

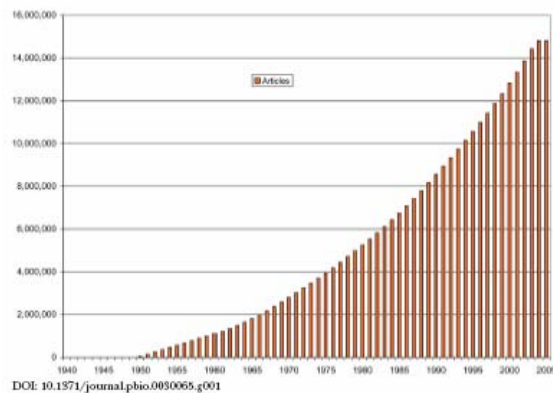
Collaboration in the Text Mine: Finding Nuggets of Knowledge in Unstructured Text

SLA PH&T Spring Meeting
March 27, 2006
Pam Kiser & Kate Lavengood
Eli Lilly and Company


Answers That Matter.

Medline added 1564 citations a day in 2004¹...

Growth of Medline, 1940-2005²



¹ <http://www.nlm.nih.gov/pubs/factsheets/medline.html>

² Schuhmann, PLoS Biology, 3(2) Feb 2005, e65.



Changing role of Info Pros

Pre-1970s:

- Paper indexes
- Pointing patrons to resources
- Maintaining the collection

Answering
questions

1970-2000:

- Dialog (and others) computerized bibliographic databases
- Helped to manage the amount of info available
- Provided stacks of documents
- Internet searching

Understanding
customer needs

2000- :

- Researchers want answers, not documents
- Want sources integrated into one answer
- Want trend analysis
- New analysis and visualization techniques

Working in
Collaboration



Text Mining: The discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.

(Marti Hearst, Proc. Assoc. Comput. Linguist., 37, 3-10 (1999)).

...allows you to find answers to questions,
not just documents that contain certain
words

Three uses of text mining

See the big picture

- What types of treatments are associated with Chronic Fatigue Syndrome?

Find answers to very specific questions

- Which genes up-regulate a specific protein?

Hypothesis generation

- What are potential new therapeutic applications for thalidomide?

Linguistics for Librarians

Don't be intimidated by the jargon!

- The language of linguistics can be a sea of confusion – all you really need is to have a basic understanding of the key concepts
- Beware of different interpretations of the same term (e.g., ontology)
- Understand just enough about linguistics to be able to:
 - influence the initial build structure
 - evaluate output on a text mining project
 - understand the underlying methodology used by the tool



And exactly what goes on “under the hood” in a text mining system?

Parsing

A) Breakdown text into ‘parts’

- Reorganize into a meaningful structure that can be explored

Tagging

B) Information Extraction

- Annotate terms with contextual meaning



A) Breakdown text into ‘parts’

TOKENIZING

Text segmentation into word-like units

Figuring out that *icecream* and *ice cream* and *ice-cream* all mean the same thing and should be recognized as the same thing

MORPHOLOGY
& STEMMING

Variant forms of the same term

Multiple forms of same word— e.g., present, past, singular, future

- Example: “to be” = will, was, were
- Example: mouse vs. mice
- Example: MeS vs. mes



A) Breakdown text into 'parts' continued

SYNTAX

Understanding sentence structure

- Example: Examine the word order [AI hit Bill vs. Bill hit AI]
- Example: Identify phrases: “memory bank” – “food bank” – “blood bank”

DISAMBIGUATION

Clarifying context

- Recognizing that *heart attack* and *myocardial infarction* mean the same thing and should be recognized as synonyms
- Recognizing correct sense of a word in context – e.g., “cool”



B) Information Extraction

POS
Tagging

Understanding sentence structure

- Example: identify *nouns* – *verbs* - *prepositions*

Semantic
Tagging

Identify entity types

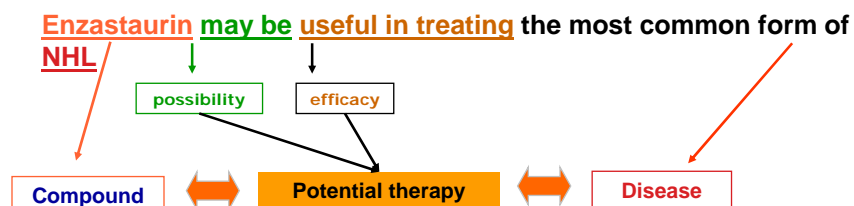
- Assign descriptive classification – *company name* or *compound name*
- Recognize variations in same entity: *LY-246736* & *LY246736*
- Identify named entities: *author*, *journal name*

Relationship
Tagging

Identify relationships between entities

- Strep* is a common cause of sore throats

What Does Text Mining Do ?



Text mining recognizes the relationships between words and allows a user to ask very specific questions with the help of codified knowledge or ontologies

Step 2: Organization of the 'parts'

- **Thesaurus**

“a list of subject headings or descriptors usually with a cross-reference system for use in the organization of a collection of documents for reference and retrieval”

(Source: Merriam-Webster Online Dictionary)

- **Taxonomy**

“in a broad sense, the science of classification, but more strictly the classification of living and extinct organisms—i.e., biological classification.”

(Source: Encyclopaedia Britannica Online)



**Data
Integration**

Building the infrastructure

- **Ontology**

“An explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them.”

(Source: The Free On-line Dictionary of Computing (2003-OCT-10))

“Ontologies have to be built and maintained, and this is a difficult, labor-intensive, and expertise-intensive task.” (IBM Systems Journal, Vol 43, No 3, 2004, p507)



Striking Gold

Considerations before you enter the mine:

- Careful selection of the core **documents** to be mined (the corpora) will increase the chances of better text mining results
- Thoughtful evaluation of the **sources** used by the vendor to create the underlying structure will also lead to better results
- Input on which **taxonomies** are incorporated can significantly influence usability of final tool

Garbage in, Garbage out (GIGO)

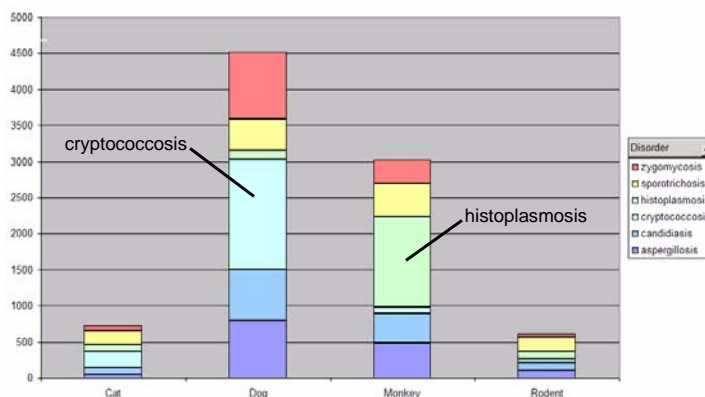


Results: Basic Excel Spreadsheet Value of Media Acquisitions

Company	verb phrase	currency	Company	sentence
MacAndrews & Forbes Holdings	acquisition	\$750 million	Deluxe film processing	MacAndrews & Forbes Holdings Inc. today announced that it has completed its approximately \$750 million acquisition of the Deluxe film processing and creative services business
Ripplewood's and ZelnickMedia's	acquired	\$60 million	Lillian Vernon Corp.	Ripplewood's and ZelnickMedia's acquired catalog retailer Lillian Vernon Corp. for \$60 million.
Time Warner Inc and Comcast Corporation	acquire	\$12.7 billion	Adelphia Communications Corporation	Time Warner Inc. and Comcast Corporation today announced that they have reached definitive agreements to acquire substantially all the assets of Adelphia Communications Corporation for a total of \$12.7 billion



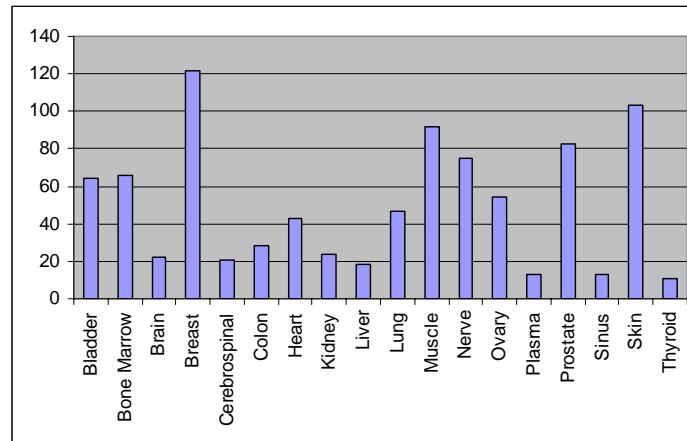
Pivot Table of Excel Results Fungal Infections by Animal





Innovation Through Information

Bar chart of Excel results Protein Occurrence in Body Tissue



March 27, 2006
Collaboration in the Text Mine

Company Confidential
Copyright © 2006 Eli Lilly and Company

17



Innovation Through Information

Clustering Who is Martha Stewart?



March 27, 2006
Collaboration in the Text Mine

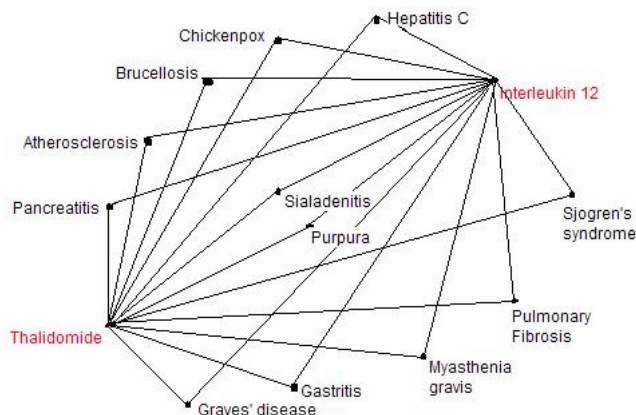
Company Confidential
Copyright © 2006 Eli Lilly and Company

18



Visualizing Indirect Relationships

What diseases might be treated with Thalidomide?



Source: Data from Weeber et al, J Am Med Inform Assoc. 2003; 10:252-259.



Danger: Don't enter the mine without a librarian

Information professionals are natural partners for text mining with existing skills

- Knowledge of the info superhighway and the ability to place information in context
- Knowledge of available products and techniques
- Strong command of scientific and/or clinical language
- Subject area experience/expertise
- Blend of analytical and creativity skills
- Problem solving skills and ability to deal with ambiguity
- Consultative and listening skills
- Ability to adapt and try different approaches to problems



Specific roles for information professionals in text mining projects

- Facilitate conversations between internal teams and vendors
- Help place tool into context with other information sources
- Advise on source selection for the specific information question
- Advise on search strategies to retrieve content set
- Consult with vendor on appropriate taxonomies/ontologies
- Evaluate what's "under the hood"
- Identify application areas most appropriate for text mining
- Set appropriate expectations
- Help customer evaluate and manipulate results



The Buddy System: Never Mine Alone

Collaboration

...with Scientists/Business SME

- Help to articulate the question (needs to be a real business question)
- Need buy-in from the scientist that they see this as a valid approach and will work through the results
- Need their domain knowledge to actually find an answer
- Need their tacit knowledge of the area



Collaboration

...with the Vendors

- Facilitate conversations (the scientist and the vendor speak different languages)
- Provide the corpus
- Monitor the process
- Help to query system and present results in a way that the expert can work with



Collaboration

...with IT

- Projects frequently result in huge files that need IT expertise
- System may be brought in-house
- IT knows IT but not the files/systems they are working with



Collaboration

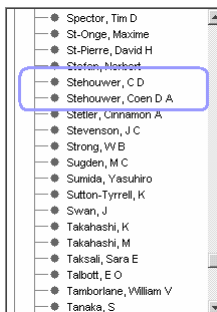
...with your Colleagues

- Need to understand what you're working on and why it's important
- Need to understand if and when they should be involved
- Help promote text mining as a new information tool
- Build relationships for future projects

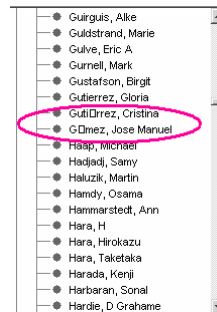


Prudent Mining Practices

Using your headlamp in the mine (or where things go wrong)



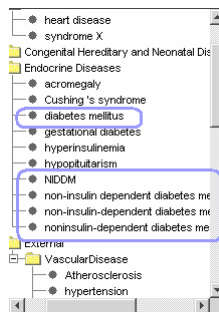
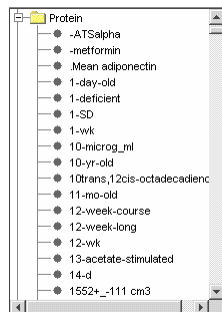
Author names not 'normalized'



Or contain non-text characters

Where things go wrong (continued)

Terms in list don't make sense



Concepts not "rolled up"

Setting user expectations:

- Be sure to define what you mean by "public sources"
- The "mine" will likely cover only article abstracts not fulltext
- The *clinical literature* is very different than *clinical data*



Lessons Learned

- You don't need a degree in computational linguistics to add value in text mining applications – and don't be intimidated by the jargon!
- Text mining is still evolving and there is much to be improved
- The initial parsing of the content set is the key!
- Customers tend to be distracted by the pretty interface
- A subject matter expert is critical for scientific direction
- Users have no idea how to apply the power of the application
- A different presentation of data is intriguing – regardless of novelty
- Text mining is not a magic button – application of intellectual capital will always be essential



Where's my PickAx? (Tools for Text Mining)

Publicly available tools

Arrowsmith (University of Illinois, Chicago)

- http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/start.cgi
- Helps discover indirect relationships:
A→B and B→C implies that A→C

ClusterMed (Vivisimo)

- <http://www.clustermed.org/>
- Clustering of search results from Medline, the web, news sources and even eBay



Innovation Through Information

Publicly available tools

MedMiner (NCI)

- <http://discover.nci.nih.gov/textmining/main.jsp>
- Filters and organizes literature based on a gene, gene-gene or gene-drug query of PubMed and GeneCards

PubFinder (German Cancer Research Ctr)

- <http://www.glycosciences.de/tools/PubFinder/>
- Improve the retrieval rate of scientific abstracts from PubMed relevant for a specific scientific topic based on an analysis of core documents



Innovation Through Information

Publicly available tools

Textpresso (California Institute of Technology)

- <http://www.textpresso.org/>
- Searches the *c. elegans* literature for particular facts with an ontology of 14,500 entries

XplorMed (Ottawa Health Research Institute)

- <http://www.ogic.ca/projects/xplormed/>
- Gives you the main associations between the words in groups of abstracts
- Might try in cases where you don't know what you're expecting to find