Text Technologies in the Mainstream: Text Analytics Solutions, Applications, and Trends

Seth Grimes

Alta Plana Corporation

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

INFORMS 2008

June 15, 2008

Introduction

Seth Grimes –

Principal Consultant with Alta Plana Corporation.

Contributing Editor, *IntelligentEnterprise.com*.

Channel Expert (text analytics), *B-Eye-Network.com*.

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

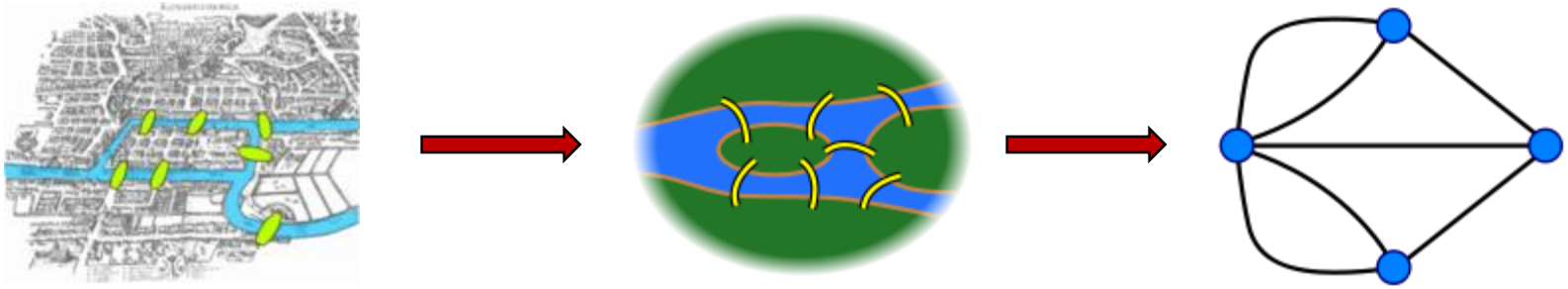
Instructor, The Data Warehousing Institute, *tdwi.org*.

©Alta Plana Corporation, 2008

What is Analytics?

$$x(t) = t$$
$$y(t) = \tfrac{1}{2}\, a\, (e^{t/a} + e^{-t/a})$$
$$= a\cosh(t/a)$$

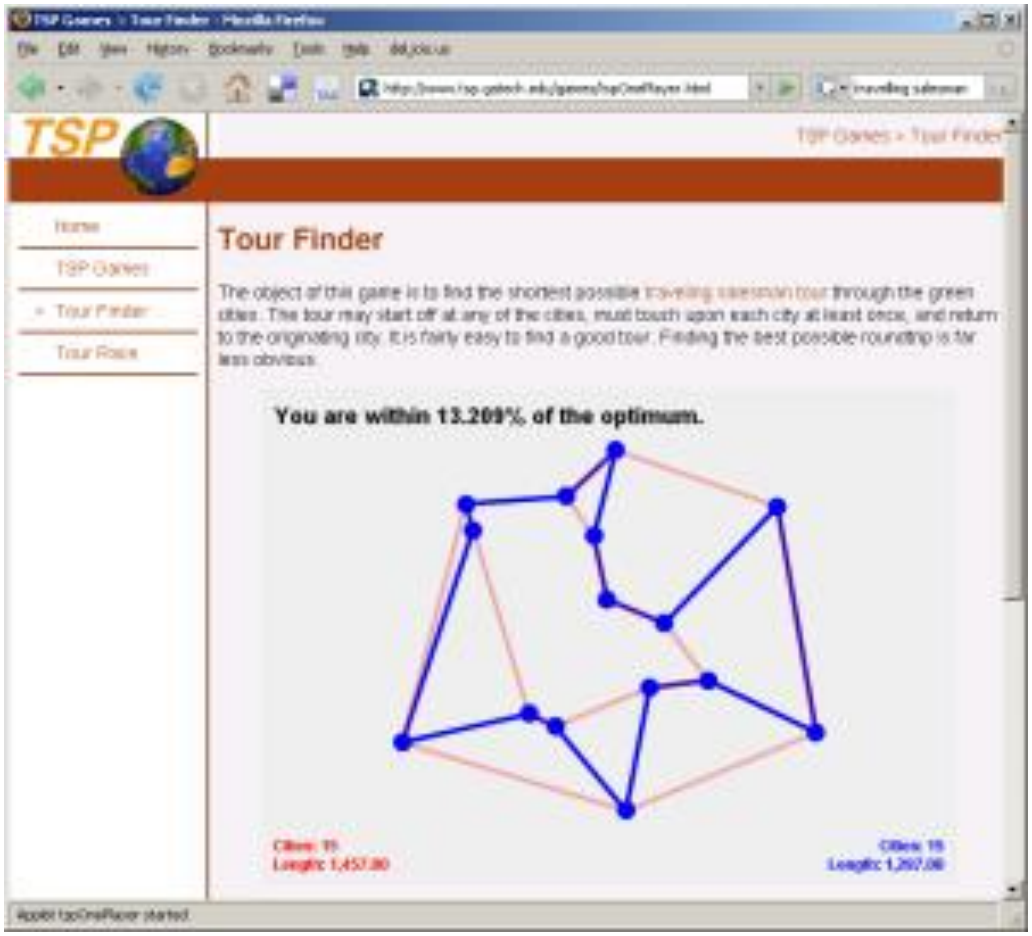
http://www.tropicalisland.de/NYC_New_York_Brooklyn_Bridge_from_World_Trade_Center_b.jpg



http://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg

©Alta Plana Corporation, 2008

What is Analytics?



©Alta Plana Corporation, 2008

What is Analytics?

What do you do when you're working with this?

```
"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
```

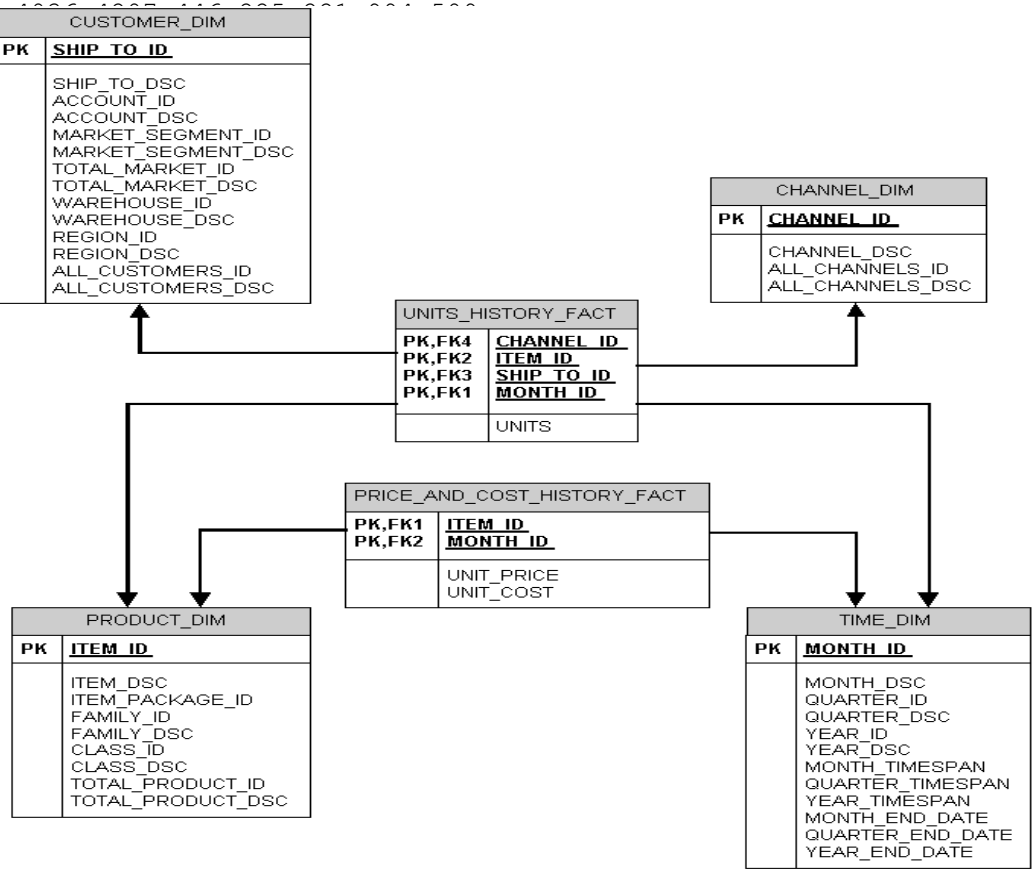

©Alta Plana Corporation, 2008

Business Intelligence

Traditional BI feeds off:

```
"SUMLEV","STATE","COUNTY","STNAME","CTYNAME","YEAR","POPESTIMATE",
50,19,1,"Iowa","Adair County",1,8243
50,19,1,"Iowa","Adair County",2,8243
50,19,1,"Iowa","Adair County",3,8212
50,19,1,"Iowa","Adair County",4,8095
50,19,1,"Iowa","Adair County",5,8003
50,19,1,"Iowa","Adair County",6,7961
50,19,1,"Iowa","Adair County",7,7875
50,19,1,"Iowa","Adair County",8,7795
50,19,1,"Iowa","Adair County",9,7714
```

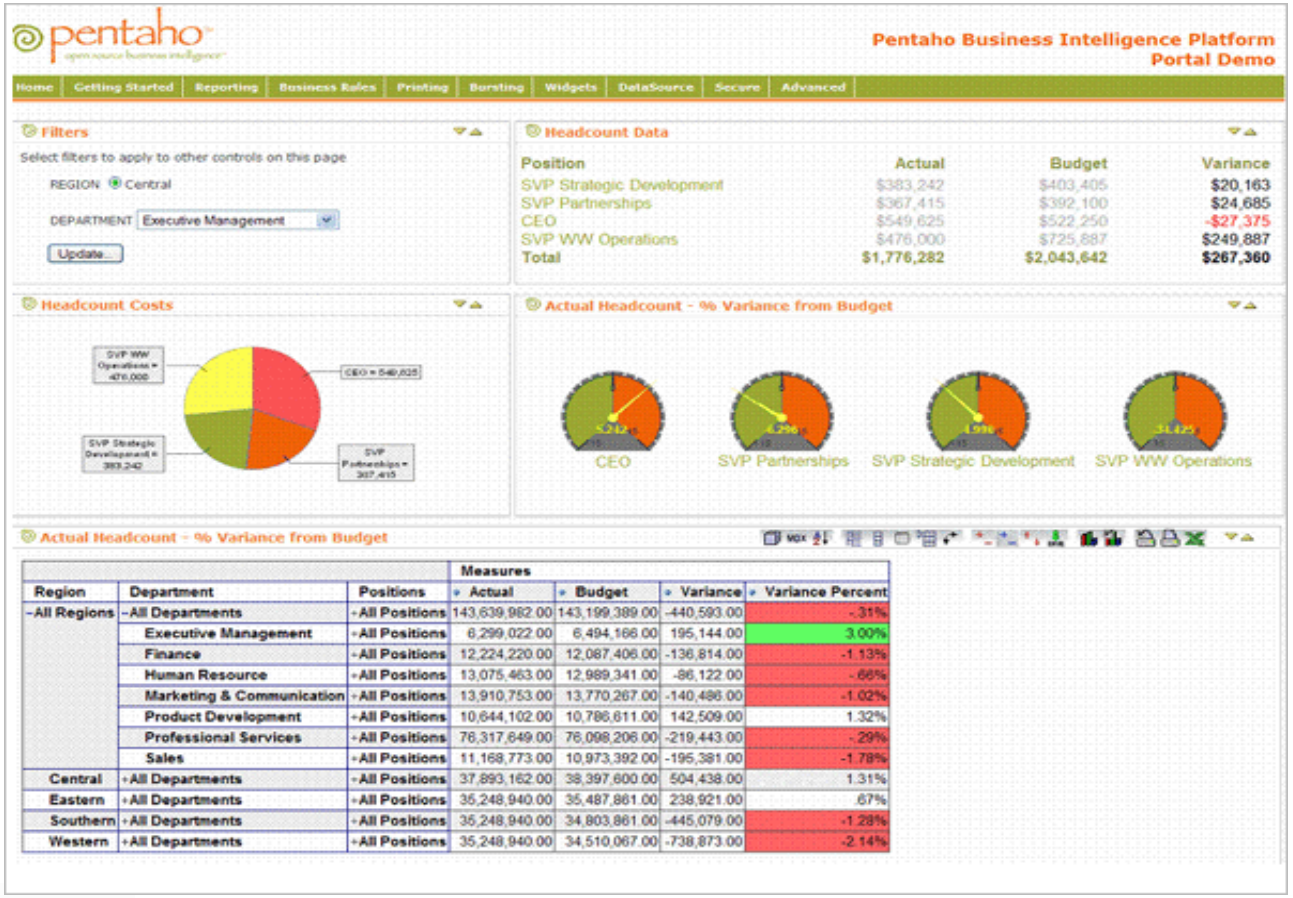
It runs off:



©Alta Plana Corporation, 2008

Business Intelligence

Traditional BI produces:



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

©Alta Plana Corporation, 2008

Business Intelligence

"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."

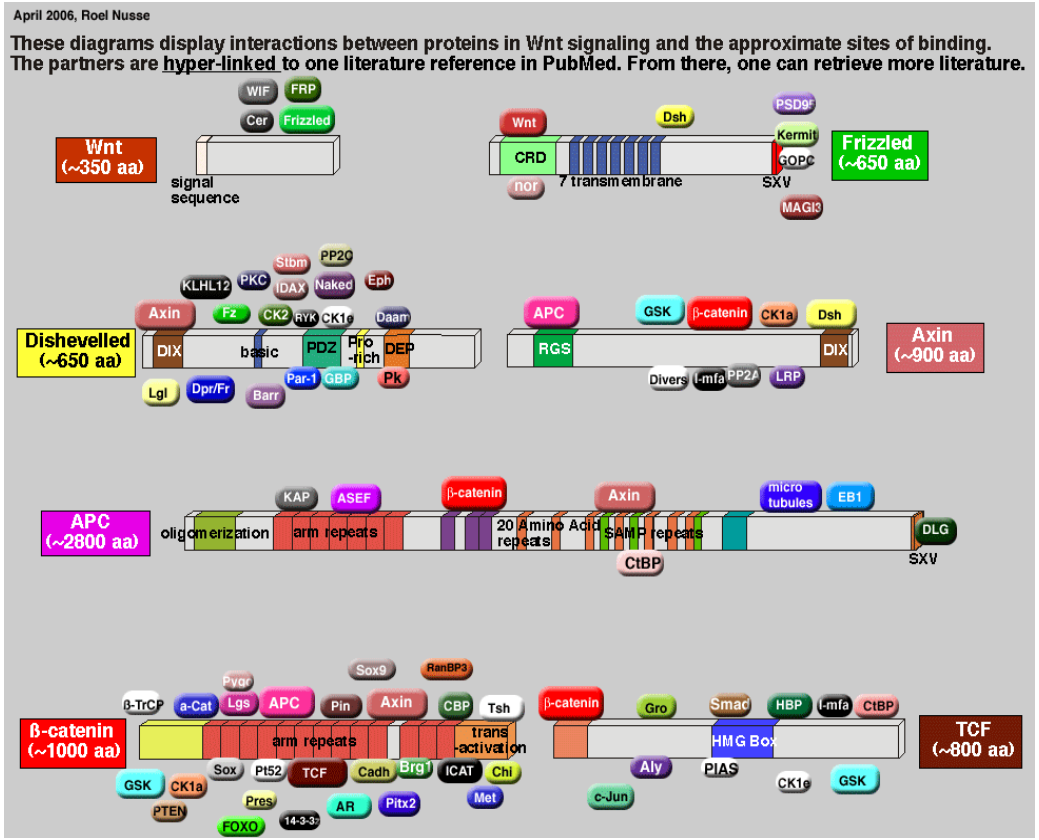
– *Prabhakar Raghavan, Yahoo Research, former CTO of enterprise-search vendor Verity (now part of Autonomy)*

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.

©Alta Plana Corporation, 2008

The "Unstructured Data" Challenge



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

©Alta Plana Corporation, 2008

Text (and Media) Technologies

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 (a.k.a. text data mining) can be automated or interactive.

©Alta Plana Corporation, 2008

The "Unstructured Data" Challenge

Consider:

E-mail, news & blog articles, forum postings, and other social media.

Contact-center notes and transcripts; recorded conversation.

Surveys, feedback forms, warranty claims.

And every other sort of document 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.

©Alta Plana Corporation, 2008

Search

Search is typically answer #1. Search involves –

- Words & phrases: search terms & natural language.
- Qualifiers: include/exclude, and/or, not, etc.

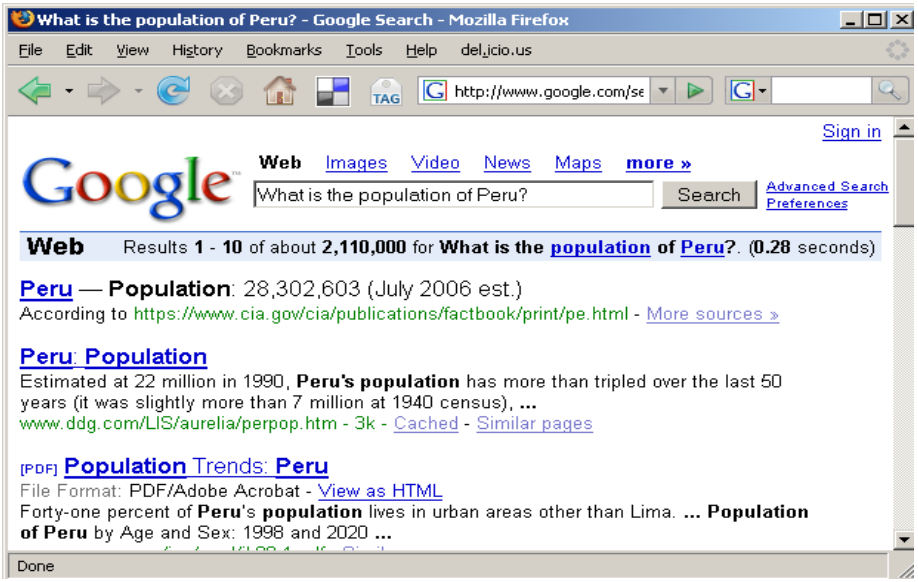
Search is not enough.

- *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.*

©Alta Plana Corporation, 2008

Search++

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

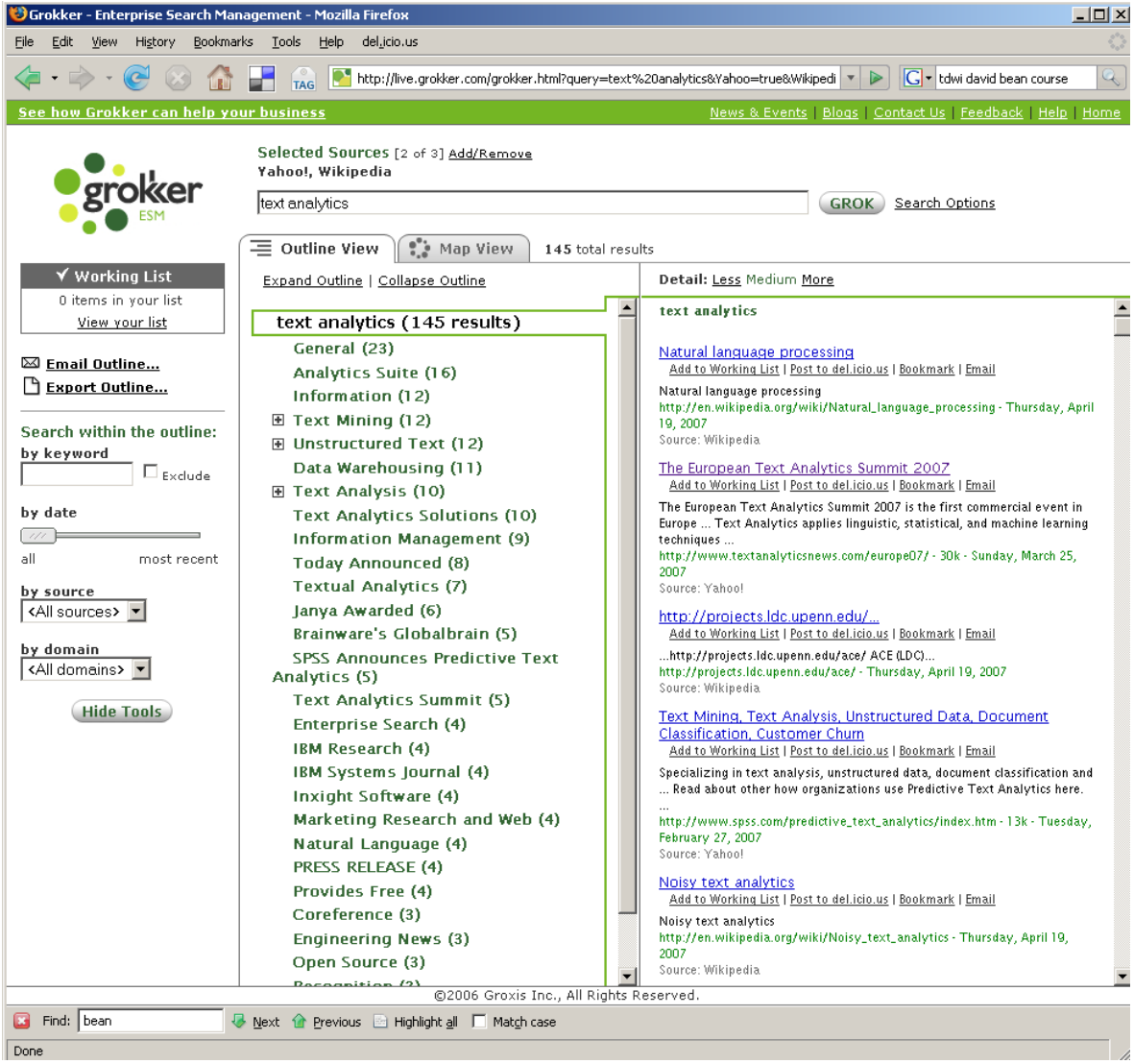


Now on to knowledge discovery, to discerning ***interrelationships of presented facts...***

©Alta Plana Corporation, 2008

Search can be pretty smart.

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

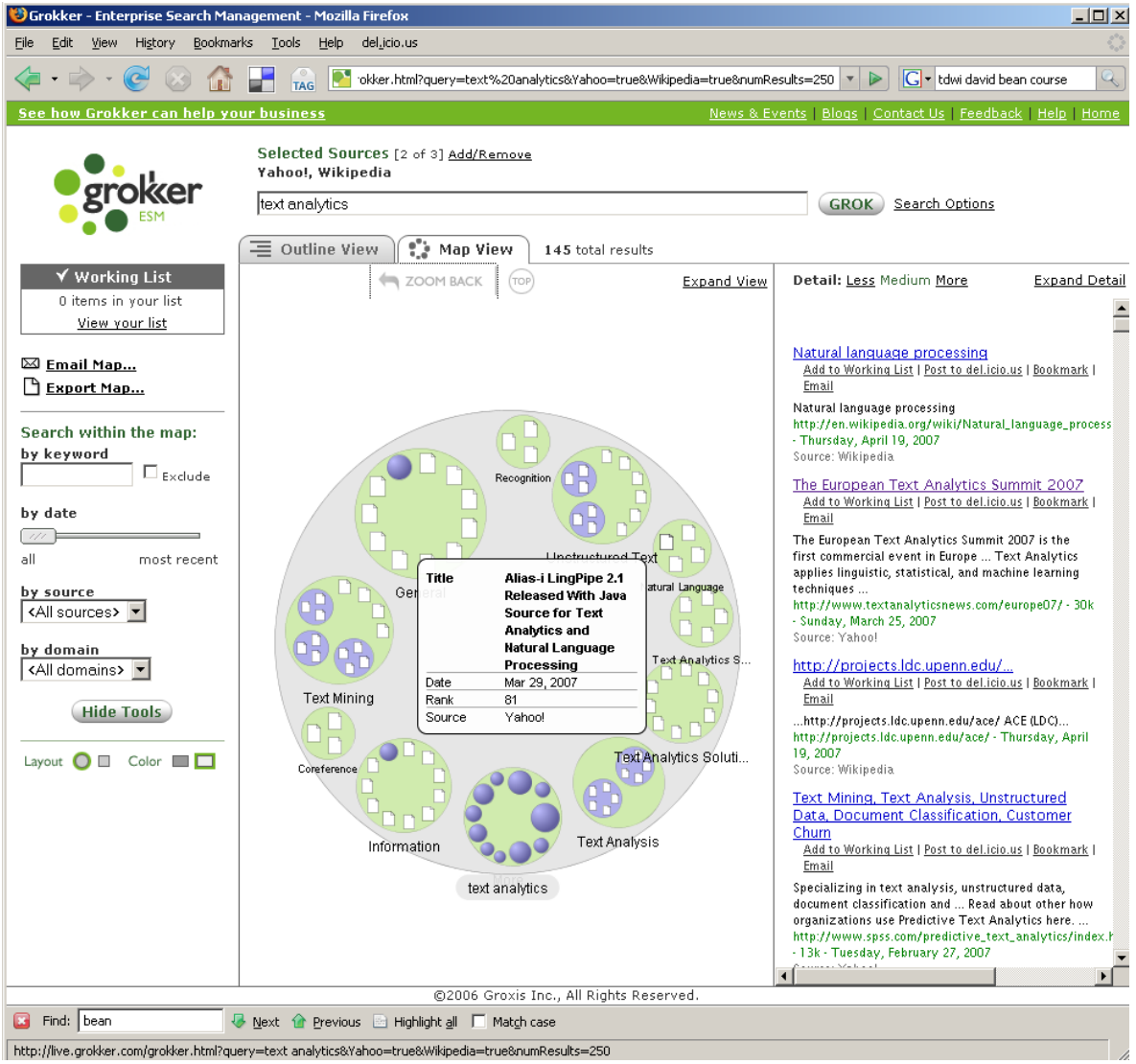


live.grokker.com/grokker.html?query=text%20analytics&Yahoo=true&Wikipedia=true&numResults=250

©Alta Plana Corporation, 2008

…with a zoomable display.

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



©Alta Plana Corporation, 2008

Search++

Text analytics can do better.

Text analytics extracts and classifies by –

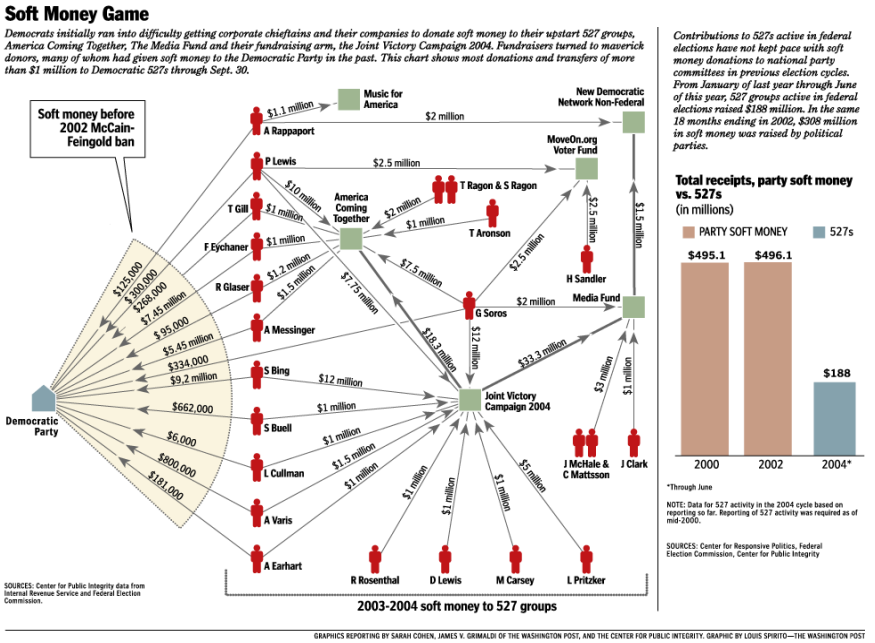
- Entities: names, e-mail addresses, phone numbers
- Concepts: abstractions of entities.
- Facts and relationships.
- Abstract attributes, e.g., “expensive,” “comfortable”
- Opinions, sentiments: attitudinal data.
- ... and sometimes data objects.

©Alta Plana Corporation, 2008

Text Analytics

Text Mining = Data Mining of textual sources.

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



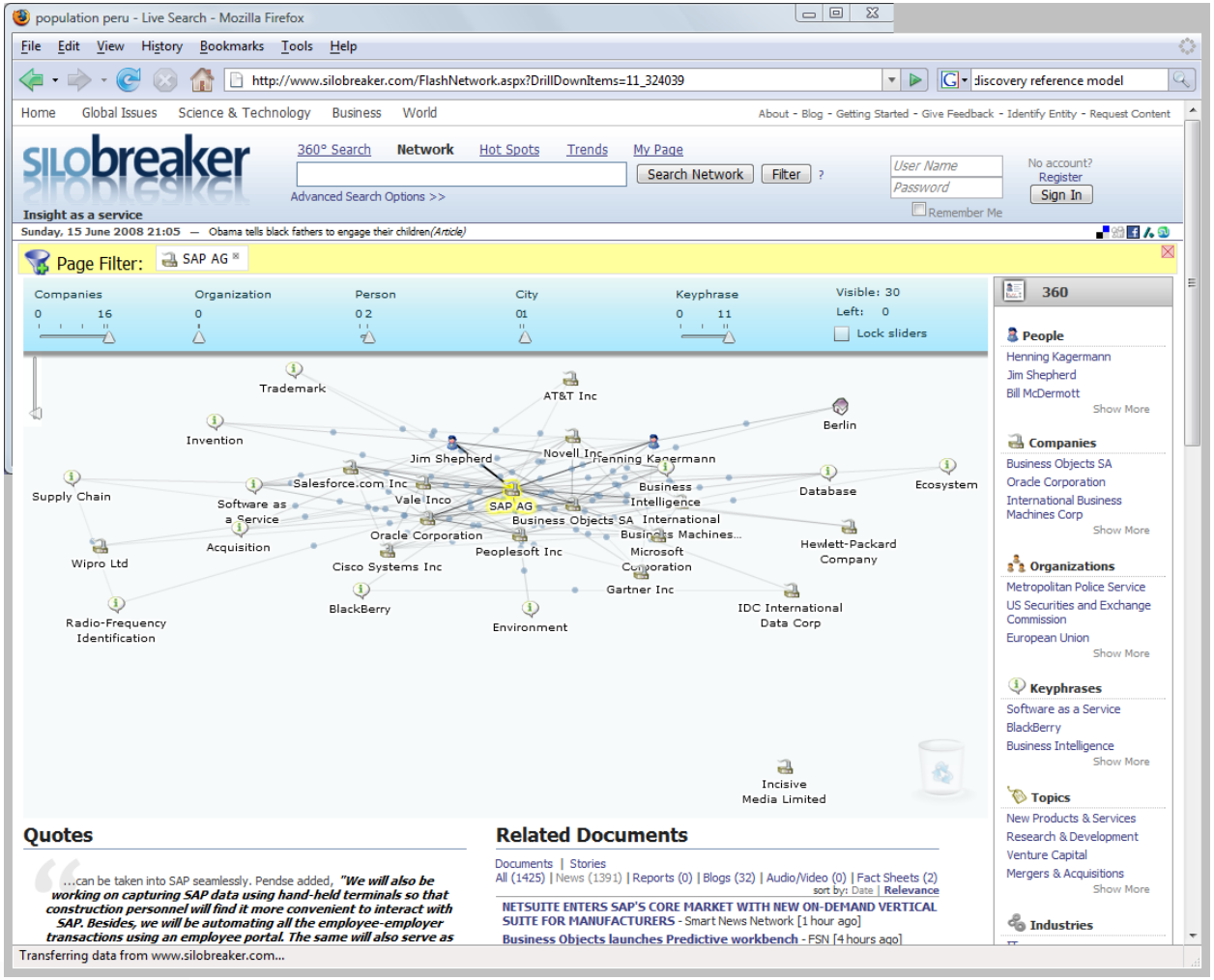
Text Mining = Knowledge Discovery in Text.

Search (Information Retrieval) is a first step.

Alta Plana

©Alta Plana Corporation, 2008

Visualizing Interrelationships



©Alta Plana Corporation, 2008

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.

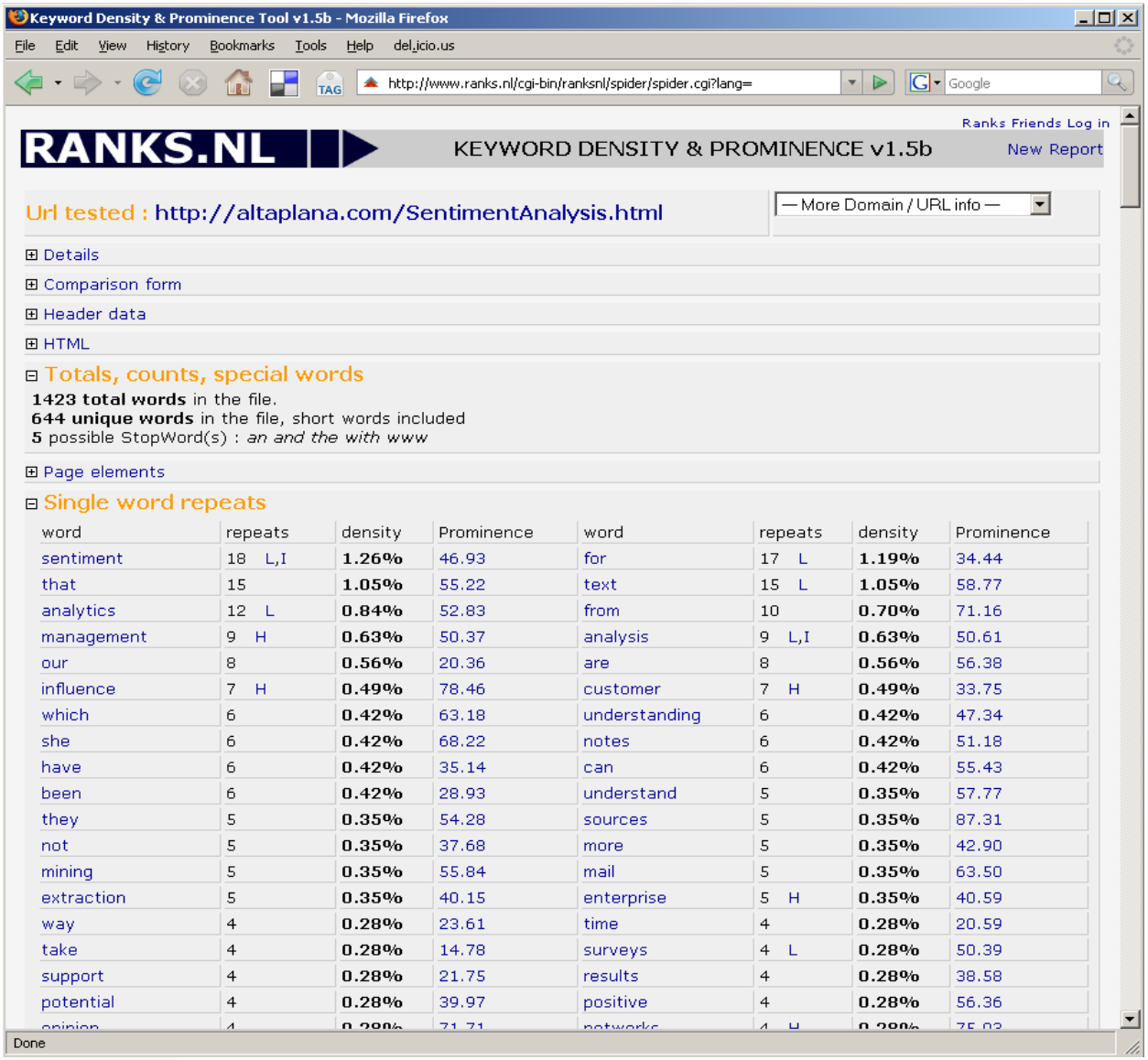
©Alta Plana Corporation, 2008

Text Analytics

Text analytics discerns linguistic and statistical structure 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, looking at an article I wrote on sentiment analysis...

©Alta Plana Corporation, 2008



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=

Ranks Friends Log in

RANKS.NL KEYWORD DENSITY & PROMINENCE v1.5b New Report

Url tested : http://altaplana.com/SentimentAnalysis.html — More Domain / URL info —

⊞ Details
⊞ Comparison form
⊞ Header data
⊞ HTML

⊟ Totals, counts, special words

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

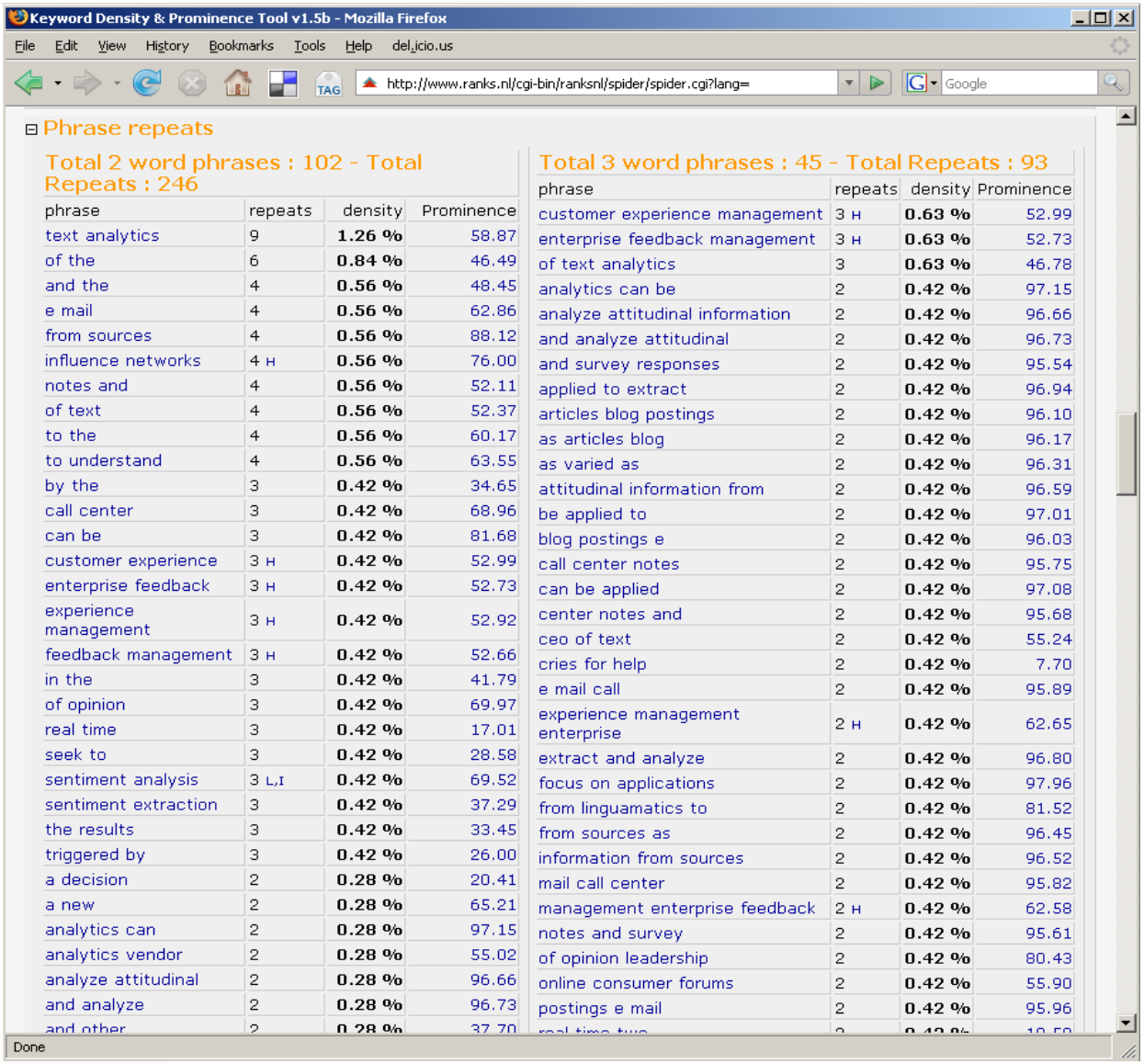
⊞ Page elements

⊟ Single word repeats

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

Done

©Alta Plana Corporation, 2008



©Alta Plana Corporation, 2008

Text Analytics

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

- Shallow parsing and statistical analysis can be enough, for instance, to support classification.
- It can help you get at meaning, for instance, by studying co-occurrence of terms.

But statistical pattern matching alone – the bag of words approach in a vector-space model – may fall short.

©Alta Plana Corporation, 2008

The Need for Linguistics

Consider –

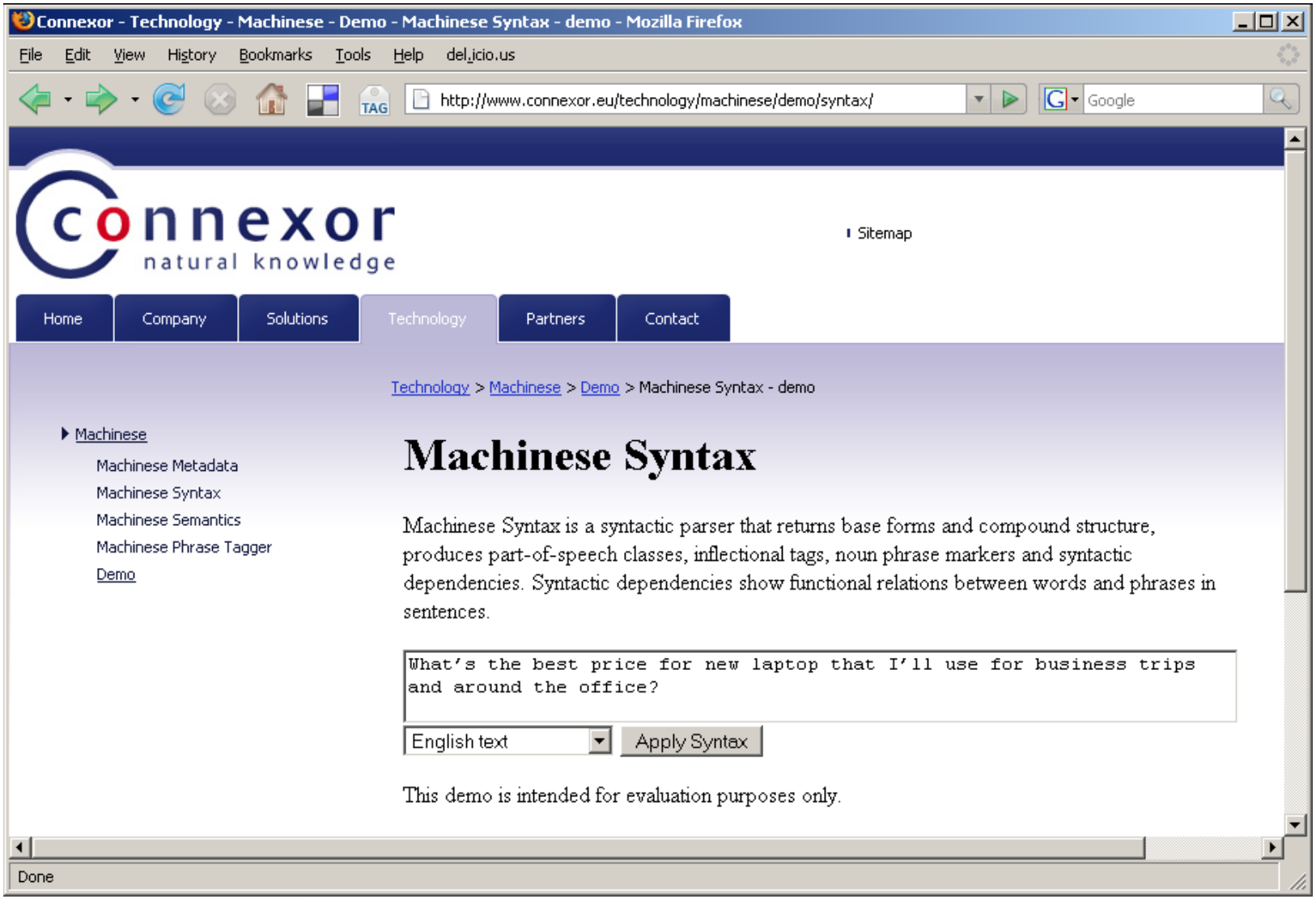
The Dow *fell* 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index fell 1.44, or 0.11 percent, to 1,263.85, and the Nasdaq composite *gained* 6.84, or 0.32 percent, to 2,162.78.

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.

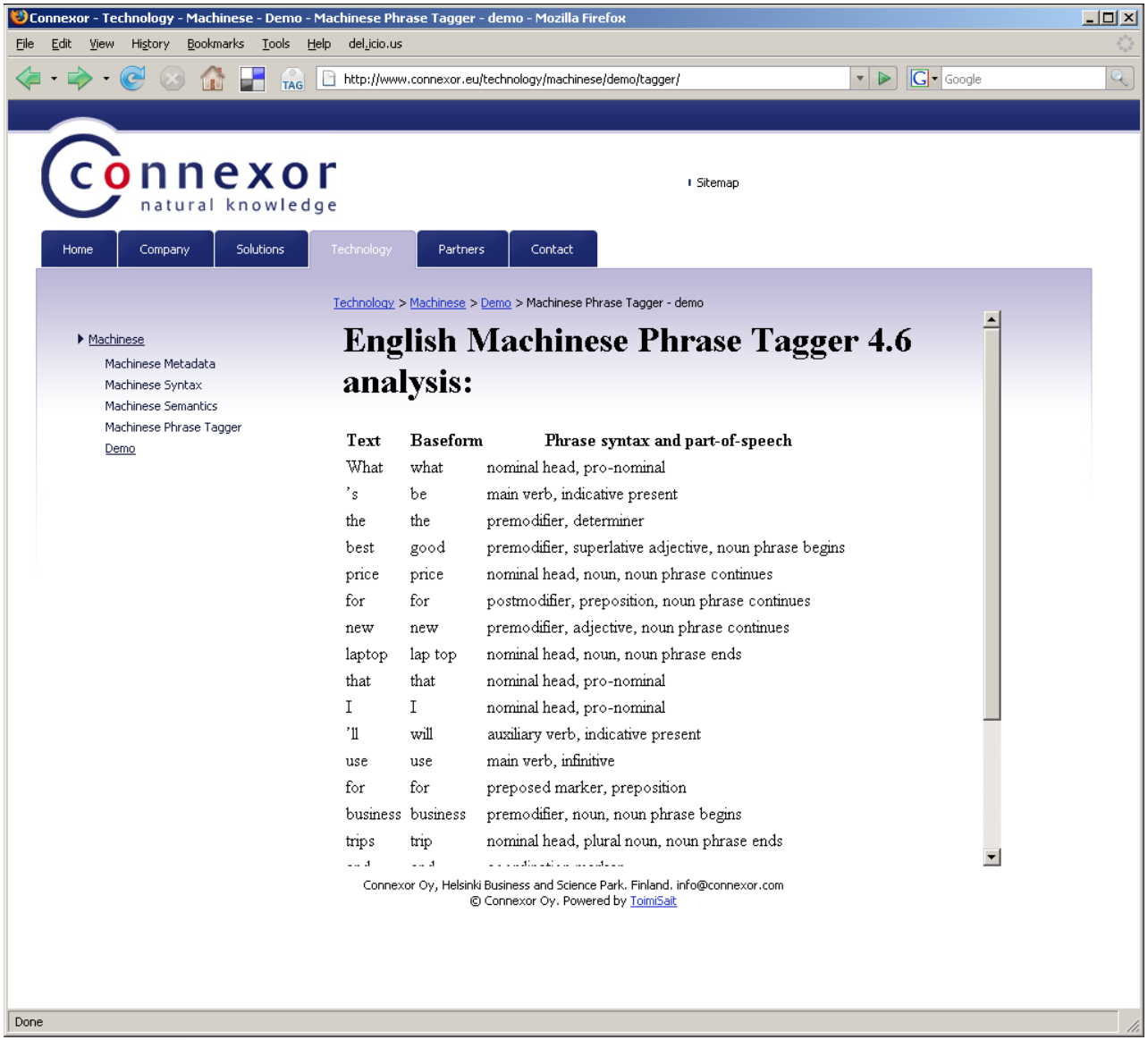
Let's try syntactic analysis of a bit of text...

©Alta Plana Corporation, 2008



©Alta Plana Corporation, 2008

©Alta Plana Corporation, 2008

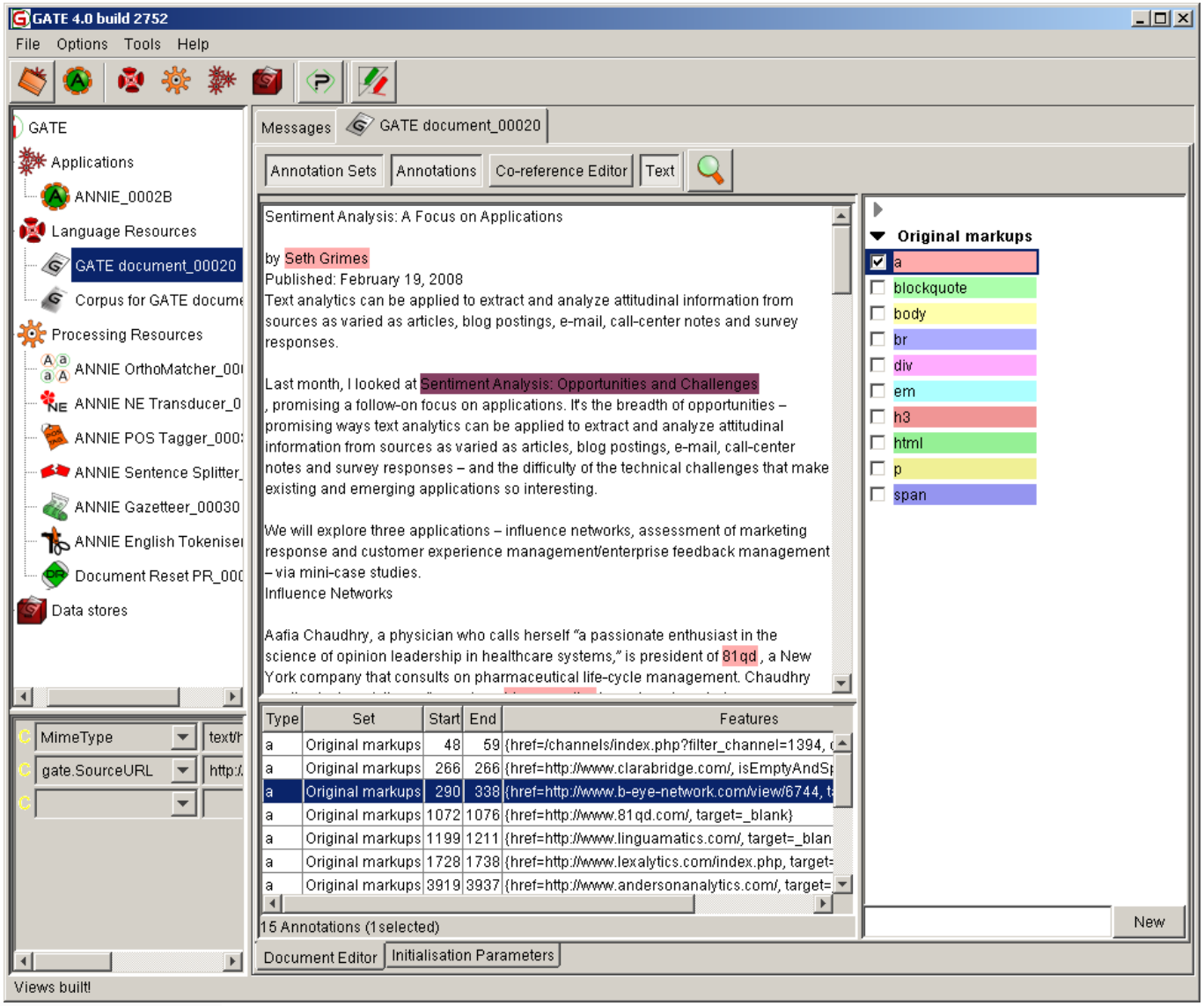


©Alta Plana Corporation, 2008

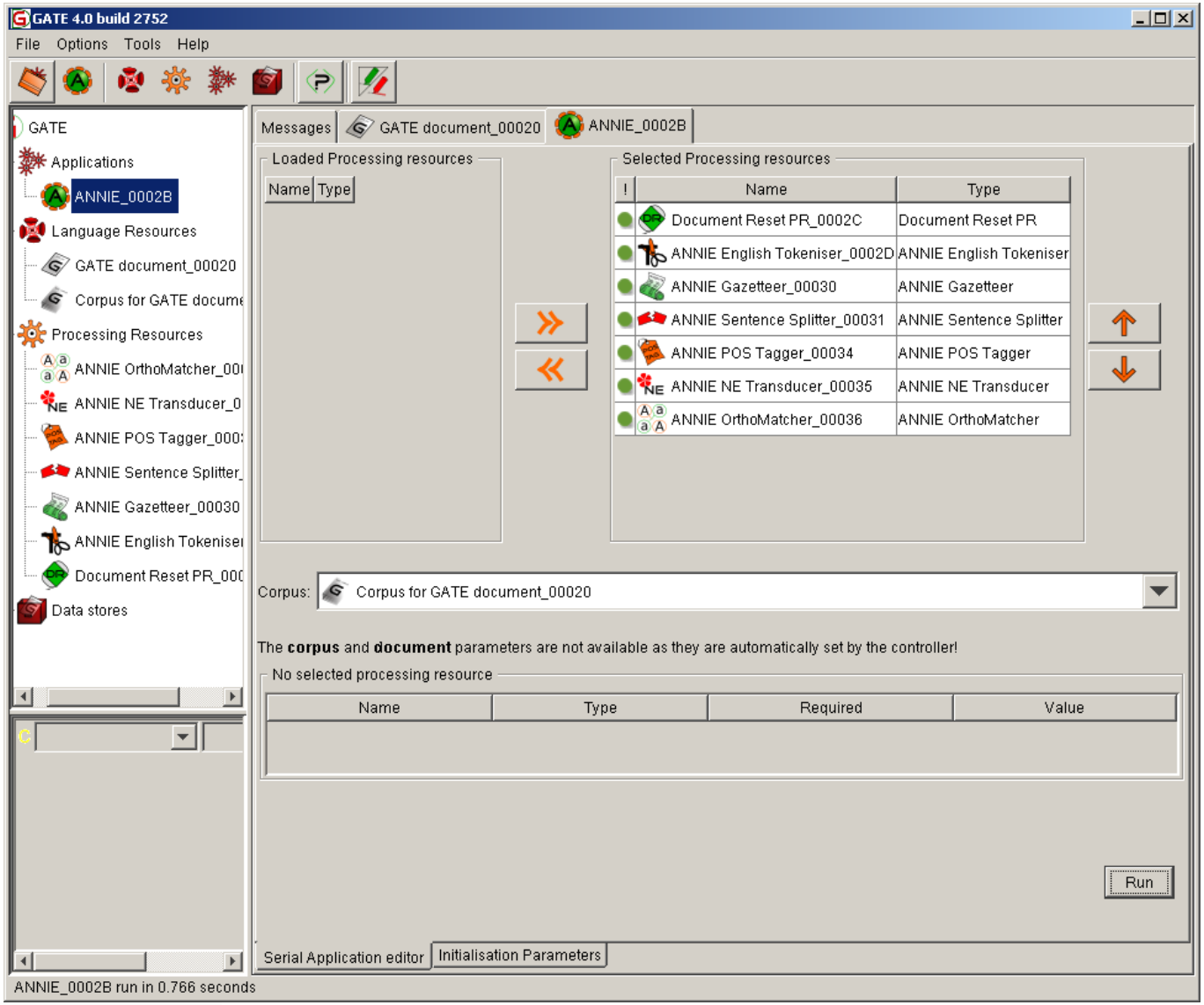
Information Extraction

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

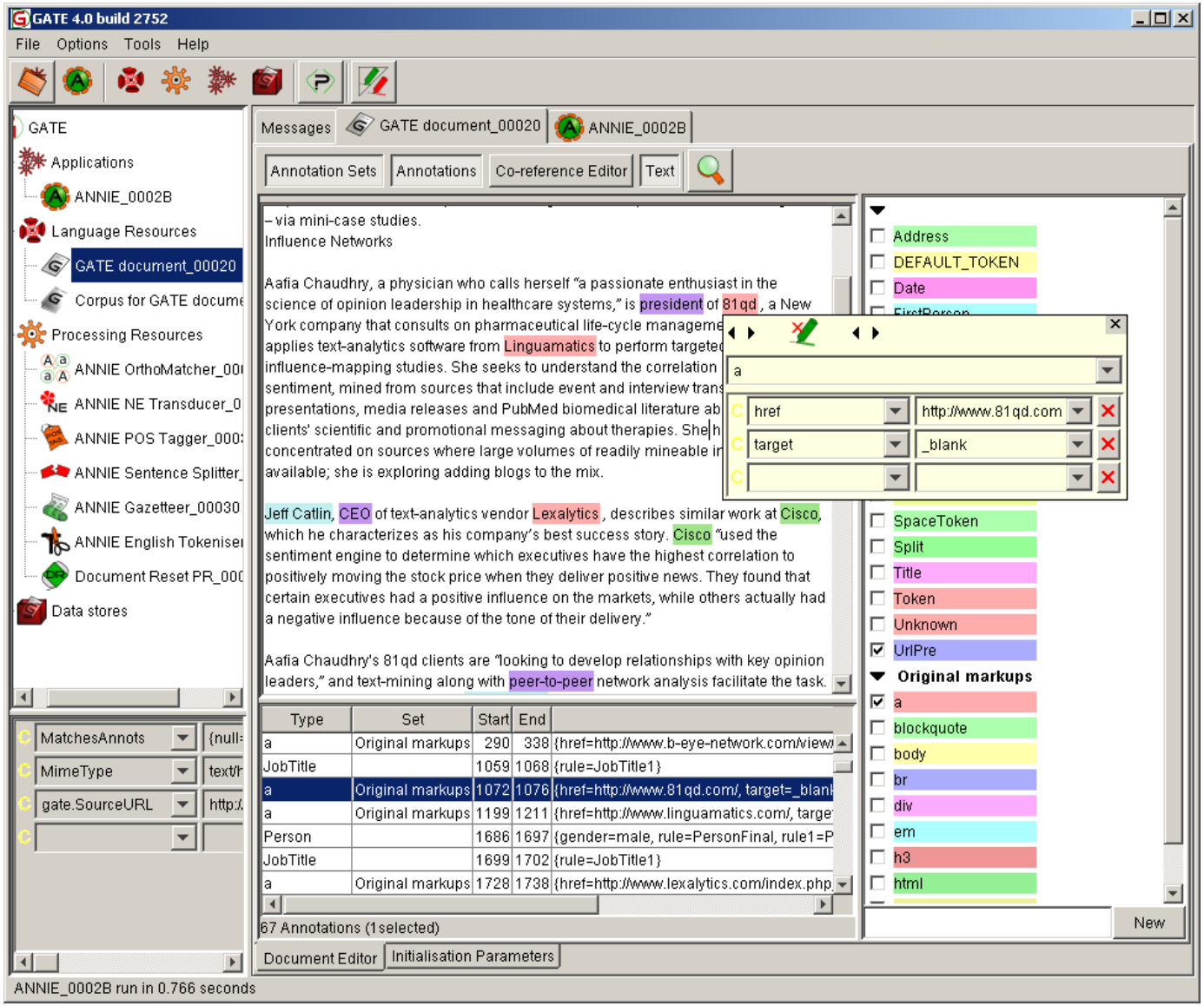
©Alta Plana Corporation, 2008



©Alta Plana Corporation, 2008



©Alta Plana Corporation, 2008



©Alta Plana Corporation, 2008

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.
- Once you've done information extraction, you can mine the data and create predictive models.

©Alta Plana Corporation, 2008

Applications

Text analytics has applications in –

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

©Alta Plana Corporation, 2008

Questions?

Discussion?

Thanks!

Seth Grimes

Alta Plana Corporation

301-270-0795 – *http://altaplana.com*

Alta Plana

©Alta Plana Corporation, 2008