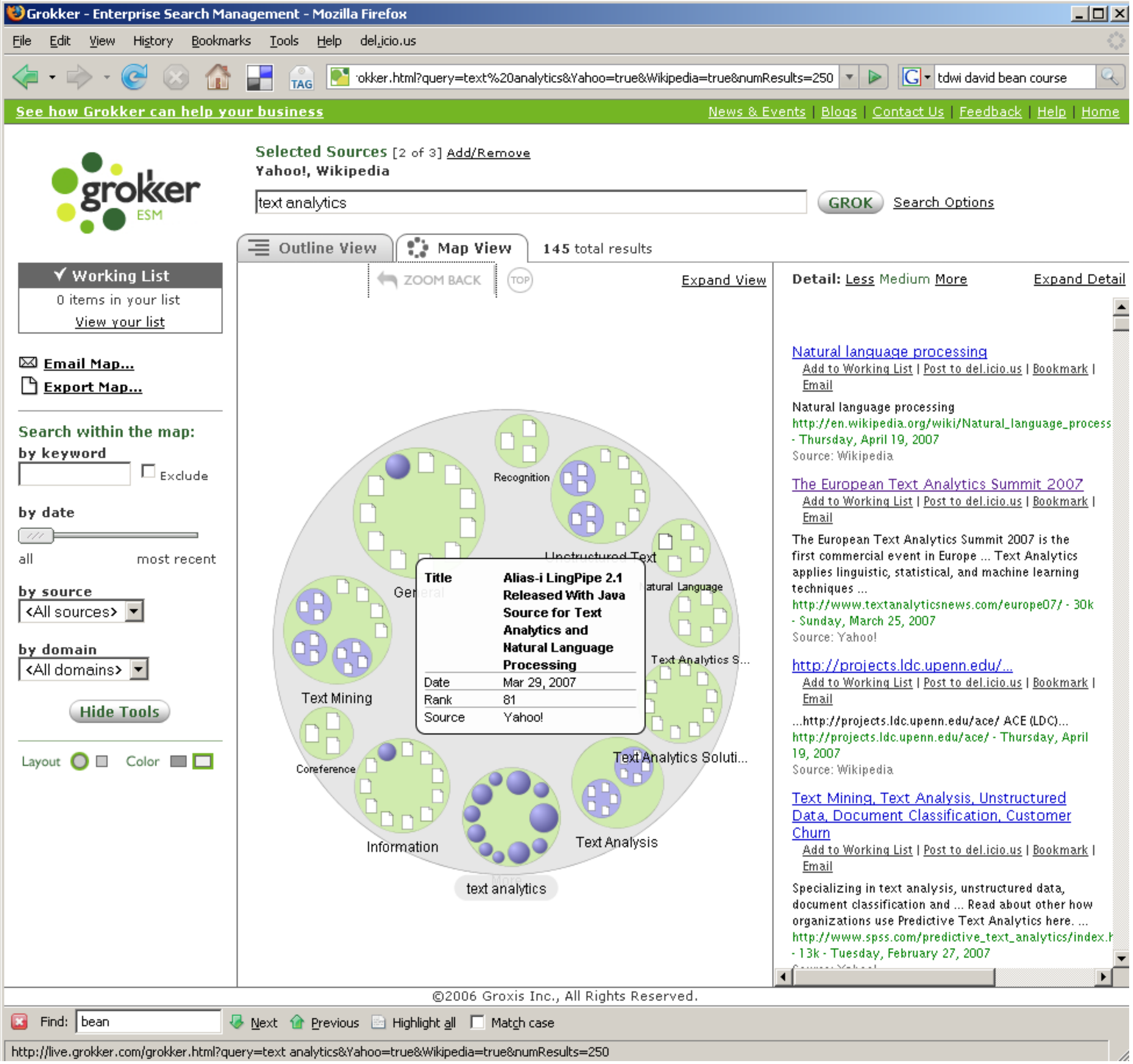
Done

live.grokker.com/grokker.html?query=text%20analytics&Yahoo=true&Wikipedia=true&numResults=250

Text Analytics for Dummies

…with a zoomable display.

Clustering here utilizes statistical (text) data mining techniques to identifying cohesive groupings of retrieved documents.



okker.html?query=text%20analytics&Yahoo=true&Wikipedia=true&numResults=250

tdwi david bean course

Selected Sources [2 of 3] Add/Remove
Yahoo!, Wikipedia

text analytics

145 total results

by source <All sources>

by domain <All domains>

Title	Alias-i LingPipe 2.1 Released With Java Source for Text Analytics and Natural Language Processing
Date	Mar 29, 2007
Rank	81
Source	Yahoo!

Natural language processing
Add to Working List | Post to del.icio.us | Bookmark | Email

Natural language processing
http://en.wikipedia.org/wiki/Natural_language_process
- Thursday, April 19, 2007
Source: Wikipedia

The European Text Analytics Summit 2007
Add to Working List | Post to del.icio.us | Bookmark | Email

The European Text Analytics Summit 2007 is the first commercial event in Europe ... Text Analytics applies linguistic, statistical, and machine learning techniques ...
http://www.textanalyticsnews.com/europe07/ - 30k
- Sunday, March 25, 2007
Source: Yahoo!

http://projects.ldc.upenn.edu/...
Add to Working List | Post to del.icio.us | Bookmark | Email

...http://projects.ldc.upenn.edu/ace/ ACE (LDC)...
http://projects.ldc.upenn.edu/ace/ - Thursday, April 19, 2007
Source: Wikipedia

Text Mining, Text Analysis, Unstructured Data, Document Classification, Customer Churn
Add to Working List | Post to del.icio.us | Bookmark | Email

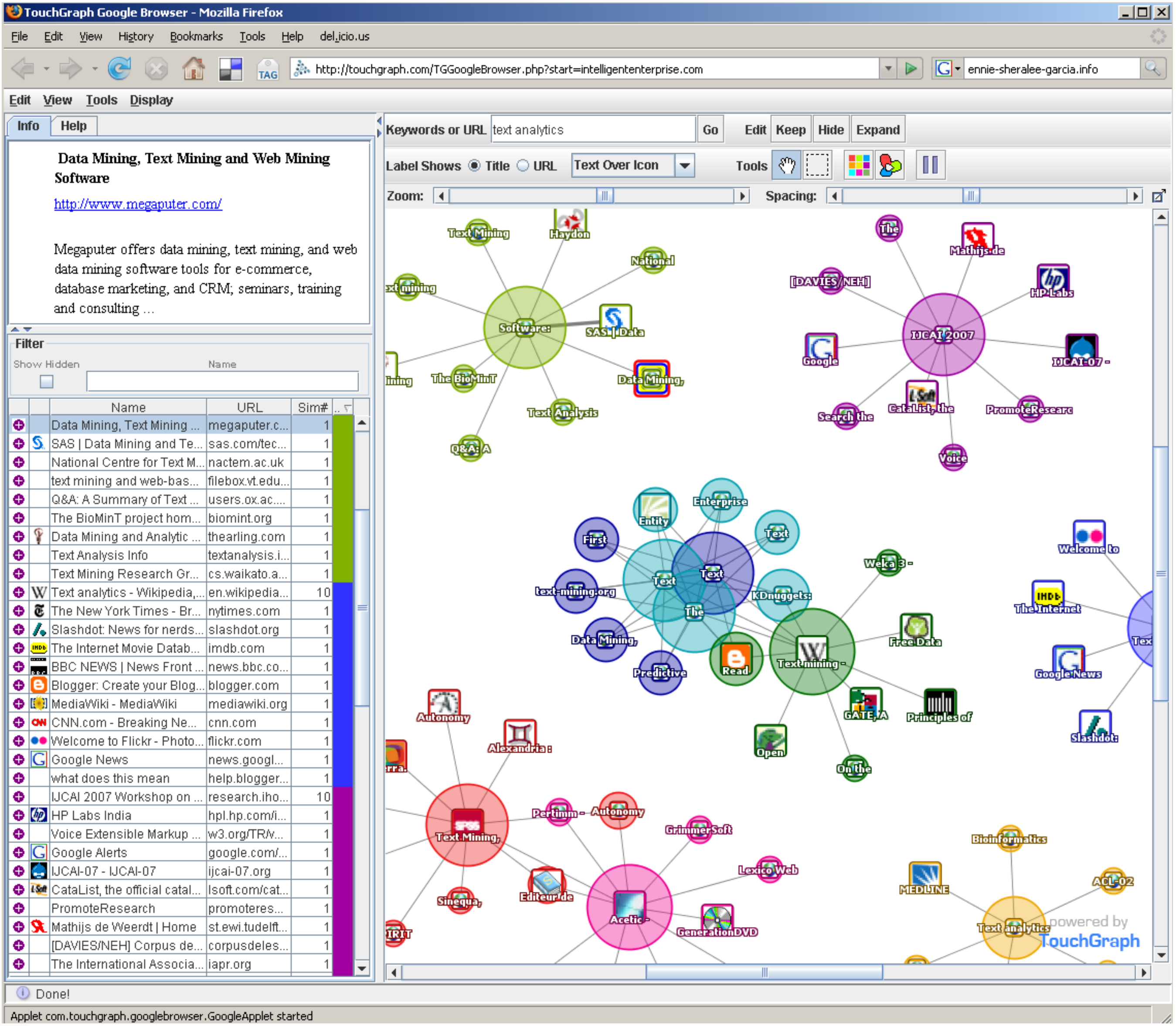
Specializing in text analysis, unstructured data, document classification and ... Read about other how organizations use Predictive Text Analytics here. ...
http://www.spss.com/predictive_text_analytics/index.h
- 13k - Tuesday, February 27, 2007

©2006 Groxis Inc., All Rights Reserved.

Find: bean

http://live.grokker.com/grokker.html?query=text analytics&Yahoo=true&Wikipedia=true&numResults=250

A dynamic network viz.: the TouchGraph GoogleBrowser applet



http://touchgraph.com/TGGoogleBrowser.php?start=intelligententerprise.com

Keywords or URL: text analytics

Label Shows ● Title ○ URL   Text Over Icon

Data Mining, Text Mining and Web Mining Software

http://www.megaputer.com/

Megaputer offers data mining, text mining, and web data mining software tools for e-commerce, database marketing, and CRM; seminars, training and consulting ...

Filter

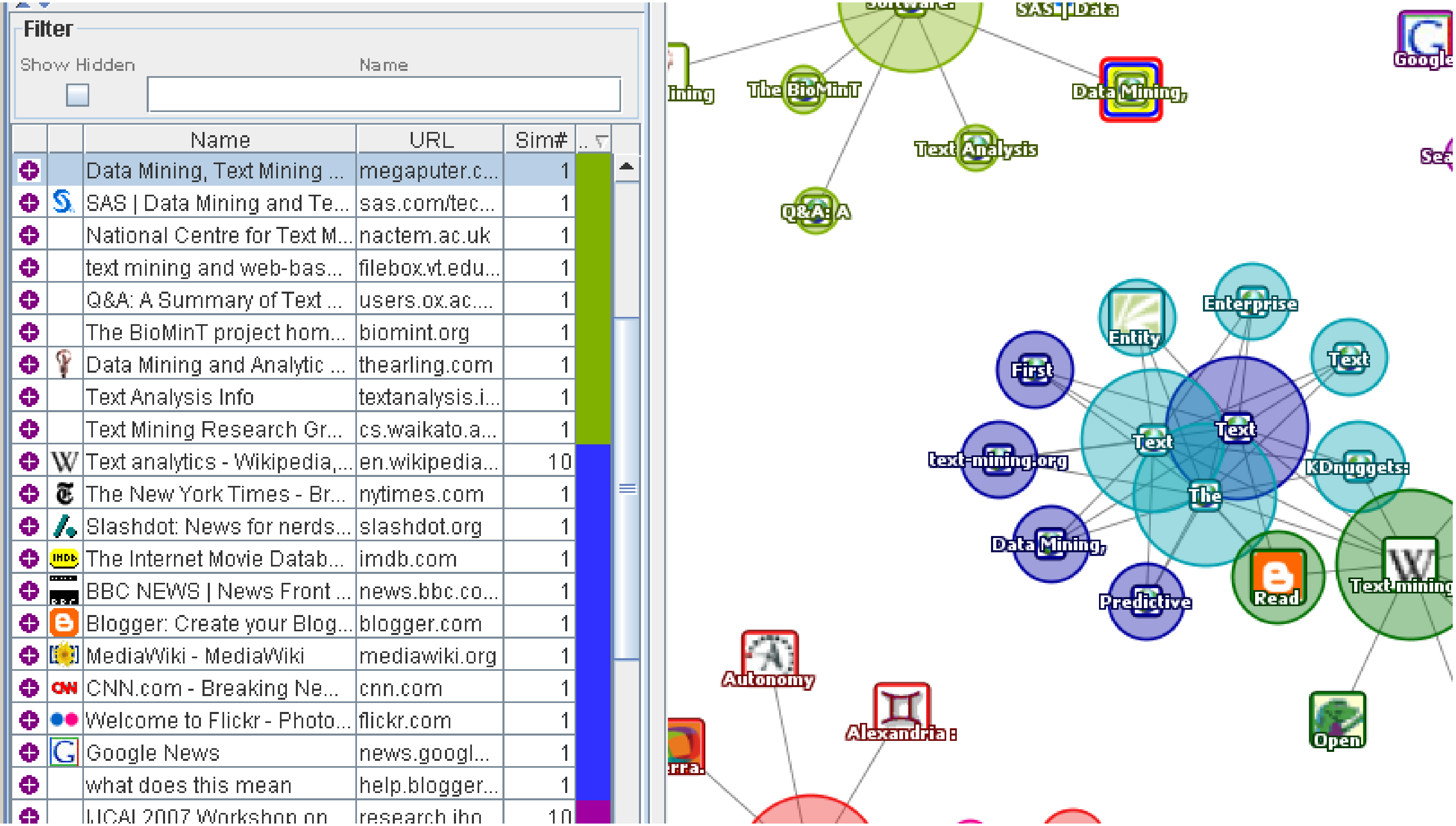
Show Hidden □   Name

Name	URL	Sim#
Data Mining, Text Mining ...	megaputer.c...	1
SAS | Data Mining and Te...	sas.com/tec...	1
National Centre for Text M...	nactem.ac.uk	1
text mining and web-bas...	filebox.vt.edu...	1
Q&A: A Summary of Text ...	users.ox.ac....	1
The BioMinT project hom...	biomint.org	1
Data Mining and Analytic ...	thearling.com	1
Text Analysis Info	textanalysis.i...	1
Text Mining Research Gr...	cs.waikato.a...	1
Text analytics - Wikipedia,...	en.wikipedia...	10
The New York Times - Br...	nytimes.com	1
Slashdot: News for nerds...	slashdot.org	1
The Internet Movie Datab...	imdb.com	1
BBC NEWS | News Front ...	news.bbc.co...	1
Blogger: Create your Blog...	blogger.com	1
MediaWiki - MediaWiki	mediawiki.org	1
CNN.com - Breaking Ne...	cnn.com	1
Welcome to Flickr - Photo...	flickr.com	1
Google News	news.googl...	1
what does this mean	help.blogger...	1
IJCAI 2007 Workshop on ...	research.iho...	10
HP Labs India	hpl.hp.com/i...	1
Voice Extensible Markup ...	w3.org/TR/v...	1
Google Alerts	google.com/...	1
IJCAI-07 - IJCAI-07	ijcai-07.org	1
CataList, the official catal...	lsoft.com/cat...	1
PromoteResearch	promoteres...	1
Mathijs de Weerdt | Home	st.ewi.tudelft...	1
[DAVIES/NEH] Corpus de...	corpusdeles...	1
The International Associa...	iapr.org	1

Done!

Applet com.touchgraph.googlebrowser.GoogleApplet started

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



Filter

Show Hidden □

Name

		Name	URL	Sim#	..
⊕		Data Mining, Text Mining ...	megaputer.c...	1	🟩
⊕	S	SAS | Data Mining and Te...	sas.com/tec...	1	🟩
⊕		National Centre for Text M...	nactem.ac.uk	1	🟩
⊕		text mining and web-bas...	filebox.vt.edu...	1	🟩
⊕		Q&A: A Summary of Text ...	users.ox.ac....	1	🟩
⊕		The BioMinT project hom...	biomint.org	1	🟩
⊕		Data Mining and Analytic ...	thearling.com	1	🟩
⊕		Text Analysis Info	textanalysis.i...	1	🟩
⊕		Text Mining Research Gr...	cs.waikato.a...	1	🟩
⊕	W	Text analytics - Wikipedia,...	en.wikipedia...	10	🟦
⊕		The New York Times - Br...	nytimes.com	1	🟦
⊕		Slashdot: News for nerds...	slashdot.org	1	🟦
⊕	IMDb	The Internet Movie Datab...	imdb.com	1	🟦
⊕	BBC	BBC NEWS | News Front ...	news.bbc.co...	1	🟦
⊕		Blogger: Create your Blog...	blogger.com	1	🟦
⊕		MediaWiki - MediaWiki	mediawiki.org	1	🟦
⊕	CNN	CNN.com - Breaking Ne...	cnn.com	1	🟦
⊕		Welcome to Flickr - Photo...	flickr.com	1	🟦
⊕	G	Google News	news.googl...	1	🟦
⊕		what does this mean	help.blogger...	1	🟦
⊕		IJCAI 2007 Workshop on	research.ibn	10	🟪

Text Analytics

So text analytics enhances search, a.k.a. Information Retrieval.

- It recognizes patterns in search queries to enable basic question answering.
- It recognizes patterns in search results to enable clustering of results.

We want to get beyond IR to Information Extraction (IE).

First, *time out* to summarize and provide some definitions...

Glossary

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.*
- ***Unlocks meaning and relationships*** *in large volumes of information that were previously unprocessable by computer.*

Glossary

Text Analytics is perhaps a superset of ***Text Mining.***

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.

Fact: A relationship between two entities.

Sentiment: A valuation at the entity or higher level.

Opinion: A fact that involves a sentiment.

Glossary

Semantics: A fancy word for meaning, as distinct from ***Syntax,*** which is structuring.

Natural Language Processing (NLP): Computers hear humans.

Parsing: Evaluating the contents of a document.

Tokenization: Identification of distinct elements within a text.

Stemming/ Lemmatization : Reducing variants of word bases created by conjugation, declension, case, pluralization, etc.

Tagging: Wrapping XML tags around distinct text elements, a.k.a. ***text augmentation***.

POS Tagging: Specifically identifying parts of speech.

Glossary

Categorization: Specification of ways like items can be grouped.

Clustering: Creating categories according to statistical 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.

Taxonomy: A hierarchical categorization of entities and concepts.

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

Text Analytics

Typical steps in text analytics include –

Retrieve documents for analysis.

Apply statistical &/ linguistic &/ structural techniques to **identify, tag, and extract** entities, concepts, relationships, and events (features) within document sets.

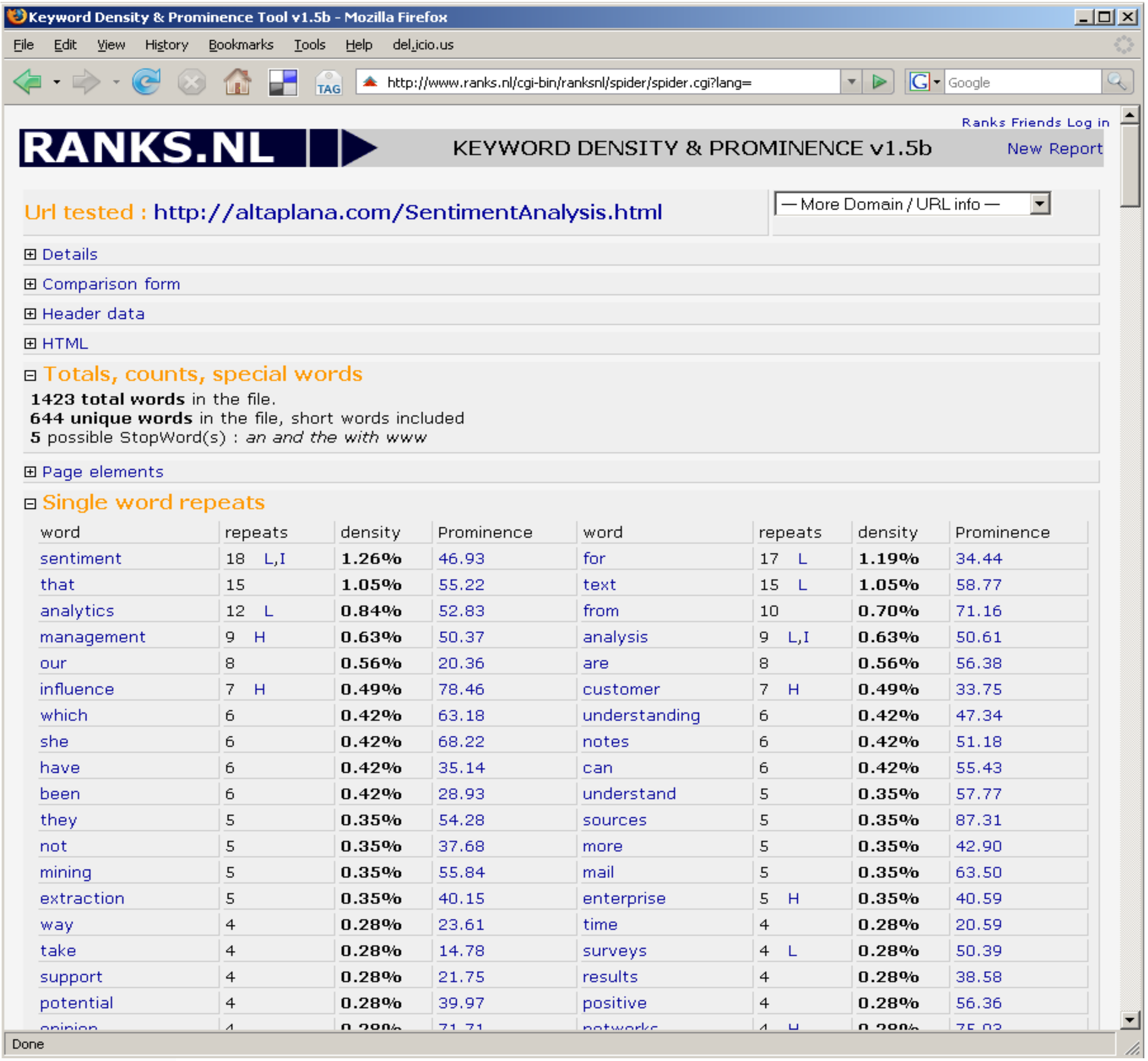
Apply statistical pattern-matching & similarity techniques to **classify** documents and organize extracted features according to a specified or generated categorization / taxonomy.

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

Text Analytics

So text analytics looks for structure that is inherent in the textual source materials. Let's look at some of the steps.

First, we'll do a lexical analysis of a text file, essentially a basic statistical analysis of the words and multi-word terms...


Keyword Density & Prominence Tool v1.5b - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
del.icio.us
TAG
http://www.ranks.nl/cgi-bin/ranksnl/spider/spider.cgi?lang=
Google
Ranks Friends Log in
RANKS.NL
New Report
— More Domain / URL info —
Done

KEYWORD DENSITY & PROMINENCE v1.5b

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

⊞ Details

⊞ Comparison form

⊞ Header data

⊞ HTML

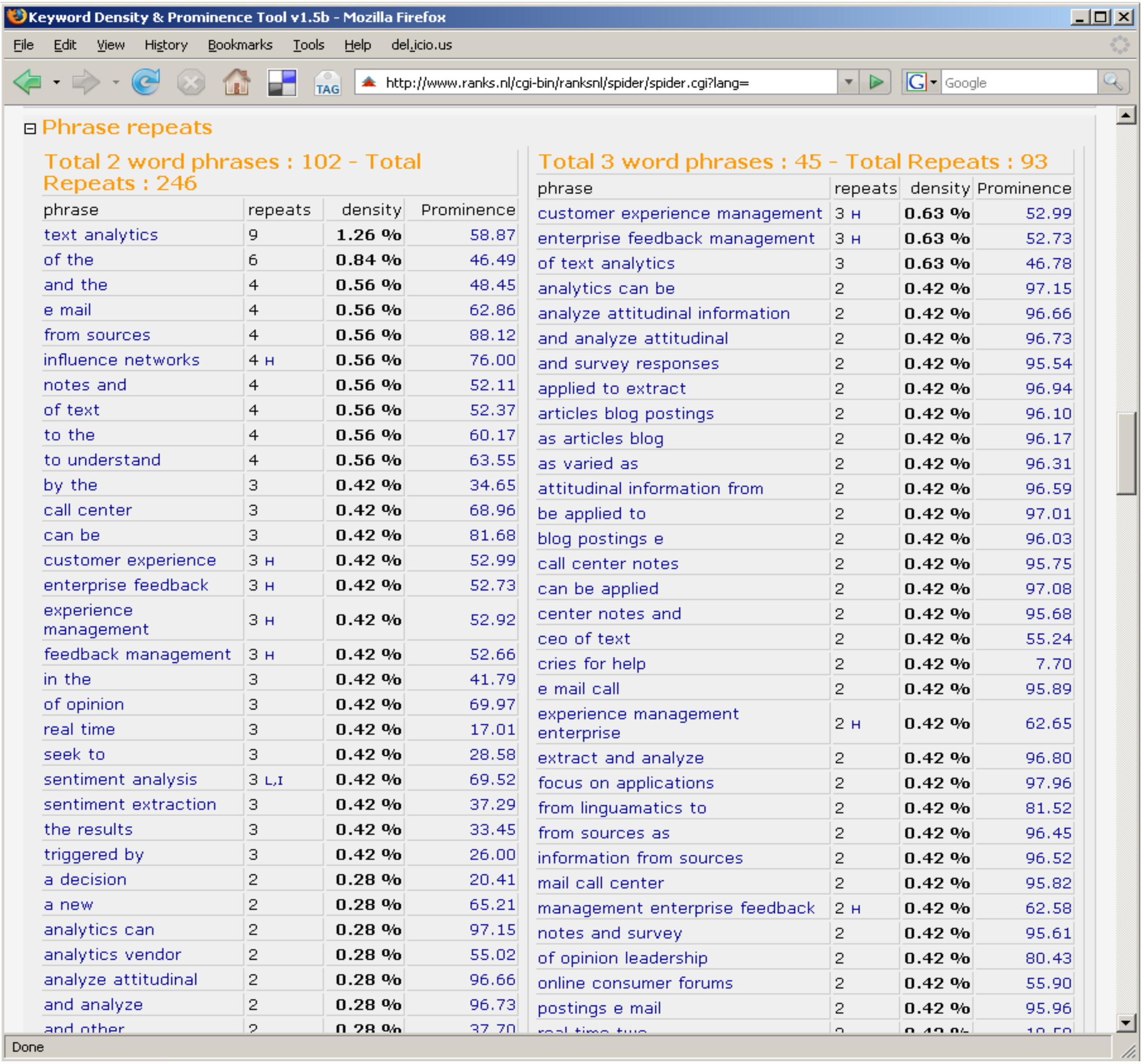
⊟ Totals, counts, special words

1423 total words in the file.
644 unique words in the file, short words included
5 possible StopWord(s) : *an and the with www*

⊞ Page elements

⊟ Single word repeats

word	repeats	density	Prominence	word	repeats	density	Prominence
sentiment	18 L,I	**1.26%**	46.93	for	17 L	**1.19%**	34.44
that	15	**1.05%**	55.22	text	15 L	**1.05%**	58.77
analytics	12 L	**0.84%**	52.83	from	10	**0.70%**	71.16
management	9 H	**0.63%**	50.37	analysis	9 L,I	**0.63%**	50.61
our	8	**0.56%**	20.36	are	8	**0.56%**	56.38
influence	7 H	**0.49%**	78.46	customer	7 H	**0.49%**	33.75
which	6	**0.42%**	63.18	understanding	6	**0.42%**	47.34
she	6	**0.42%**	68.22	notes	6	**0.42%**	51.18
have	6	**0.42%**	35.14	can	6	**0.42%**	55.43
been	6	**0.42%**	28.93	understand	5	**0.35%**	57.77
they	5	**0.35%**	54.28	sources	5	**0.35%**	87.31
not	5	**0.35%**	37.68	more	5	**0.35%**	42.90
mining	5	**0.35%**	55.84	mail	5	**0.35%**	63.50
extraction	5	**0.35%**	40.15	enterprise	5 H	**0.35%**	40.59
way	4	**0.28%**	23.61	time	4	**0.28%**	20.59
take	4	**0.28%**	14.78	surveys	4 L	**0.28%**	50.39
support	4	**0.28%**	21.75	results	4	**0.28%**	38.58
potential	4	**0.28%**	39.97	positive	4	**0.28%**	56.36
opinion	4	**0.28%**	71.71	networks	4 H	**0.28%**	75.03


Keyword Density & Prominence Tool v1.5b - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
http://www.ranks.nl/cgi-bin/ranksnl/spider/spider.cgi?lang=
Google
Done

⊟ Phrase repeats

Total 2 word phrases : 102 - Total Repeats : 246

phrase	repeats	density	Prominence
text analytics	9	1.26 %	58.87
of the	6	0.84 %	46.49
and the	4	0.56 %	48.45
e mail	4	0.56 %	62.86
from sources	4	0.56 %	88.12
influence networks	4 H	0.56 %	76.00
notes and	4	0.56 %	52.11
of text	4	0.56 %	52.37
to the	4	0.56 %	60.17
to understand	4	0.56 %	63.55
by the	3	0.42 %	34.65
call center	3	0.42 %	68.96
can be	3	0.42 %	81.68
customer experience	3 H	0.42 %	52.99
enterprise feedback	3 H	0.42 %	52.73
experience management	3 H	0.42 %	52.92
feedback management	3 H	0.42 %	52.66
in the	3	0.42 %	41.79
of opinion	3	0.42 %	69.97
real time	3	0.42 %	17.01
seek to	3	0.42 %	28.58
sentiment analysis	3 L,I	0.42 %	69.52
sentiment extraction	3	0.42 %	37.29
the results	3	0.42 %	33.45
triggered by	3	0.42 %	26.00
a decision	2	0.28 %	20.41
a new	2	0.28 %	65.21
analytics can	2	0.28 %	97.15
analytics vendor	2	0.28 %	55.02
analyze attitudinal	2	0.28 %	96.66
and analyze	2	0.28 %	96.73
and other	2	0.28 %	37.70

Total 3 word phrases : 45 - Total Repeats : 93

phrase	repeats	density	Prominence
customer experience management	3 H	0.63 %	52.99
enterprise feedback management	3 H	0.63 %	52.73
of text analytics	3	0.63 %	46.78
analytics can be	2	0.42 %	97.15
analyze attitudinal information	2	0.42 %	96.66
and analyze attitudinal	2	0.42 %	96.73
and survey responses	2	0.42 %	95.54
applied to extract	2	0.42 %	96.94
articles blog postings	2	0.42 %	96.10
as articles blog	2	0.42 %	96.17
as varied as	2	0.42 %	96.31
attitudinal information from	2	0.42 %	96.59
be applied to	2	0.42 %	97.01
blog postings e	2	0.42 %	96.03
call center notes	2	0.42 %	95.75
can be applied	2	0.42 %	97.08
center notes and	2	0.42 %	95.68
ceo of text	2	0.42 %	55.24
cries for help	2	0.42 %	7.70
e mail call	2	0.42 %	95.89
experience management enterprise	2 H	0.42 %	62.65
extract and analyze	2	0.42 %	96.80
focus on applications	2	0.42 %	97.96
from linguamatics to	2	0.42 %	81.52
from sources as	2	0.42 %	96.45
information from sources	2	0.42 %	96.52
mail call center	2	0.42 %	95.82
management enterprise feedback	2 H	0.42 %	62.58
notes and survey	2	0.42 %	95.61
of opinion leadership	2	0.42 %	80.43
online consumer forums	2	0.42 %	55.90
postings e mail	2	0.42 %	95.96
real time two	2	0.42 %	19.50

Alta Plana

Text Analytics

Those "tri-grams" are pretty good at describing the *Whatness* of the source text.

Lesson: "Structure" may not matter.

- Shallow parsing and statistical analysis can be enough, for instance, to support classification. (But that's not BI.)
- It can help you get at meaning, for instance, by studying co-occurrence of terms.

But statistical pattern matching – the bag/vector of words approach – may fall short.

The Need for Linguistics

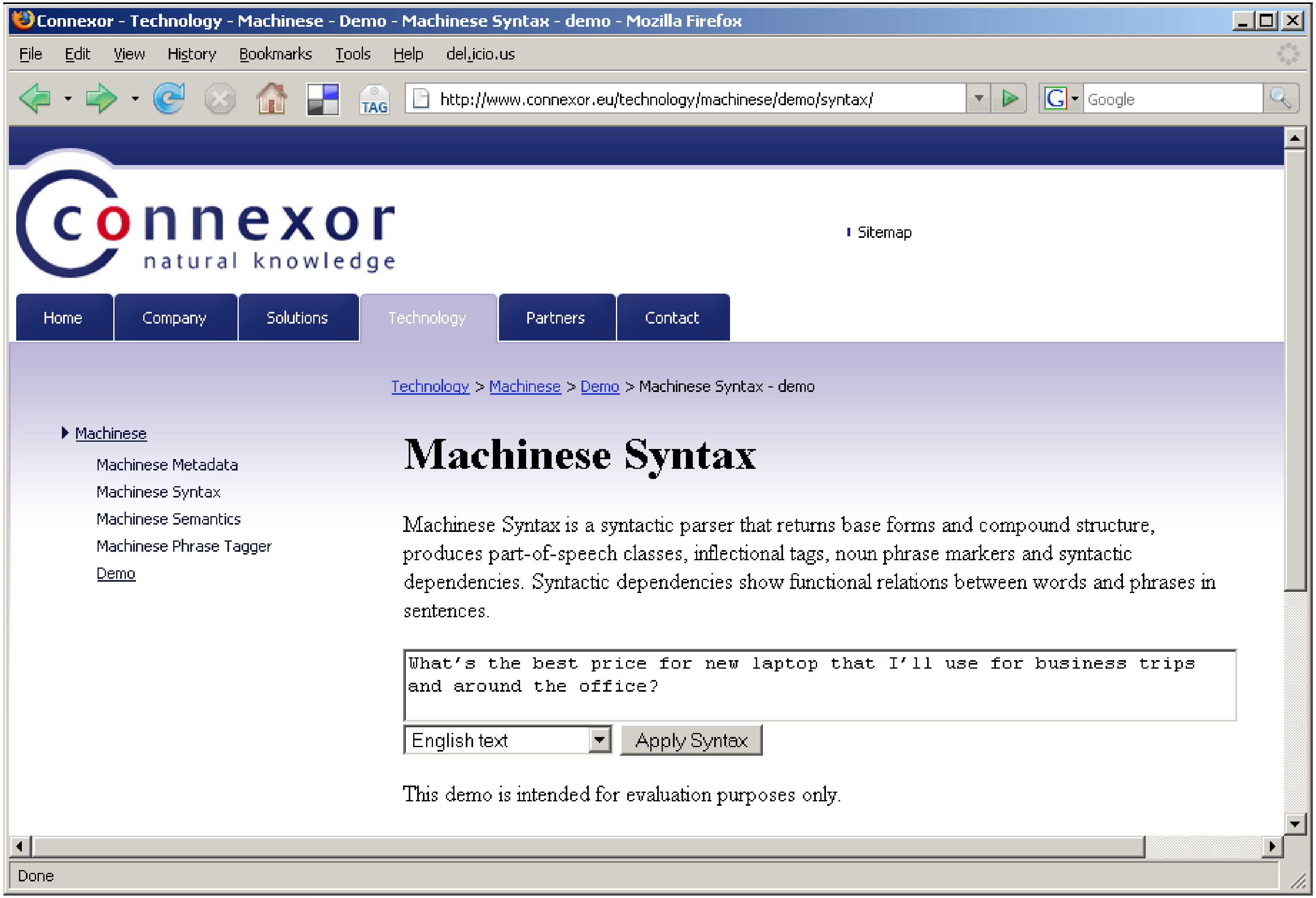
Consider –

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 the Nasdaq composite *gained* 6.84, or 0.32 percent, to 2,162.78.

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, and the Nasdaq composite *fell* 6.84, or 0.32 percent, to 2,162.78.

Example from Luca Scagliarini, Expert System.

Let’s try syntactic analysis of a bit of text...


Connexor - Technology - Machinese - Demo - Machinese Syntax - demo - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
http://www.connexor.eu/technology/machinese/demo/syntax/
Google
connexor natural knowledge
Sitemap
Home Company Solutions Technology Partners Contact
Machinese
Machinese Metadata
Machinese Syntax
Machinese Semantics
Machinese Phrase Tagger
Demo
Done

Technology > Machinese > Demo > Machinese Syntax - demo

Machinese Syntax

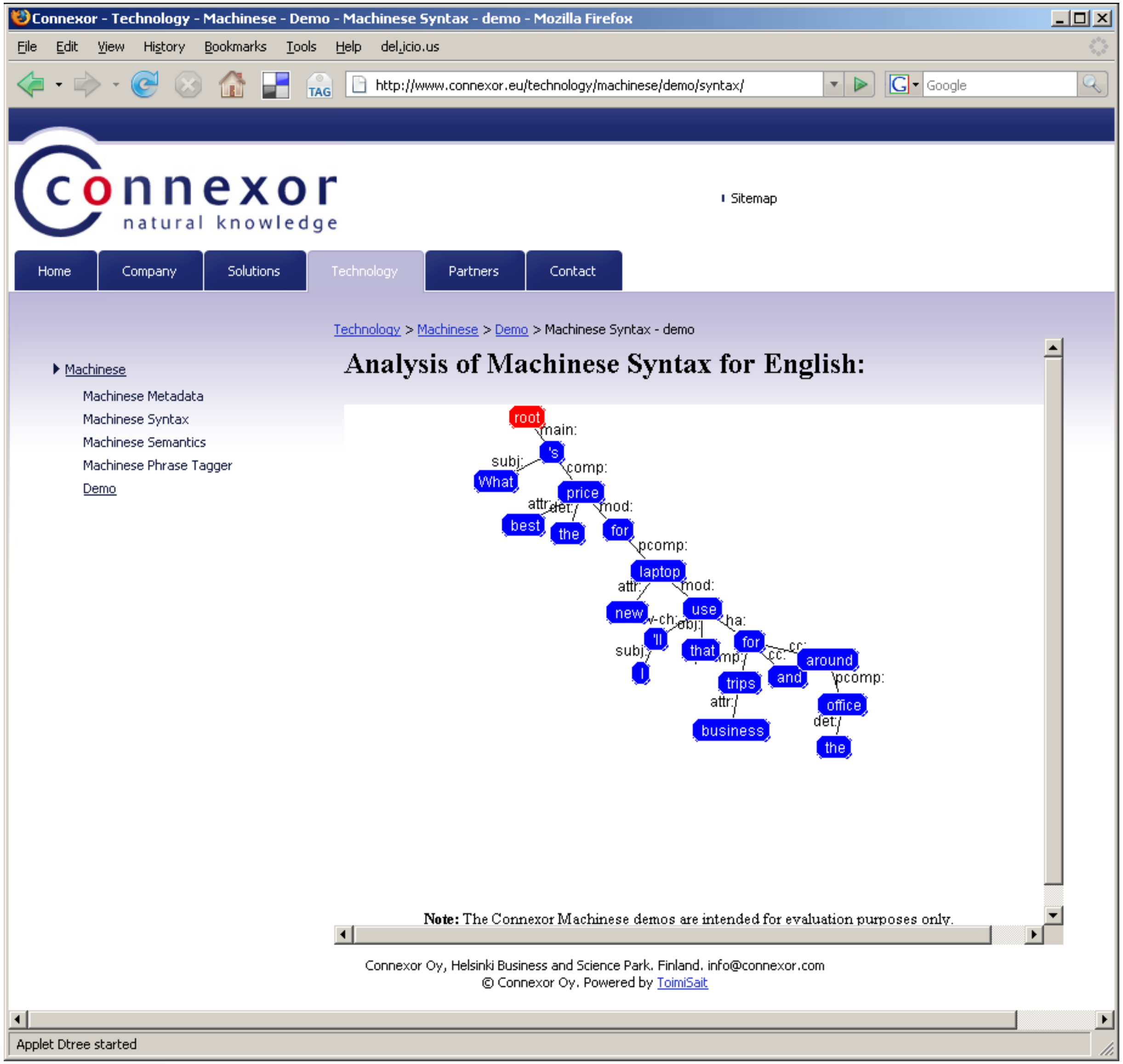
Machinese Syntax is a syntactic parser that returns base forms and compound structure, produces part-of-speech classes, inflectional tags, noun phrase markers and syntactic dependencies. Syntactic dependencies show functional relations between words and phrases in sentences.

```
What's the best price for new laptop that I'll use for business trips
and around the office?
```

English text ▾ | Apply Syntax

This demo is intended for evaluation purposes only.

Alta Plana



Technology > Machinese > Demo > Machinese Syntax - demo

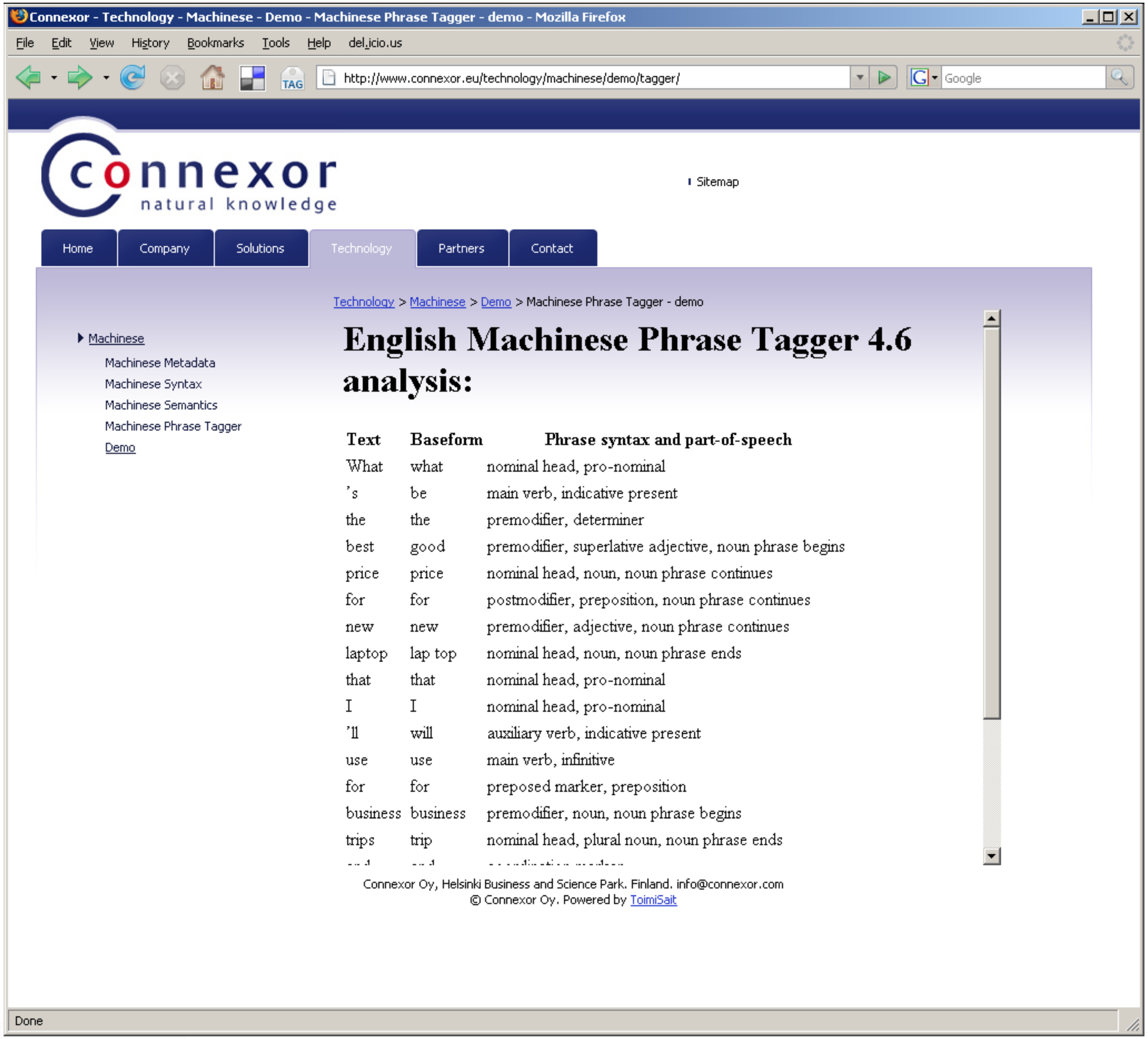
Analysis of Machinese Syntax for English:

- Machinese
  - Machinese Metadata
  - Machinese Syntax
  - Machinese Semantics
  - Machinese Phrase Tagger
  - Demo

Note: The Connexor Machinese demos are intended for evaluation purposes only.

Connexor Oy, Helsinki Business and Science Park, Finland. info@connexor.com
© Connexor Oy. Powered by ToimiSait

Applet Dtree started


Connexor - Technology - Machinese - Demo - Machinese Phrase Tagger - demo - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
TAG
http://www.connexor.eu/technology/machinese/demo/tagger/
Google
connexor
natural knowledge
Sitemap
Home
Company
Solutions
Technology
Partners
Contact
Machinese
Machinese Metadata
Machinese Syntax
Machinese Semantics
Machinese Phrase Tagger
Demo
Done

Technology > Machinese > Demo > Machinese Phrase Tagger - demo

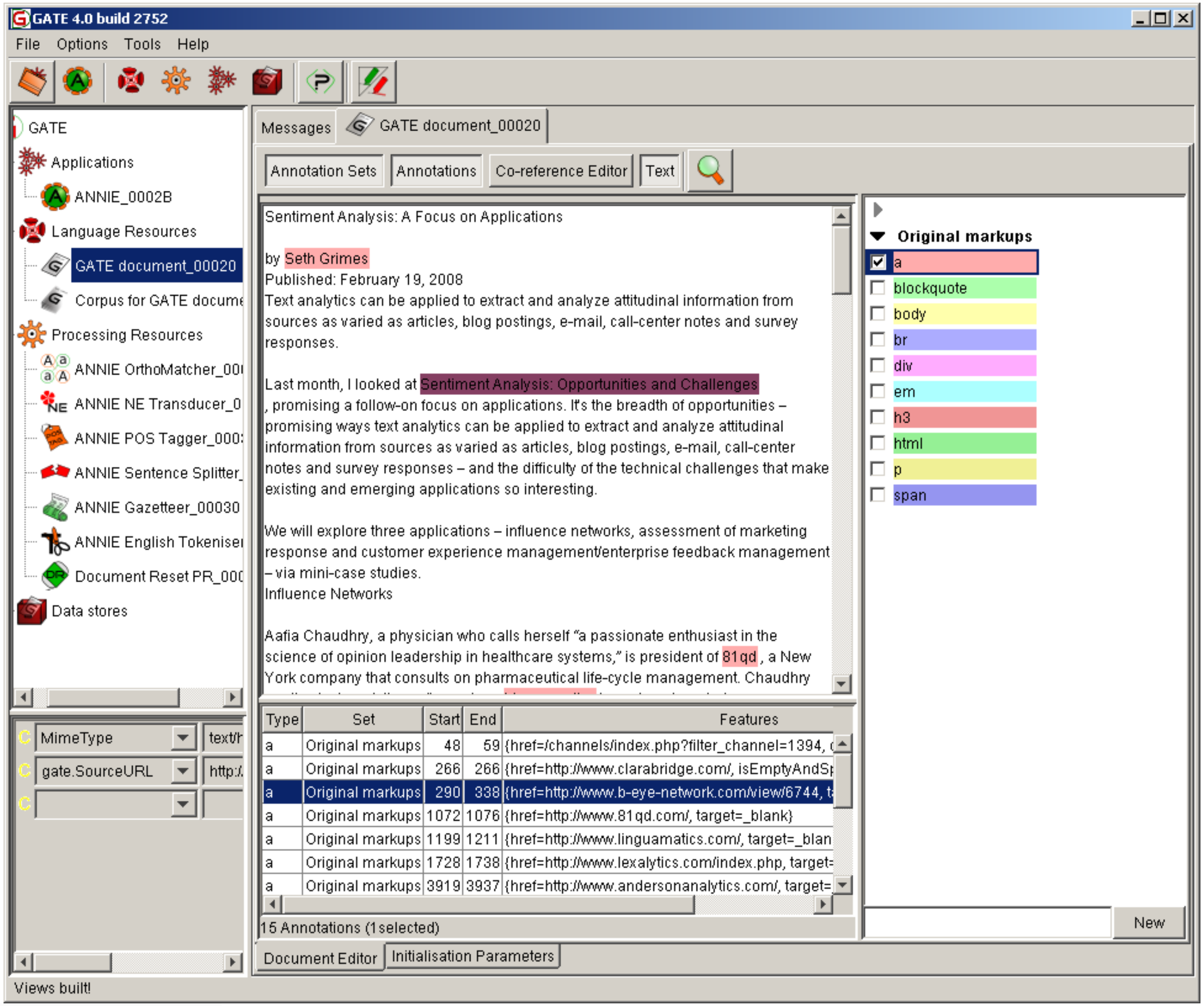
English Machinese Phrase Tagger 4.6 analysis:

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 present
use	use	main verb, infinitive
for	for	preposed marker, preposition
business	business	premodifier, noun, noun phrase begins
trips	trip	nominal head, plural noun, noun phrase ends
and	and	coordination marker

Connexor Oy, Helsinki Business and Science Park. Finland. info@connexor.com
© Connexor Oy. Powered by ToimiSait

Information Extraction

Let's see tagging in action. We'll use GATE, an open-source tool...



MimeType text/h

gate.SourceURL http:/

Sentiment Analysis: A Focus on Applications

by Seth Grimes
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.

Last month, I looked at Sentiment Analysis: Opportunities and Challenges
, promising a follow-on focus on applications. It's the breadth of opportunities – promising ways 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 – and the difficulty of the technical challenges that make existing and emerging applications so interesting.

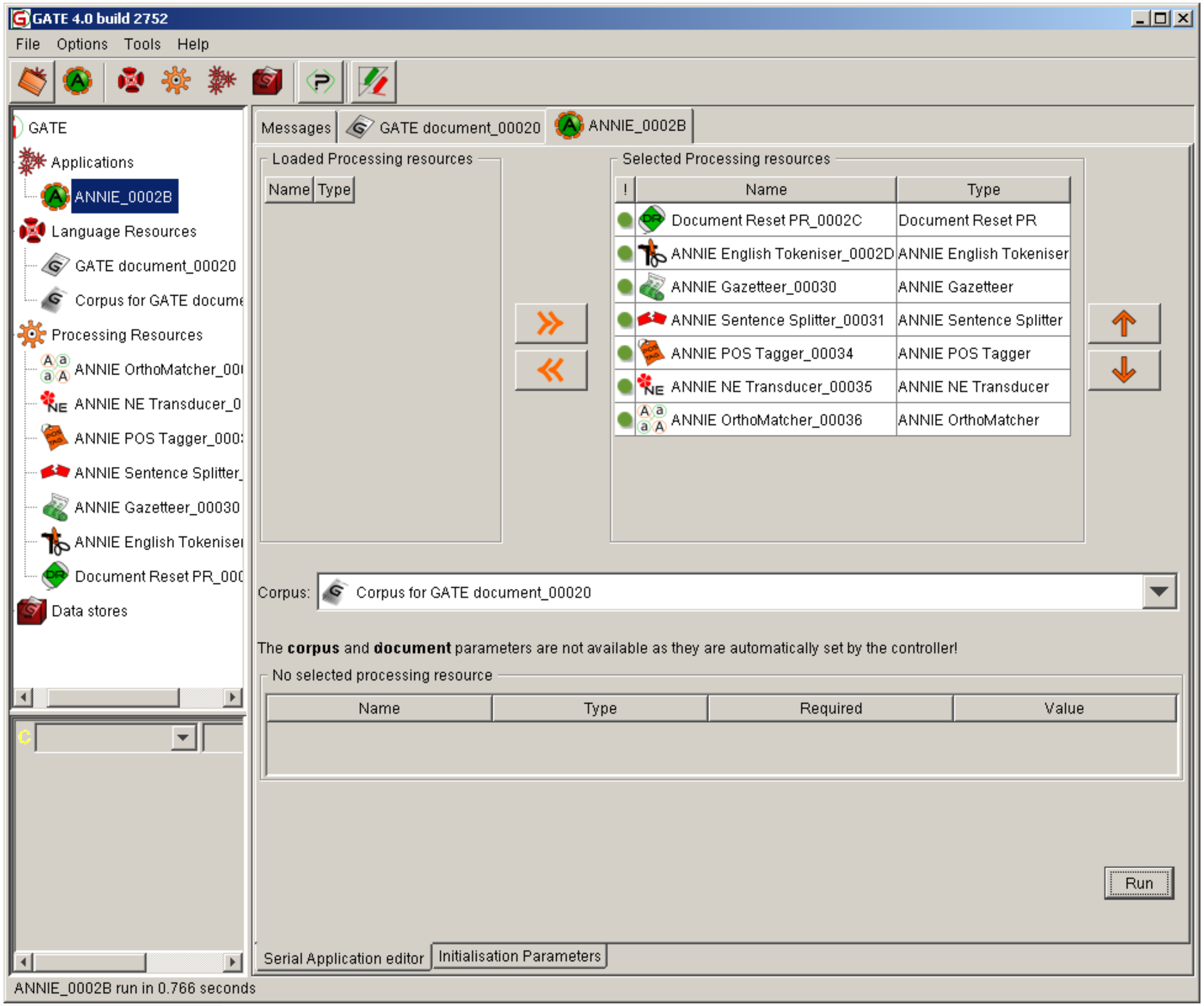
We will explore three applications – influence networks, assessment of marketing response and customer experience management/enterprise feedback management – via mini-case studies.
Influence Networks

Aafia Chaudhry, a physician who calls herself "a passionate enthusiast in the science of opinion leadership in healthcare systems," is president of 81qd , a New York company that consults on pharmaceutical life-cycle management. Chaudhry

Type	Set	Start	End	Features
a	Original markups	48	59	{href=/channels/index.php?filter_channel=1394, c
a	Original markups	266	266	{href=http://www.clarabridge.com/, isEmptyAndSp
a	Original markups	290	338	{href=http://www.b-eye-network.com/view/6744, ta
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=

15 Annotations (1selected)

Text Analytics for Dummies



!	Name	Type
●	Document Reset PR_0002C	Document Reset PR
●	ANNIE English Tokeniser_0002D	ANNIE English Tokeniser
●	ANNIE Gazetteer_00030	ANNIE Gazetteer
●	ANNIE Sentence Splitter_00031	ANNIE Sentence Splitter
●	ANNIE POS Tagger_00034	ANNIE POS Tagger
●	ANNIE NE Transducer_00035	ANNIE NE Transducer
●	ANNIE OrthoMatcher_00036	ANNIE OrthoMatcher

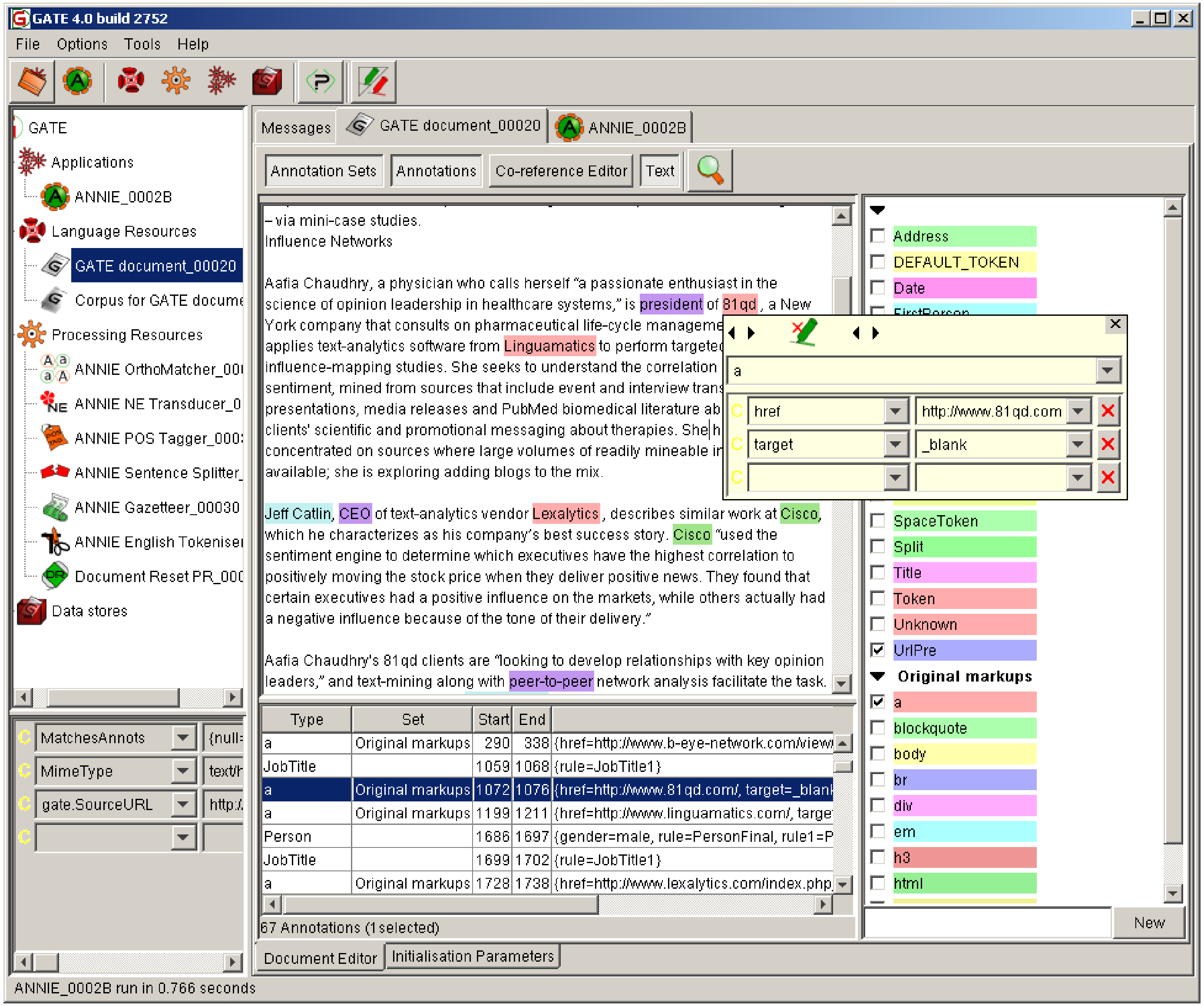
Corpus: Corpus for GATE document_00020

The **corpus** and **document** parameters are not available as they are automatically set by the controller!

No selected processing resource

Name	Type	Required	Value

ANNIE_0002B run in 0.766 seconds



– via mini-case studies.
Influence Networks

Aafia Chaudhry, a physician who calls herself “a passionate enthusiast in the science of opinion leadership in healthcare systems,” is president of 81qd , a New York company that consults on pharmaceutical life-cycle manageme
applies text-analytics software from Linguamatics to perform targetec
influence-mapping studies. She seeks to understand the correlation
sentiment, mined from sources that include event and interview trans
presentations, media releases and PubMed biomedical literature ab
clients' scientific and promotional messaging about therapies. She h
concentrated on sources where large volumes of readily mineable in
available; she is exploring adding blogs to the mix.

Jeff Catlin, CEO of text-analytics vendor Lexalytics , describes similar work at Cisco, which he characterizes as his company’s best success story. Cisco “used the sentiment engine to determine which executives have the highest correlation to positively moving the stock price when they deliver positive news. They found that certain executives had a positive influence on the markets, while others actually had a negative influence because of the tone of their delivery.”

Aafia Chaudhry's 81qd clients are “looking to develop relationships with key opinion leaders,” and text-mining along with peer-to-peer network analysis facilitate the task.

a

href	http://www.81qd.com
target	_blank

Type	Set	Start	End	
a	Original markups	290	338	{href=http://www.b-eye-network.com/view
JobTitle		1059	1068	{rule=JobTitle1}
a	Original markups	1072	1076	{href=http://www.81qd.com/, target=_blank
a	Original markups	1199	1211	{href=http://www.linguamatics.com/, targe
Person		1686	1697	{gender=male, rule=PersonFinal, rule1=P
JobTitle		1699	1702	{rule=JobTitle1}
a	Original markups	1728	1738	{href=http://www.lexalytics.com/index.php

67 Annotations (1 selected)

MatchesAnnots	{null=
MimeType	text/h
gate.SourceURL	http:/

□ Address
□ DEFAULT_TOKEN
□ Date
□ FirstPerson
□ SpaceToken
□ Split
□ Title
□ Token
□ Unknown
☑ UrlPre

Original markups
☑ a
□ blockquote
□ body
□ br
□ div
□ em
□ h3
□ html

ANNIE_0002B run in 0.766 seconds

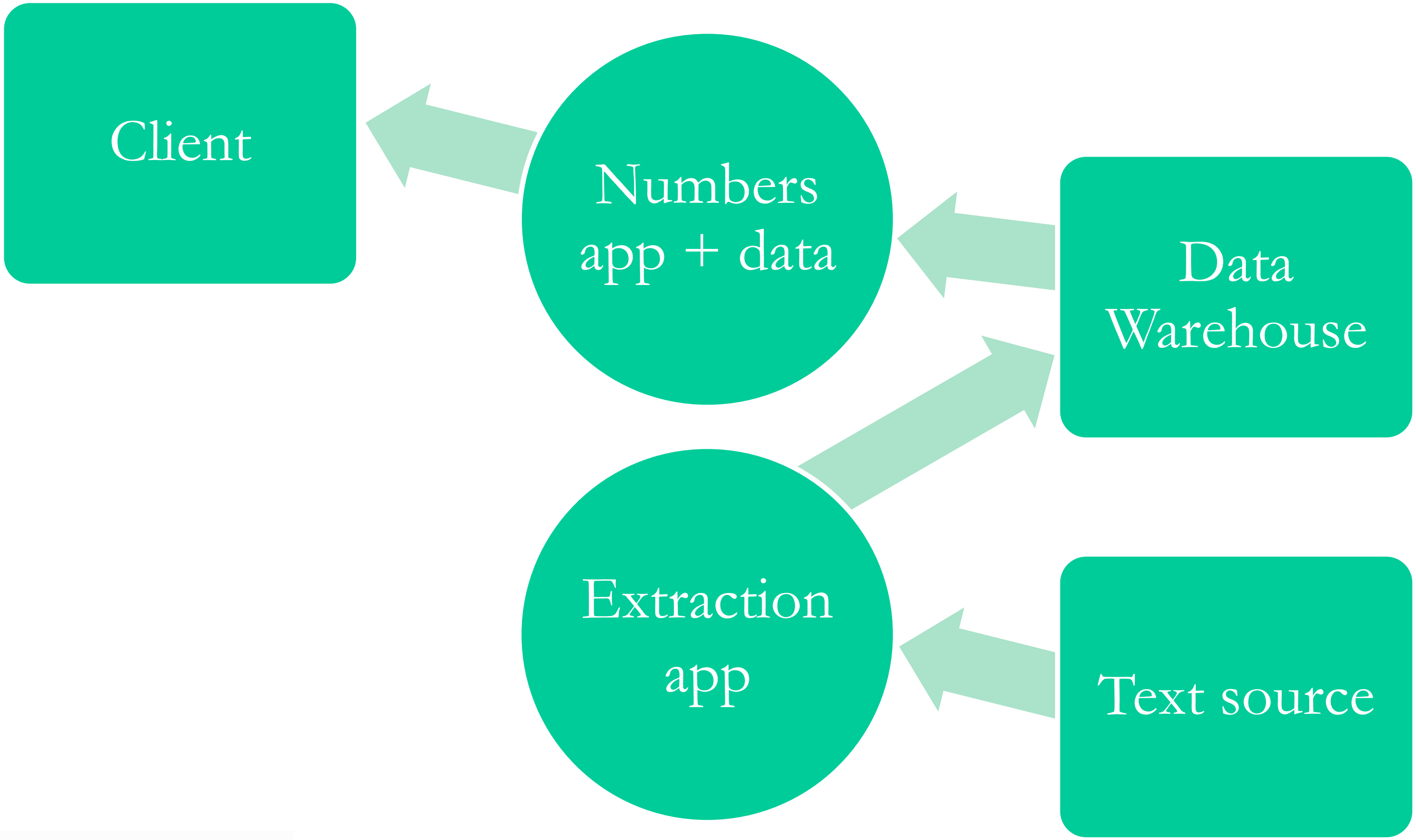
Information Extraction

For content analysis, key in on extracting information to databases.

- Entities and concepts (features) are like 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.

Information Extraction

Data integration via information extraction.



Information Extraction

XML-annotated text is an intermediate format.

```
<?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" />
<Node id="60" />Published<Node id="69" />:<Node id="70" /> <Node id="71" />February<Node id="79" />
<Node id="80" />19<Node id="82" />,<Node id="83" /> <Node id="84" />2008<Node id="88" />
<Node id="89" />Text<Node id="93" /> <Node id="94" />analytics<Node id="103" />
                                                                                    <material cut>
</TextWithNodes>
```

Information Extraction

XML-annotated text...

```
<!-- The default annotation set -->
<AnnotationSet>
                                                                                         <material cut>
<Annotation Id="67" Type="Token" StartNode="48" EndNode="52">
      <Feature>
            <Name className="java.lang.String">length</Name>
            <Value className="java.lang.String">4</Value>
      </Feature>
      <Feature>
            <Name className="java.lang.String">category</Name>
            <Value className="java.lang.String">NNP</Value>
      </Feature>
      <Feature>
            <Name className="java.lang.String">orth</Name>
            <Value className="java.lang.String">upperInitial</Value>
      </Feature>
      <Feature>
            <Name className="java.lang.String">kind</Name>
            <Value className="java.lang.String">word</Value>
      </Feature>
      <Feature>
            <Name className="java.lang.String">string</Name>
            <Value className="java.lang.String">Seth</Value>
      </Feature>
</Annotation>
                                                                                         <material cut>
</AnnotationSet>
</GateDocument>
```

Alta Plana

Example: E-mail

What else can we extract? Let's look at an e-mail message –

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

Example: E-mail

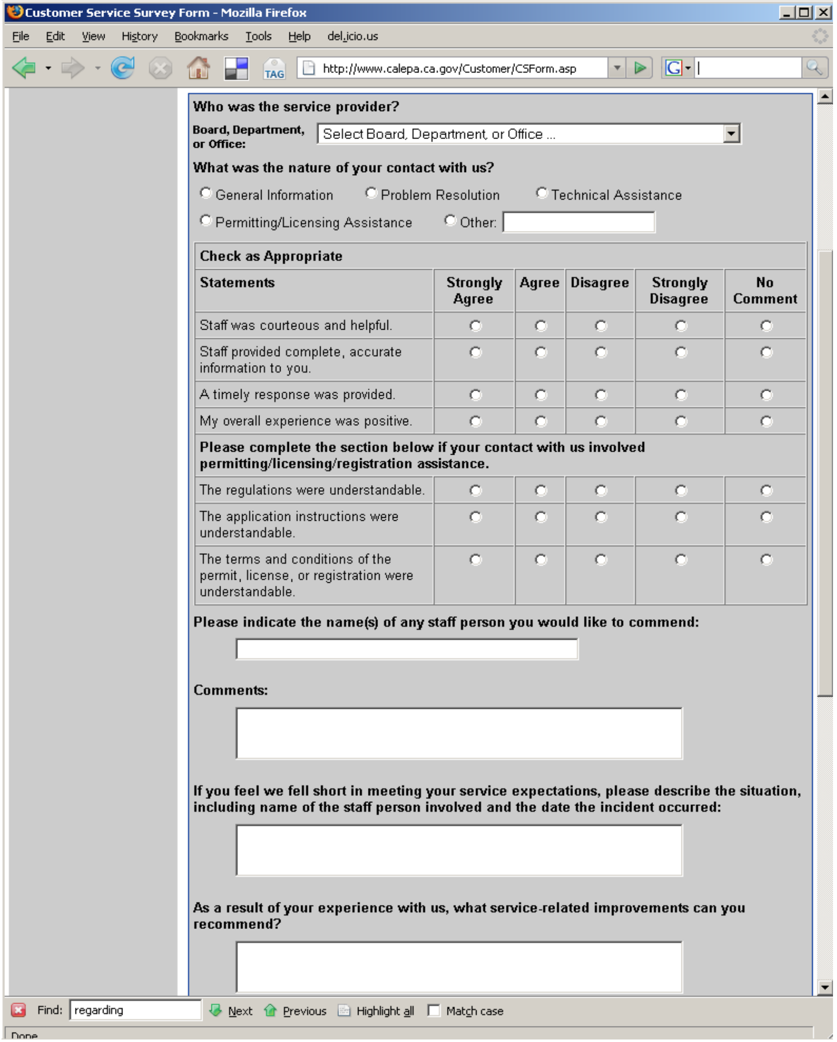
An e-mail message is "semi-structured."

- Semi=half. What's "structured" and what's not?
- Is augmentation/tagging and entity extraction enough?
- What categorization might you create from that example message?

From semi-structured text, it's especially easy to extract metadata.

There are many forms of s-s information...

Example: Survey


Customer Service Survey Form - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
TAG
Next
Previous
Highlight all
Match case
Done

http://www.calepa.ca.gov/Customer/CSForm.asp

Find: regarding

Who was the service provider?

Board, Department, or Office: Select Board, Department, or Office ...

What was the nature of your contact with us?

○ General Information ○ Problem Resolution ○ Technical Assistance

○ Permitting/Licensing Assistance ○ Other:

Check as Appropriate

Statements	Strongly Agree	Agree	Disagree	Strongly Disagree	No Comment
Staff was courteous and helpful.	○	○	○	○	○
Staff provided complete, accurate information to you.	○	○	○	○	○
A timely response was provided.	○	○	○	○	○
My overall experience was positive.	○	○	○	○	○
Please complete the section below if your contact with us involved permitting/licensing/registration assistance.					
The regulations were understandable.	○	○	○	○	○
The application instructions were understandable.	○	○	○	○	○
The terms and conditions of the permit, license, or registration were understandable.	○	○	○	○	○

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

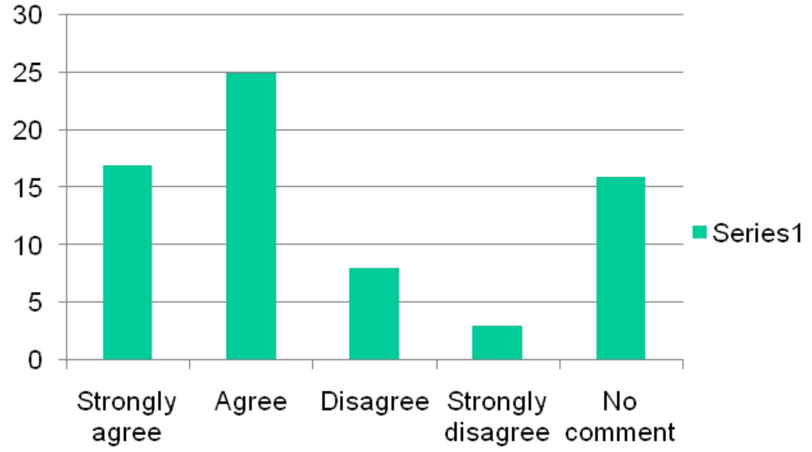
Comments:

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

As a result of your experience with us, what service-related improvements can you recommend?

Example: Survey

In analyzing surveys, we typically look at frequencies and distributions:

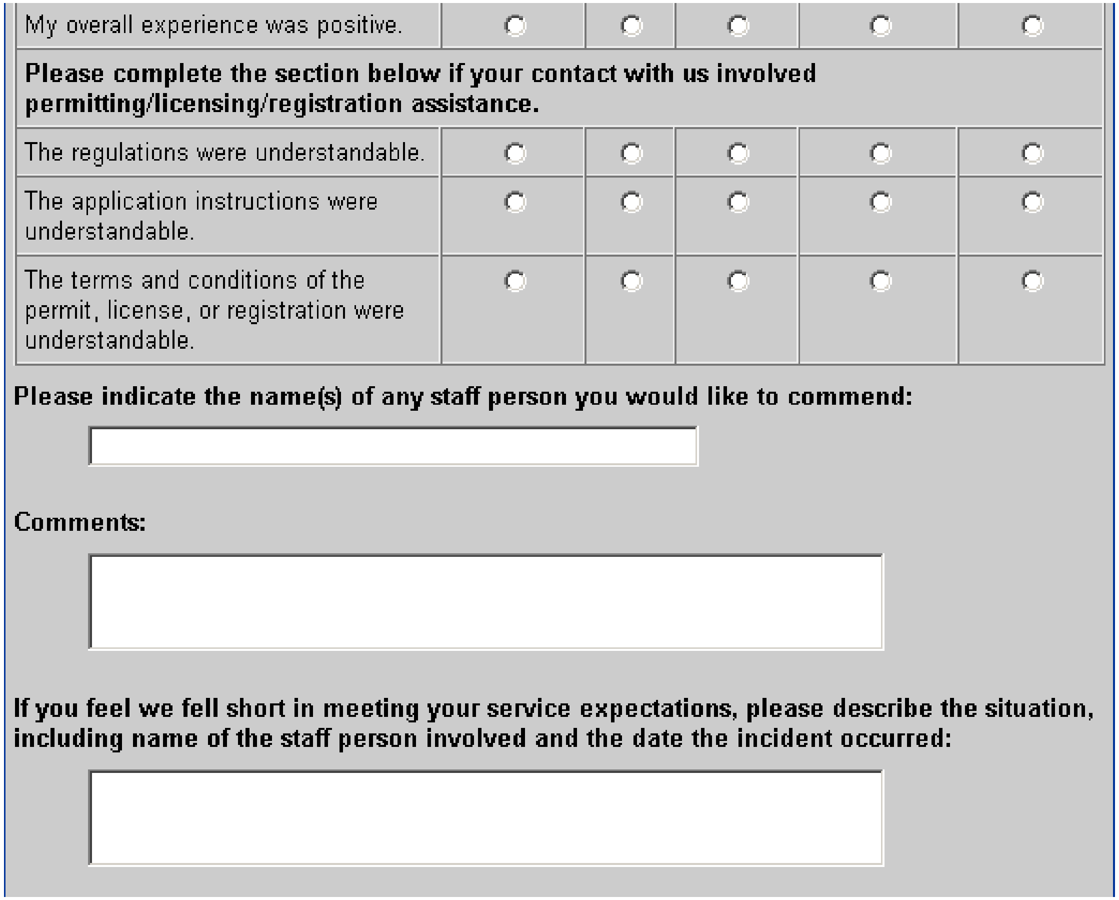


There may be fields that indicate what product/service/person the coded rating applies to.

Comments may be linked to coded ratings.

Example: Survey

The respondent is invited to explain his/her attitude:



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

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

Comments:

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

Alta Plana

Example: Survey

A survey of this type, like an e-mail message, is "semi-structured."

- Exploit what is structured in interpreting and using the free text.
- Use the *metadata* that describes the information and its provenance.
- Sentiment extraction comes into play for Voice of the Customer / Customer Experience Management applications.

Sentiment Extraction

Sentiment (opinion) extraction –

Applications include:

- Reputation management.
- Competitive intelligence.
- Quality improvement.
- Trend spotting.

Sources include:

- Wikis, blogs, forums, and newsgroups.
- Media stories and product reviews.
- Contact-center notes and transcripts.
- Customer feedback via Web-site forms and e-mail.
- Survey verbatims.

Sentiment Extraction

We need to –

- Identify and access candidate sources.
- Extract sentiment to databases.
- Correlate expressed sentiment to measures such as:
  - Sales by product, location, time, etc.
  - Defects by part, circumstances, etc.
- And information such as –
  - Customer information and customer's transactions.
- Correlation depends on semantic agreement: are we talking about the same things?

Unified Analytics

Approaches build on familiar BI tools and approaches...

Adding data and text mining...

Extracting entities, facts, sentiment, etc....

Relying on semantic integration...

...for true, 360° enterprise views.

You'll learn about lots of applications over the next two days. Good luck.

Questions?

Discussion?

Thanks!

Seth Grimes

Alta Plana Corporation

301-270-0795 – *http://altaplana.com*

Alta Plana