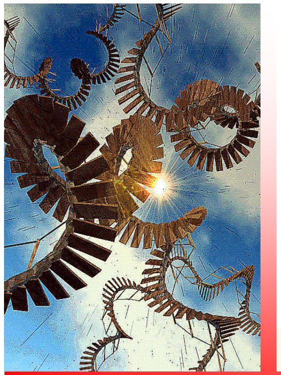
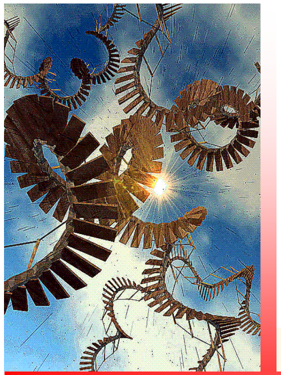
Information Retrieval and Text Mining Opportunities in Bioinformatics

Dr. N. JEYAKUMAR, M.Sc., Ph.D.,
Dept. of Bioinformatics
Bharathiar University
Coimbatore - 641046



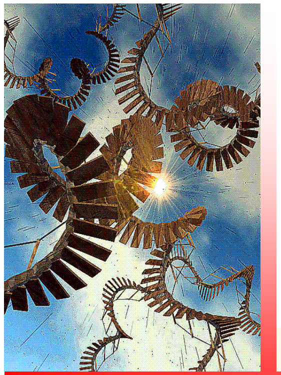
Purpose & Targeted Audience

- **Purpose**: broad overview of information retrieval and text mining and its application to bioinformatics
  - An attempt at a definition
  - A brief history of use in Bioinformatics literature
  - Outline of key applications, papers & emerging areas
- **Audience**: people with good background
  - Biology
  - Computer science
  - Neither of the two disciplines



Outline

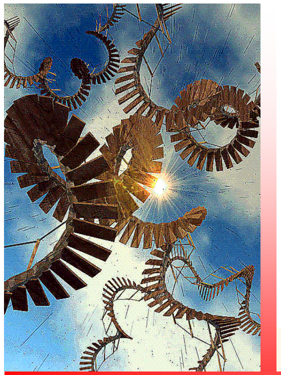
- Introduction to IR and TM
- Biomedical Literature Resources
- Two basic tasks – Bio-Entity and Entity-Relation Identification
- Knowledge Discovery with text
- Text data integration
- Outlook



Information Reterival and Text Mining:

Biology – why?

- Rich sources of text in the form of
  - Abstracts
  - Full text
  - Patients’ records
  - Annotations in data sources (sequence and structure databases)
- For example abstract database Medline contains
  - 18 million records (abstracts)
  - ~50,000 records are added every month
- Novel biomedical information are hidden across the text
  - such as protein interactions, protein localization, gene annotations, molecular pathways etc



Information Extraction

Sample PubMed Record

TI - Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein

AB - Originally identified as a ‘mitotic cyclin’, cyclin A exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an S-phase-promoting factor (SPF) as well as a candidate proto-oncogene.

Other recent studies have identified human cyclin D1 (PRAD1) as a putative G1 cyclin and candidate proto-oncogene.

However, the specific enzymatic activities and, hence, the precise biochemical mechanisms through which cyclins function to govern cell cycle progression remain unresolved.

In the present study we have investigated the coordinate interactions between these two potentially oncogenic cyclins, cyclin-dependent protein kinase subunits (cdks) and the Rb tumor-suppressor protein.

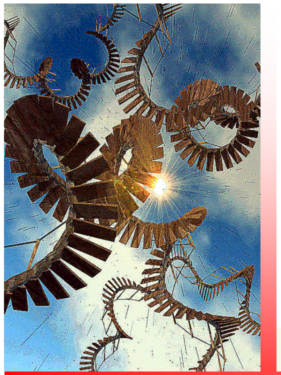
The distribution of cyclin D isoforms was modulated by serum factors in primary fetal rat lung epithelial cells.

Moreover, cyclin D1 was found to be phosphorylated on tyrosine residues in vivo and, like cyclin A, was readily phosphorylated by pp60c-src in vitro.

In synchronized human osteosarcoma cells, cyclin D1 is induced in early G1 and becomes associated with p9Ckshs1, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing’s sarcoma cells demonstrated that cyclin D1 is associated with both p34cdc2 and p33cdk2, and that cyclin D1 immune complexes exhibit appreciable histone H1 kinase activity.

Immobilized, recombinant cyclins A and D1 were found to associate with cellular proteins in complexes that contain the p105Rb protein.



Information Extraction

Sample PubMed Record with Named Entites

TI - Two potentially oncogenic cyclins, **cyclin A** and **cyclin D1**, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the **Rb** protein

AB - Originally identified as a ‘mitotic cyclin’, **cyclin A** exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an **S-phase-promoting factor** (**SPF**) as well as a candidate proto-oncogene.

Other recent studies have identified human **cyclin D1** (**PRAD1**) as a putative G1 cyclin and candidate proto-oncogene.

However, the specific enzymatic activities and, hence, the precise biochemical mechanisms through which cyclins function to govern cell cycle progression remain unresolved.

In the present study we have investigated the coordinate interactions between these two potentially oncogenic cyclins, cyclin-dependent protein kinase subunits (cdks) and the **Rb** tumor-suppressor protein.

The distribution of **cyclin D** isoforms was modulated by serum factors in primary fetal rat lung epithelial cells.

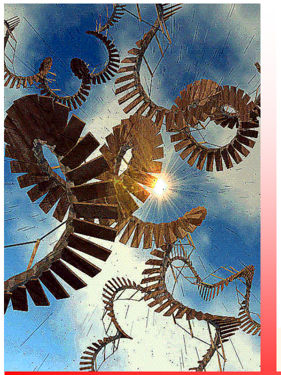
Moreover, **cyclin D1** was found to be phosphorylated on tyrosine residues in vivo and, like **cyclin A**, was readily phosphorylated by **pp60c-src** in vitro.

In synchronized human osteosarcoma cells, **cyclin D1** is induced in early G1 and becomes associated with **p9Ckshs1**, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing’s sarcoma cells demonstrated that **cyclin D1** is associated with both **p34cdc2** and **p33cdk2**, and that **cyclin D1** immune complexes exhibit appreciable histone H1 kinase activity.

Immobilized, recombinant cyclins A and D1 were found to associate with cellular proteins in complexes that contain the **p105Rb** protein.

.



Text Mining:

Genetic Basics

- **Gene/Protein – Associate/interact – Gene/protein => pathway**
  (concept) (conceptual relation) (concept) =>(Biological process)
  (e.g) STAT3 interact BCL-X => apoptosis (cell death)

- **Gene/protein – symptom– disease**
  (concept) (function) (concept)

  (e.g.) p53 tumor suppressor cancer
  TNFRSF1B Insulin resistance diabetes

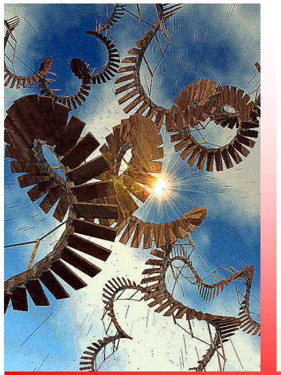
So, the main goal of any text mining/information extraction system in biomedical domain is identify the bio-entitles and their relationship

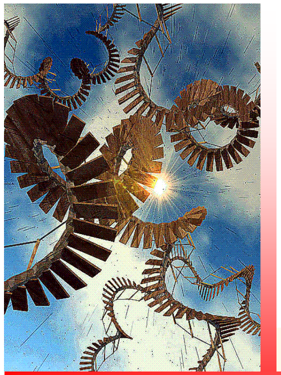
Part I: Information Retrieval and Text Mining

Information Retrieval:

Introduction and overview

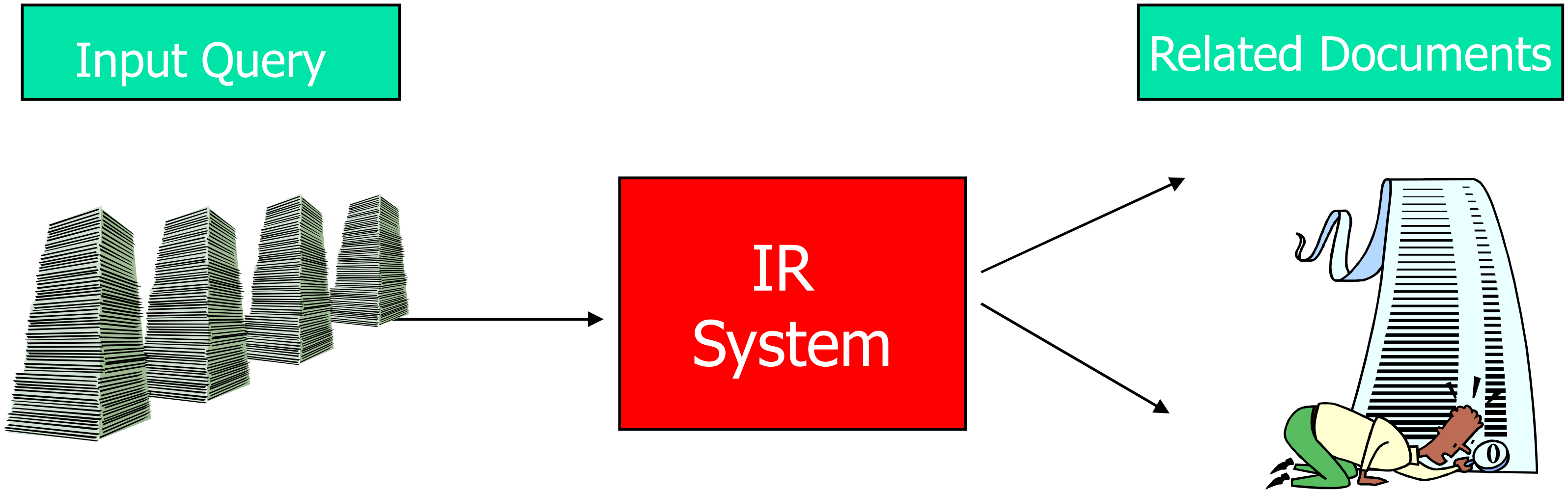
- Information retrieval (IR) is the science of searching for documents, for information within documents and for metadata about documents, as well as that of searching the World Wide Web.
- (e.g.) Google, Google Scholar, PUBMED, PUBMED CENTRAL
- Component Tasks
  - Document indexing
    - Sentence tokenization/word tolenization
    - Steaming
    - Stop word removal
  - Query Types:
    - Boolean queries
    - Bag of words/Vector space model
- Related Tasks
  - Text classification
  - Text Clustering





Information Retrieval:

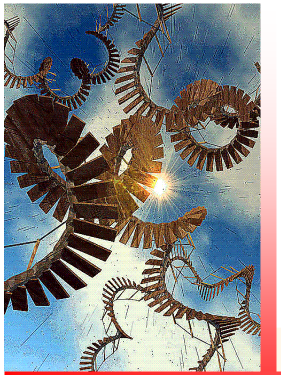
Information Retrieval - Example



Input Query

IR System

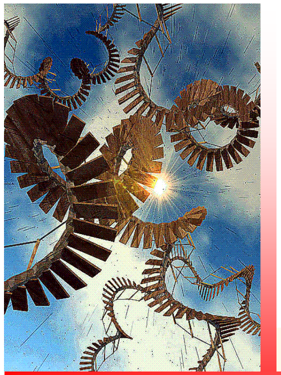
Related Documents



Information Retrieval:

IR Stages of processing – Lexical Analysis

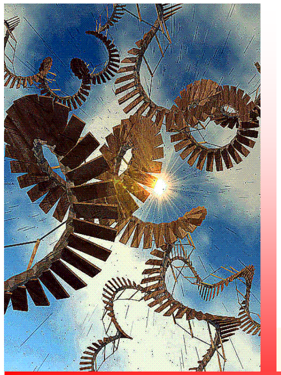
- Sentence tokenization
  - separates text into individual sentences.
- Word tokenization
  - breaks pieces of text into word-sized chunks; in biology this is a difficult task as the definition of what a word is can be quite complex and it is further complicated by heavy use of punctuation (e.g., ERD-1/2, endothelin-1).
- Stemming
  - is a process that determines the stem of a word; a word stem is the main part and excludes elements that used to indicate plurality, tense, case, gender, person, etc.
  - (e.g.) activate is the stem of the words activation, activated, activates, and activating.
  - Porter stemmer – may implementations available in Net
- Stop word removal
  - The most common words that unlikely to help text mining such as prepositions, articles, and pro-nouns
  - (e.g.) "the", "a", "an", with, "you" …
  - many stop word list are available on net



Information Retrieval:

IR stages of processing – Query Types

- **Boolean Queries**
  - Based on combination of terms using Boolean operators
  - Basic Boolean operators: AND, OR, NOT
  - Queries matched against the terms in the inverted index file
  - Fast and easy to implement but retrieves many irreverent documents



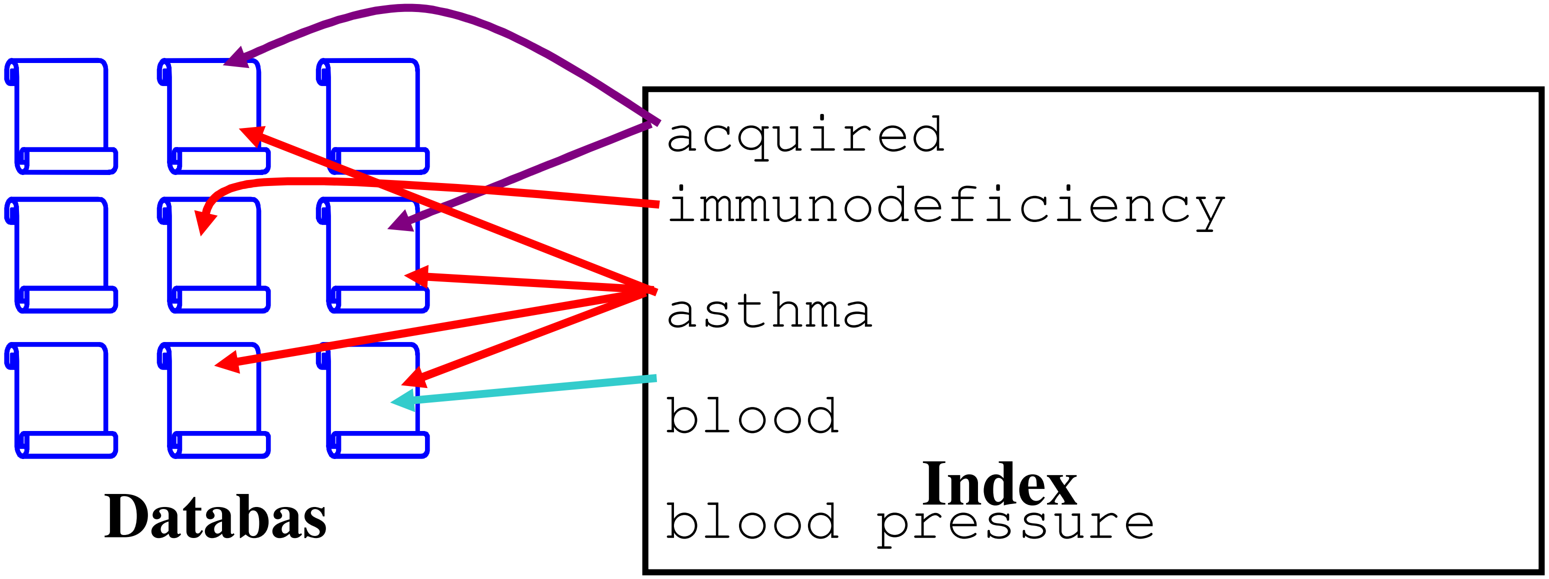
Information Retrieval:

Boolean Queries

DB: Database of documents.

Vocabulary: ***$\{t_1,\ldots,t_M\}$*** (Terms in DB, produced by the tokenization stage)

Index Structure: A term → all the documents containing it.



acquired

immunodeficiency

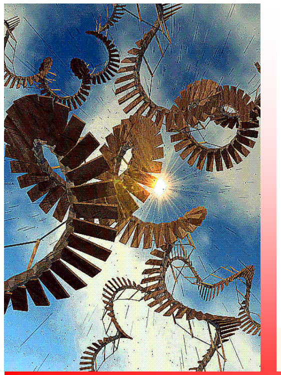
asthma

blood

blood pressure

Index

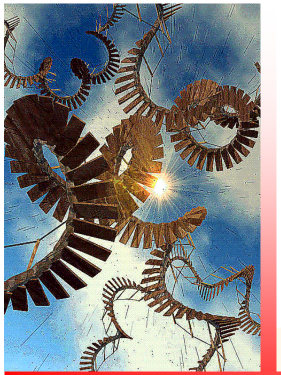
Databas
e



Information Retrieval:

IR stages of processing – Query Types

- **Bag of words/ Vector space model**
  - text document is represented by the words it contains (and their occurrences)
  - (e.g.) “Lord of the rings” {“the”, “Lord”, “rings”, “of”}
  - Highly efficient
  - Makes learning far simpler and easier
  - Order of words is not that important for certain applications
  - Each sentence is represented as vector of word frequencies
  - Relations betwteen the sentences identified by cosine angles



Information Retrieval:

Vector space model

(a)

Documents *a*, *b*, and *x*

A	Gene *BRCA1* and *BRCA2* participate in repairing radiation-induced breaks in DNA ... and other genes.
B	Cancer genes *BRCA1* on chromosome 17 and *BRCA2* on chromosome 13 might disable mechanisms ... gene and drug. But *BRCA1* and *BRCA2* are also implicated ...
X	Gene therapy using novel drug to treat breast and ovarian cancer ... of *BRCA1*.

Vector space representation of *a*, *b*, and *x*

	Gene	**BRCA 1**	**BRCA 2**	**Cancer**	**…**	**drug**
V(*a*)	2	1	1	0	…	0
V(*b*)	2	2	2	1	…	1
V(*x*)	1	1	0	1	…	1

(b)

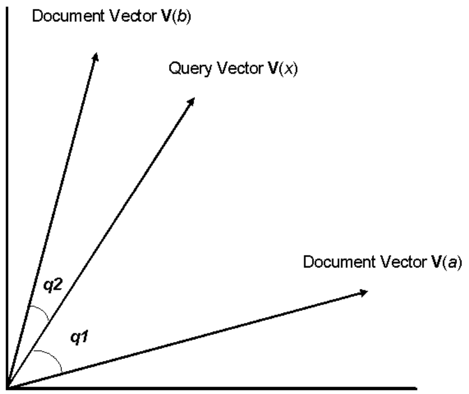
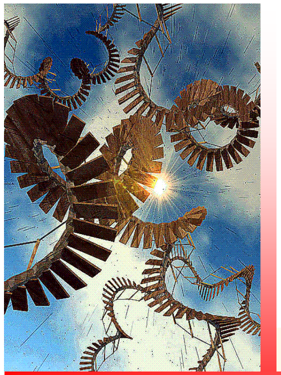


Figure 1: Vector space representation: (a) Coding of texts as weighted vectors—each entry represents the weight of the corresponding term in the vector representing a document, (b) Illustration of the cosine coefficient similarity q_1 and q_2 of query vector V(*x*) with the two vectors V(*a*) and V(*b*) in vector space. Notice that **V**(*x*) is closer to **V**(*b*) than to **V**(*a*).



Information Retrieval: Vector space model

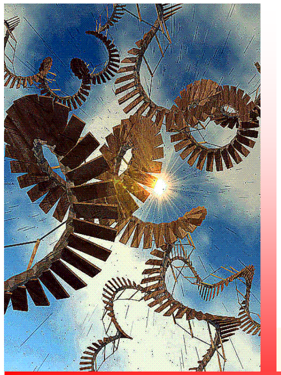
DB: Database of documents.

Vocabulary: **{*$v_1,\ldots,v_M$*}** {Terms in DB}

Document d∈DB: **Vector, *$<w_1^d,\ldots,w_M^d>$, of weights.***

Weighting Principles

- **Document frequency:** Terms occurring in a *few* documents are *more useful* than terms occurring in *many*.
- **Local term frequency:** Terms occurring *frequently* within a document are likely to be *significant* for the document.
- **Document length:** A term occurring the same # of times in a long document and in a short one has *less significance* in the *long* one.
- **Relevance:** Terms occurring in documents judged as *relevant* to a query, are *likely to be significant* (WRT the query).



Information Retrieval: Vector space model

Some Weighting Schemes:

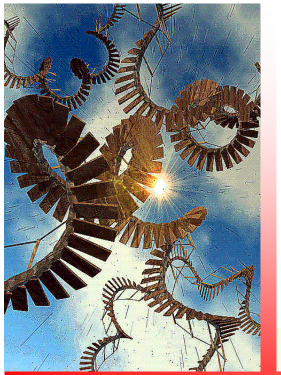
Binary $W_i^d = \begin{cases} 1 & \text{if } t_i \in d \\ 0 & \text{otherwise} \end{cases}$

TF $W_i^d = f_i^d =$ # of times t_i occurs in d.

Consider Local term frequency

TF X IDF
(one version...) $W_i^d = \frac{f_i^d}{f_i}$ ($f_i =$ # of docs containing t_i)

Consider Local term frequency
and Document frequency



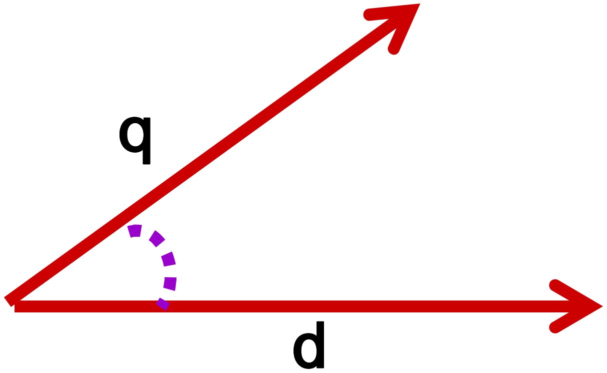
Information Retrieval: Vector space model

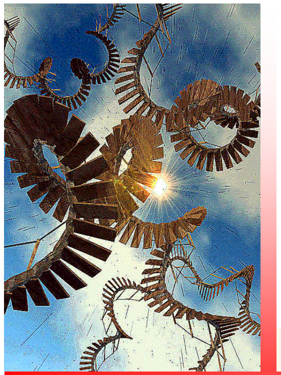
Document $d = \langle w_1^d, \ldots, w_M^d \rangle \in DB$

Query $q = \langle w_1^q, \ldots, w_M^q \rangle$ (q could itself be a document in DB...)

$$\mathrm{Sim}(q, d) = \mathit{cosine}(q, d)$$

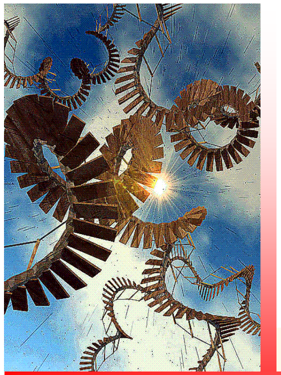
$$= \frac{q \cdot d}{|q||d|}$$





Information Retrieval: IR Evaluation

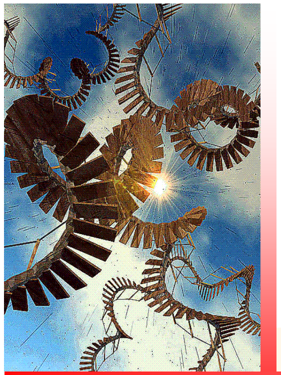
- Precision: fraction of relevant documents retrieved divided by the total returned documents
- Recall: proportion of relevant documents returned divided by the total number of relevant documents
- F-score: the harmonic mean of precision and recall
- Precision-recall curves



Information Retrieval: IR Evaluation

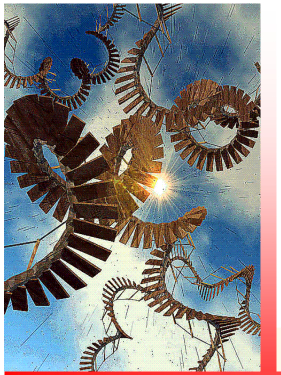
- $\text{precision} = TP / (TP + FP)$
- $\text{recall} = TP / (TP + FN)$

- $\text{F-measure} = 2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$



Text Clustering

- Find which documents have many words in common, and place the documents with the most words in common into the same groups.
- Similarity of documents instead of similarity of sequences, expression profiles or structures
- Cluster documents into topics, for instance: clinical, biochemical and microbiology articles
- A clustering program tries to find the groups in the data.

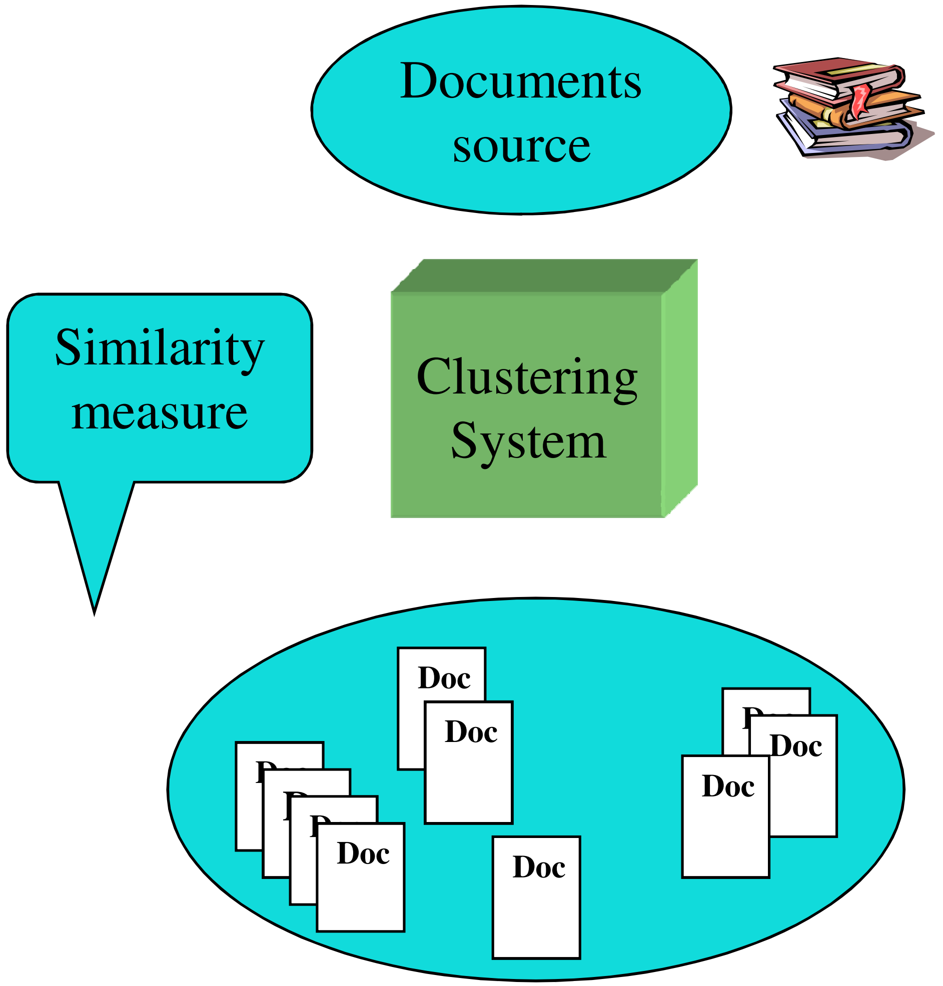


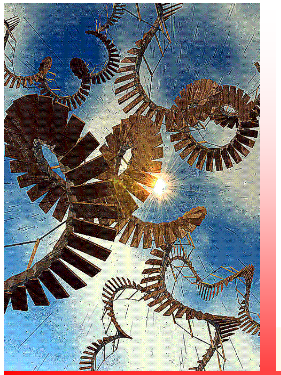
Text Clustering

- Idea
  - Frequent terms carry more information about the “cluster” they might belong to
  - Highly co-related frequent terms probably belong to the same cluster
- $D = \{D_1, \ldots, D_n\}$ – the set of documents
  - D_j *subsetOf* T, the set of all terms
- Then candidate clusters are generated from $F = \{F_1, \ldots , F_k\}$, where each F_i is a set of all frequent terms *which occur together*.

Text Mining:

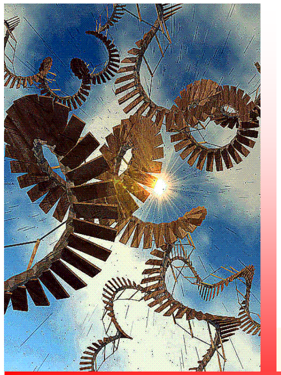
Text Clustering- Example





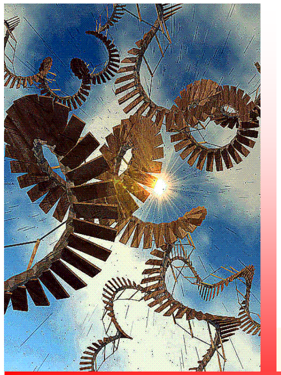
Text Clustering

- Techniques used
  - Partitioning
  - Hierarchical
    - Agglomerative
    - Divisive
  - Grid based
  - Model based



Text Classification

- The problem statement
  - Given a set of documents, each with a *label* called the class label for that document
  - Given, a classifier which ***learns*** from the above data set
  - For a new, unseen document, the classifier should be able to “predict” with a high degree of accuracy the correct class to which the new document belongs

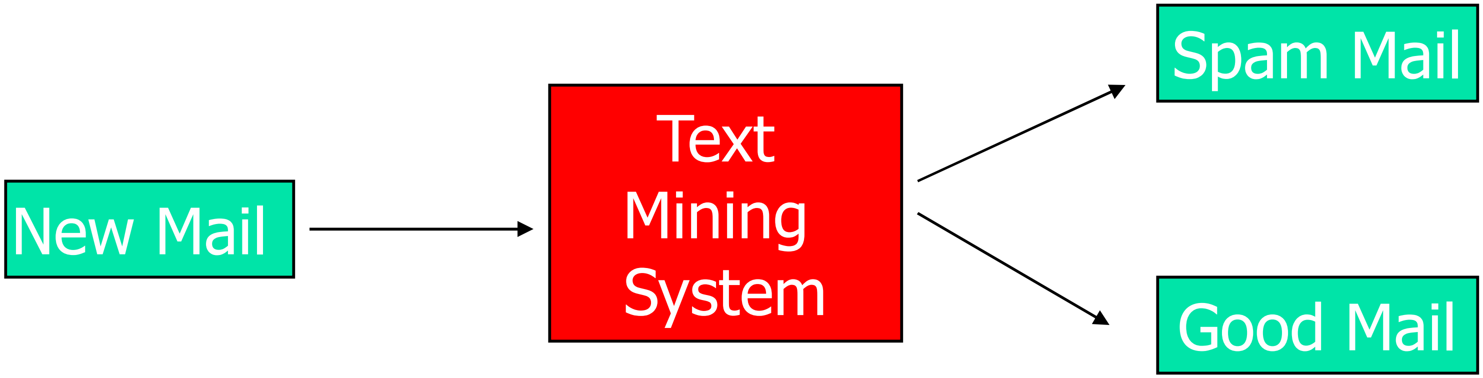


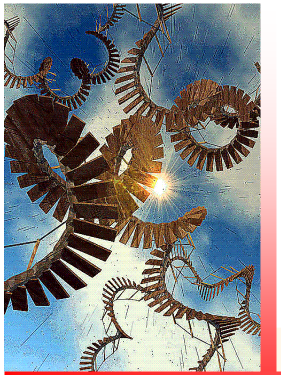
Text Classification

- Common problem in information science.
- Assignment of an electronic document to one or more categories, based on its contents (words).
- Supervised document classification where training examples of document classification are provided and the correct classification
- model is learnt based on one of the following techniques:
  - naive Bayes classifier
  - tf-idf
  - latent semantic indexing
  - support vector machines
  - artificial neural network
  - kNN
  - decision trees, such as ID3
- Classification techniques have been applied to spam filtering

Text Classification - Example

(e.g.) Spam mail filtering

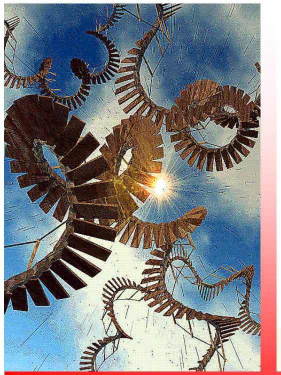




Text Mining:

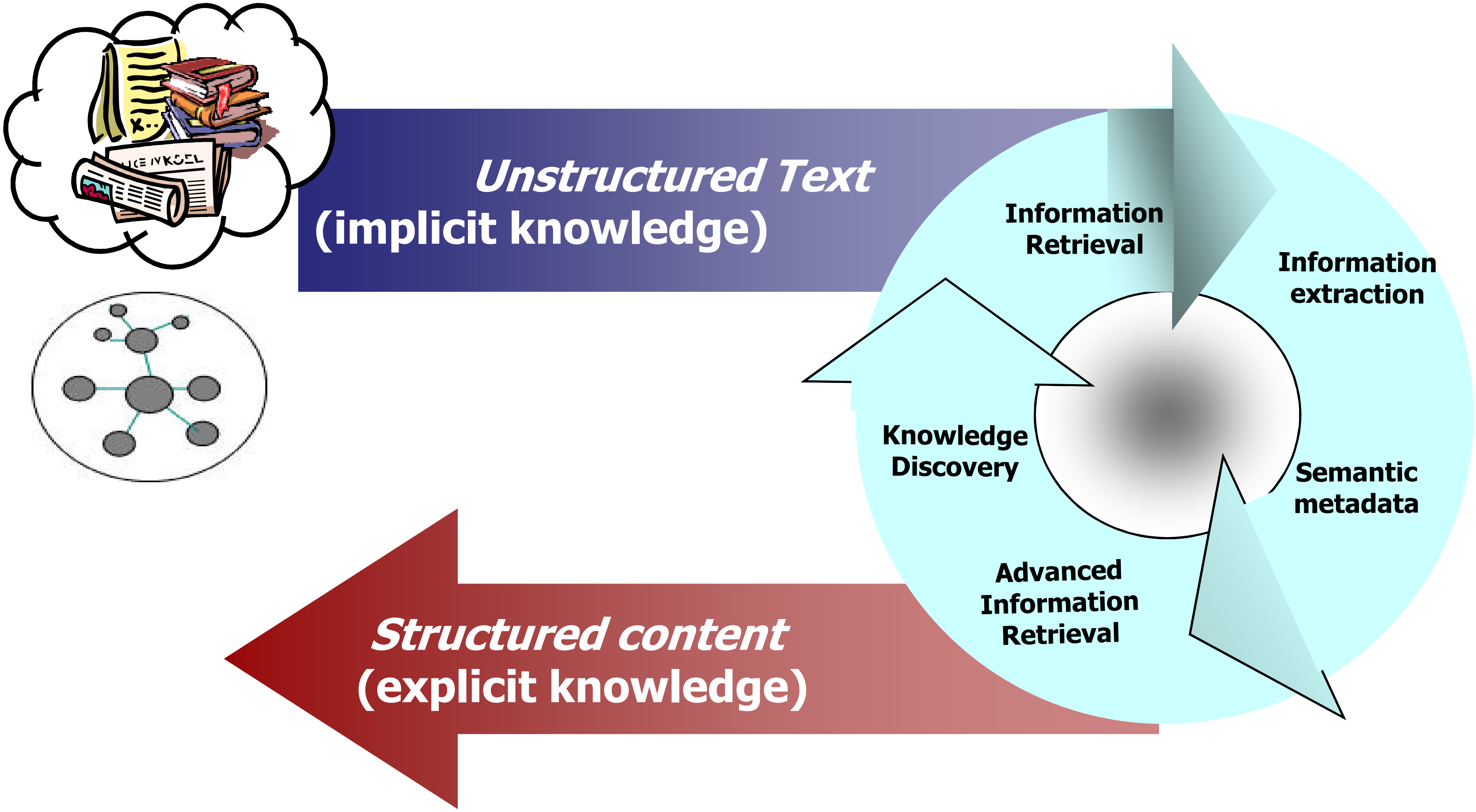
Introduction and overview

- Text mining aims to identify non-trivial, implicit, previously unknown, and potentially useful patterns in text (e.g. classification system, summarization, association rules, hyphothesis etc.)
- Includes more established research areas such as
  - Information Retrieval (IR),
  - Natural Language Processing (NLP),
  - Information Extraction (IE),
  - and traditional Data Mining (DM)
- Related Tasks
  - Text Summarization
  - Question and Answering



IR and Text Mining:

The Big Picture



Unstructured Text
(implicit knowledge)

Information Retrieval

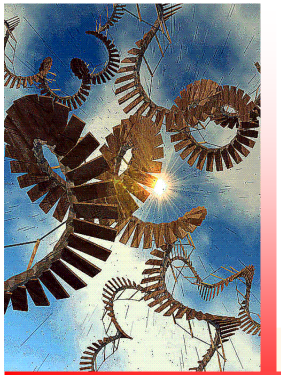
Information extraction

Semantic metadata

Advanced Information Retrieval

Knowledge Discovery

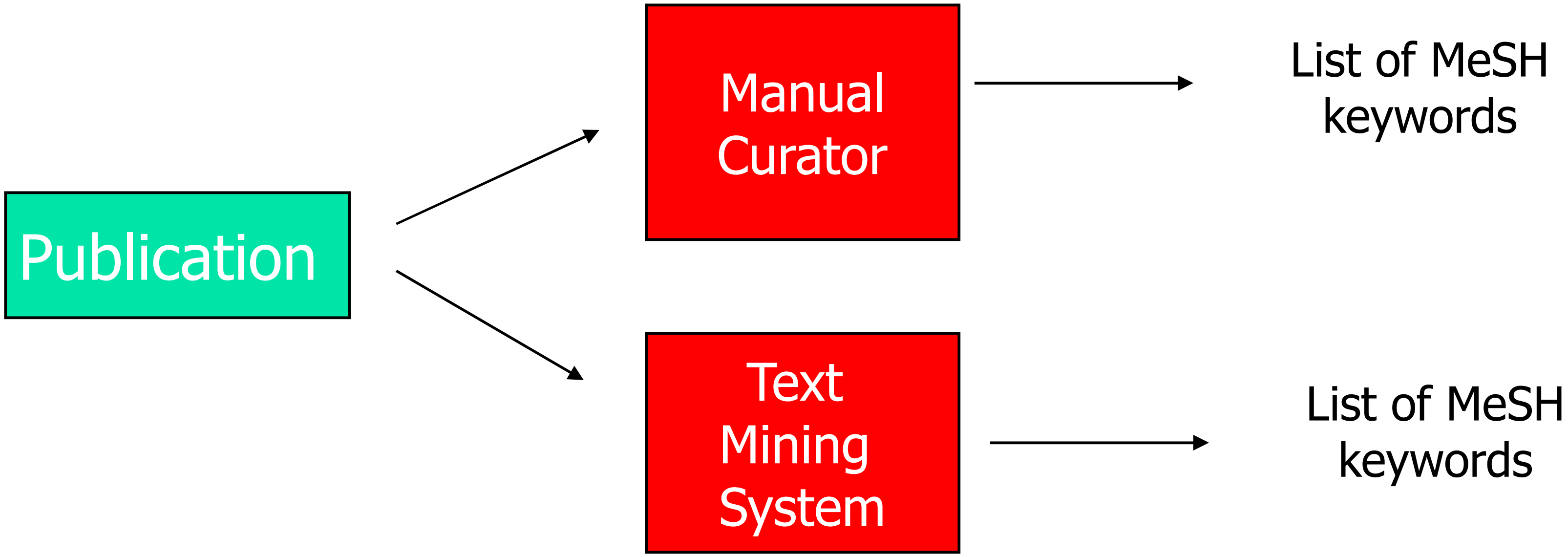
Structured content
(explicit knowledge)



Text Mining:

Text Mining – Simple Example

Automatically curating literature information



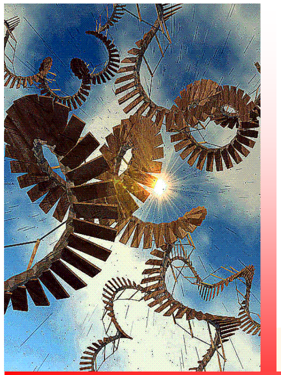
Publication

Manual Curator

Text Mining System

List of MeSH keywords

List of MeSH keywords



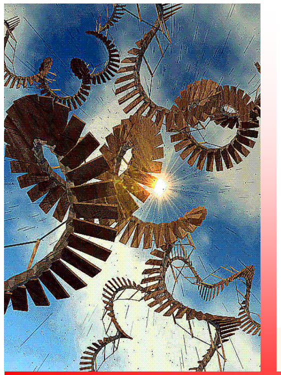
Text Mining:

Pattern or Knowledge Discovery - Example

Hypothesis generation

(e.g.1) Ram and Ravi are friends
(e.g.2) Ram and Rajiv are friends
=> Ravi and Rajiv may be friend or known to each other

(e.g.1) gene A regulate gene B
(e.g.2) gene B induce gene C
=> gene A, B, C are in same pathway



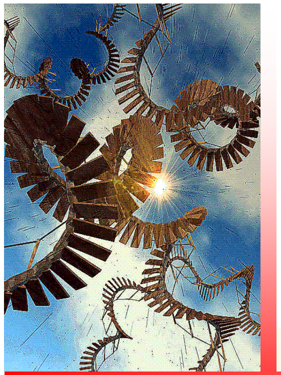
Text Mining:

Related Fields

- Information retrieval aims to identify to identify relevant documents in response to a query (e.g. Google search, PubMeD search etc.)

- Natural language processing, also called computational linguistics attempts to use automated means to process text and deduce its syntactic and semantic structure

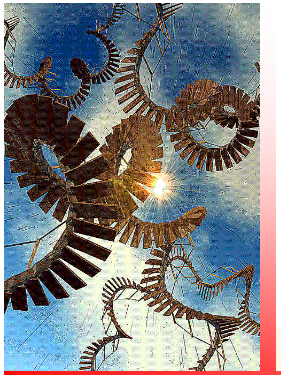
- Information extraction aims to identify automatically specific predefined classes of entities (e.g. protein and gene names), relations (e.g. protein interactions) or known facts (cell localization) in natural language text



Text Mining:

Natural Language processing and Component Tasks

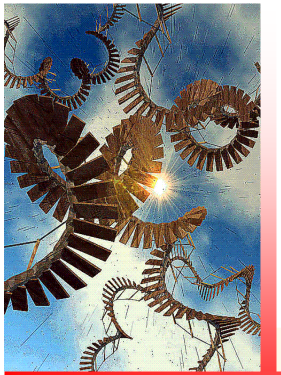
- Syntactic and semantic relation of text
- Gives sentence structure and how word are form the sentence
- (e.g.) noun, verb, adverb, pro-noun, prepositions etc and complete sentence structure
- Component Tasks
  - Part of speech (pos) tagging
  - Shallow parsing
  - Full parsing



Text Mining:

NLP stages of processing

- Part-of-speech tagging
  - involves the assignment of part-of-speech information or labels such as word categories (e.g., adjective, article, noun, proper noun, preposition, verb) and other lexical class markers to individual tokens a text corpus.
  - e.g., John (noun) gave (verb) the (det) ball (noun)
- Shallow parsing
  - refers to a class of techniques concerned with the identification of phrasal chunks (noun, noun phrase, verb, verb phrase) in each sentence of a corpus without assignment of ‘deep’ hierarchical structures (graph).
- Full parsing
  - is concerned with the construction of a complete parse tree (deep hierarchical structures) for a sentence in a corpus



Text Mining:

NLP - POS tagging

- Part of Speech (POS) tagging - involves the assignment of part-of-speech information or labels such as word categories (e.g., adjective, article, noun, proper noun, preposition, verb)

<sentence>

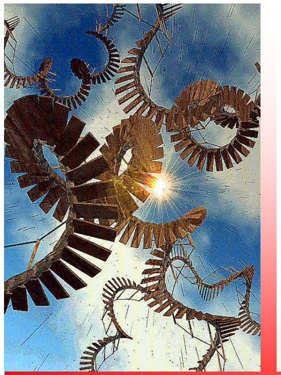
BRCA1 physically associates with p53 and stimulates its transcriptional activity.

</sentence>

<POS Sentence>

BRCA1/NNP physically/RB associates/VBZ with/IN p53/NN and/CC stimulates/VBZ its/PRP$ transcriptional/JJ activity/NN

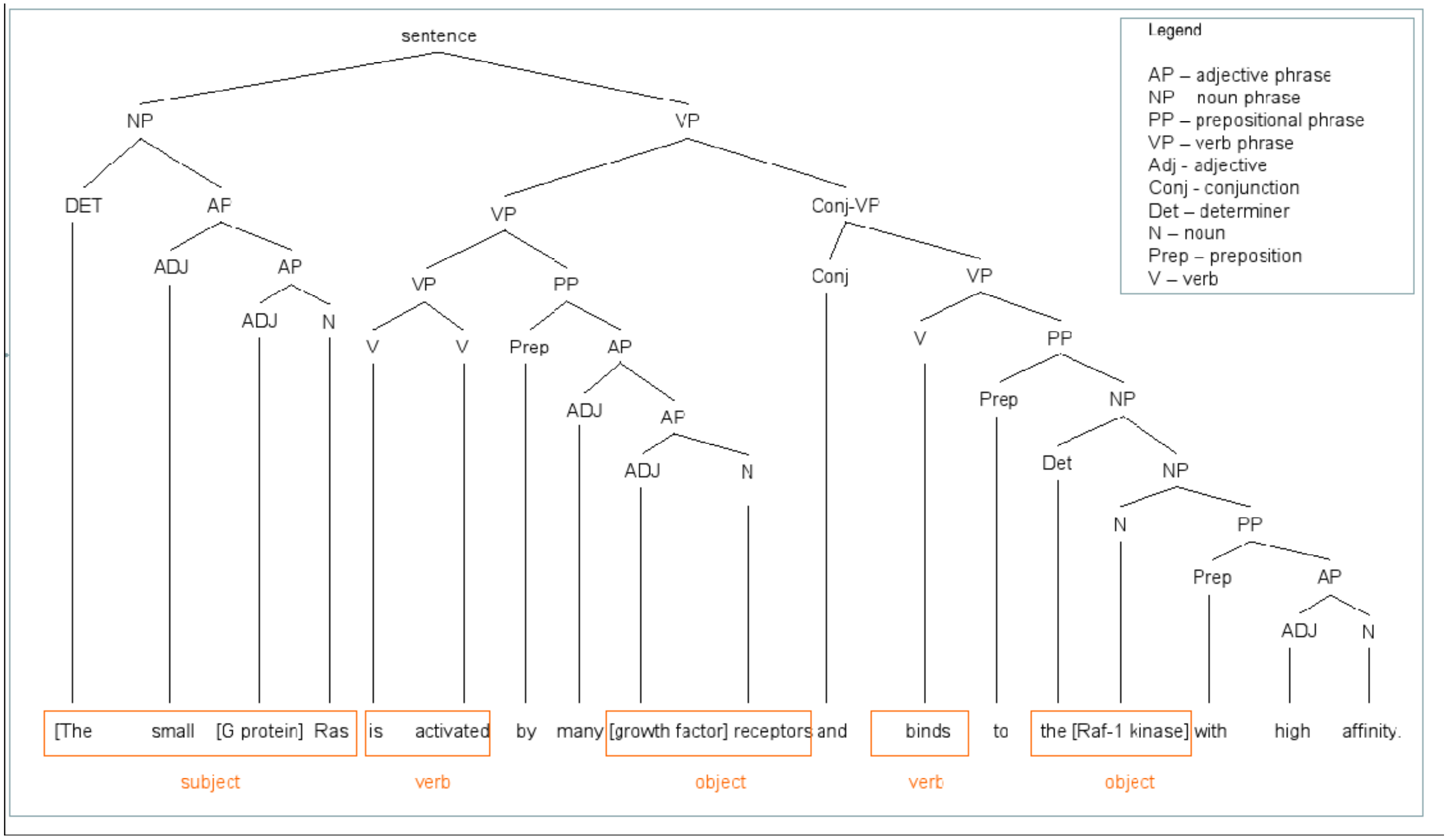
</POS Sentence>



Text Mining:

NLP - Full Parser

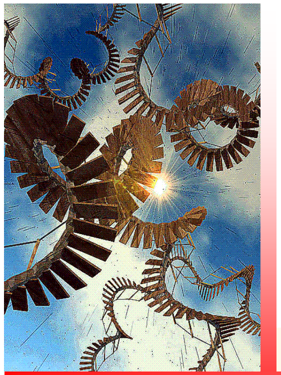
- Full parsing - Complete understanding of sentence structure



Legend

AP – adjective phrase
NP   noun phrase
PP – prepositional phrase
VP – verb phrase
Adj - adjective
Conj - conjunction
Det – determiner
N – noun
Prep – preposition
V – verb

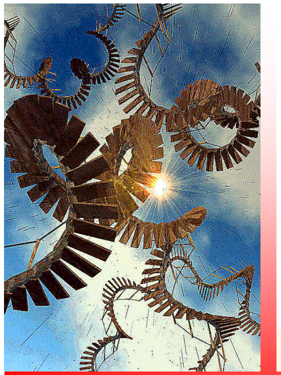
[The small [G protein] Ras is activated by many [growth factor] receptors and binds to the [Raf-1 kinase] with high affinity.



Text Mining:

Information Extraction and Component Tasks

- Find concepts
- Pro-noun concepts
- Concept relations, scenario relations
  - (e.g.) genes, protein names, relations, cross relations
- Component Tasks
  - Named entity recognition (NER)
  - Co-reference resolution
  - Template element extraction
  - Template relation extraction
  - Scenario template extraction



Text Mining:

IE – Named Entity Tagging

- Named entity tagging in Text. (identifying concepts such as protein/gene names etc.)

<sentence>

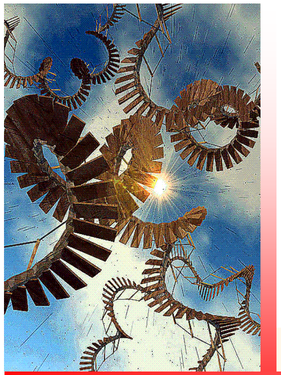
It has been show that genistein induces phosphorylation of ATM on serine 1981 and phosphorylation of histone H2AX on serine 13 in B cells.

</sentence>

<Tagged Sentence>

It has been shown that **<smallmol>**genistein**</smallmol>** induces phosphorylation of **<protein>**ATM**</protein>** on **<enzyme>**serine 1981**</enzyme>** and phosphorylation of **<protein>**histone H2AX**<protein>** on **<enzyme>**serine 13**</enzyme>** in **<celltype>**B cells**</celltype>**.

</Tagged Sentence>



Text Mining:

IE – Template Relation Extraction

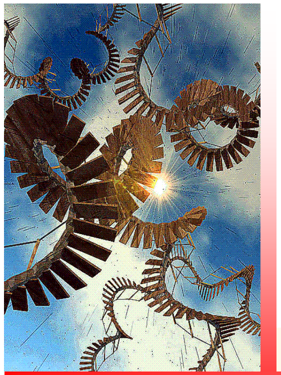
- Template relation extraction (identifying relation between the concepts such as protein-protein interactions etc.)

```
<sentence>
It has been show that genistein induces phosphorylation of ATM on serine
1981 and phosphorylation of histone H2AX on serine 13 in B cells.
</sentence>

<protein id=p1>ATM</protein>
<protein id=p2>histone H2AX</protein>

<smallmol id=s1>genistein</smallmol>

<relation id=r1 type='induce' node1=s1 node2=p1>
<relation id=r2 type='induce' node1=s1 node2=p2>
```



Text Mining:

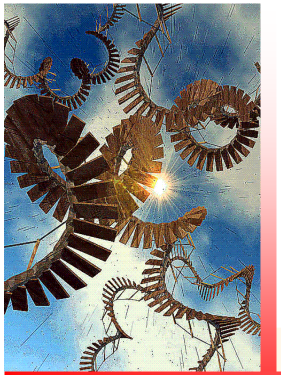
IE – Methodology

- Rule based approaches
- Context-free grammar approaches
- Full parsing approaches
- Sublanguage driven IE
- Ontology-driven IE

Text Mining:

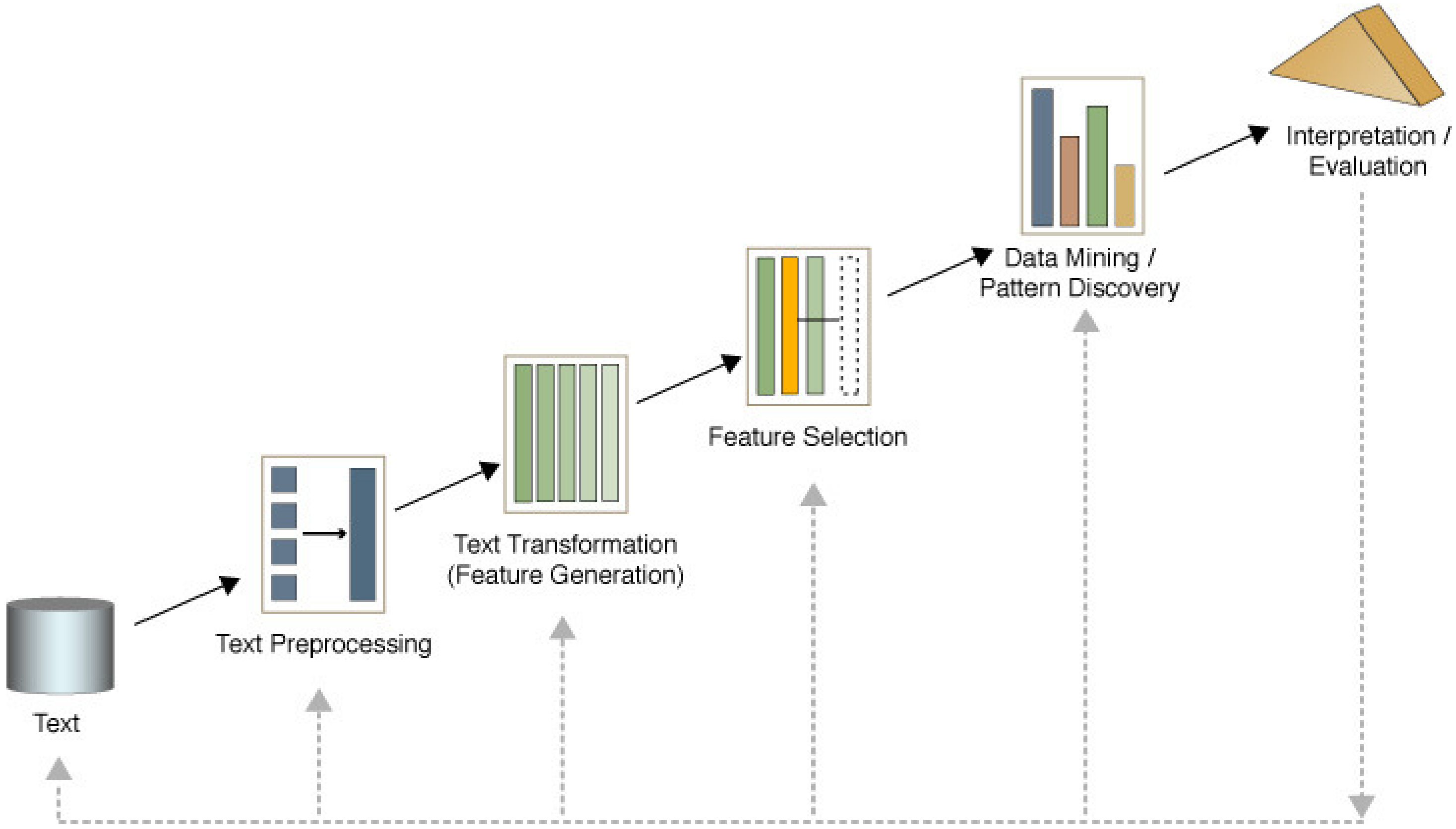
Text Mining from Related Fields

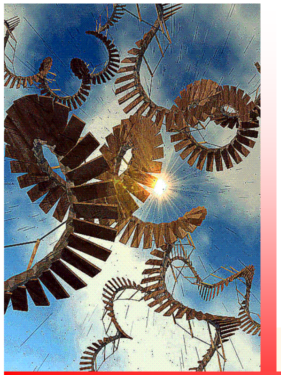
- Data collection (gathering documents related to specific problem) (IR)
- Data pre-processing (tokenization, normalization, parsing, stemming, stop word removal etc.) (NLP/IR)
- Finding entities (named objects like proteins, genes etc.) (IE)
- Finding facts (relationships among entities) (IE)
- Mining (more complex relationship among entities and concept to concept relationships) (TM)
    - (e.g.1) gene A regulate gene B
    - (e.g.2) gene B induce gene C
    - => gene A, B, C are in same pathway



Text Mining:

Text mining stages of processing

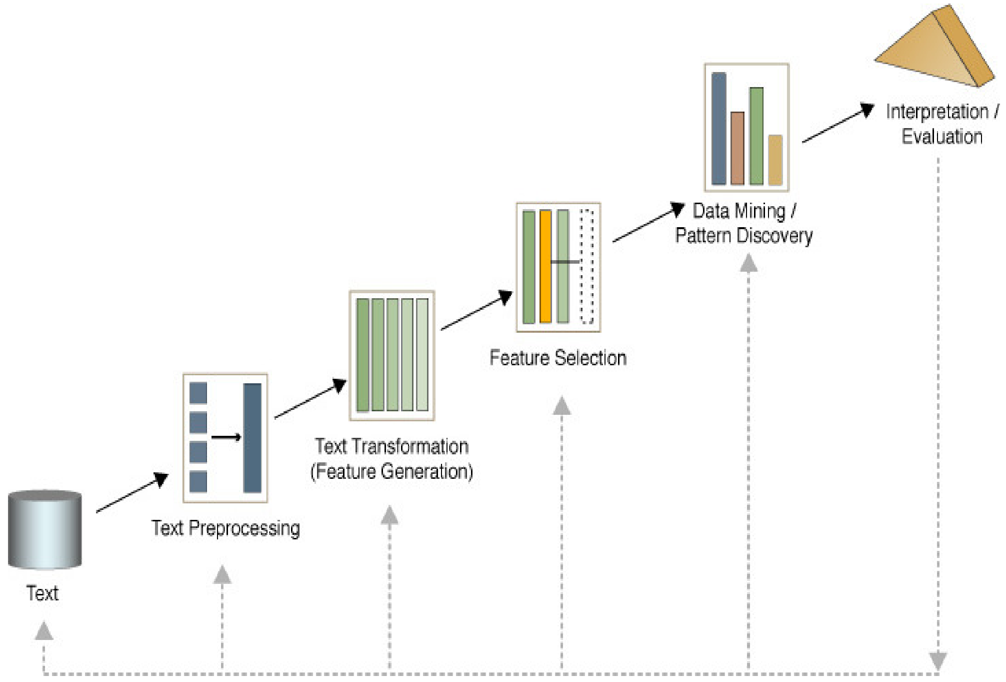




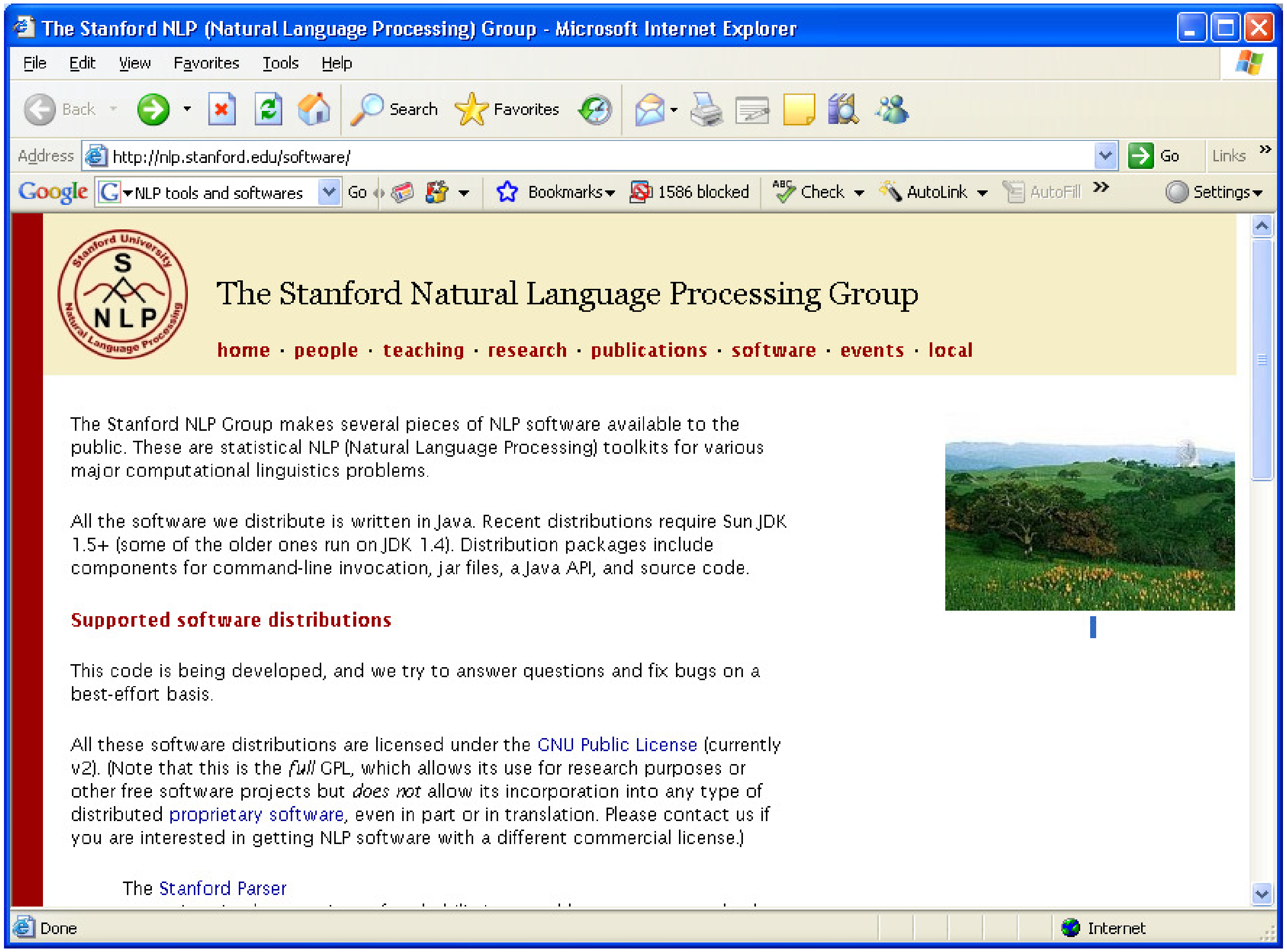
Text Mining:

Text mining stages of processing

- Text preprocessing
  - Stemming, stop word removal
  - Syntactic/Semantic text analysis
- Features Generation
  - Bag of words
- Features Selection
  - Simple counting
  - Statistics
- Text/Data Mining
  - Classification- Supervised learning
  - Clustering- Unsupervised learning
- Post-processing
  - Analyzing results
  - Evaluation


Text
Text Preprocessing
Text Transformation (Feature Generation)
Feature Selection
Data Mining / Pattern Discovery
Interpretation / Evaluation

Text Mining:
Resources Example


The Stanford NLP (Natural Language Processing) Group - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://nlp.stanford.edu/software/
Go Links
Google NLP tools and softwares Go Bookmarks 1586 blocked Check AutoLink AutoFill Settings
Done
Internet

The Stanford Natural Language Processing Group

home · people · teaching · research · publications · software · events · local

The Stanford NLP Group makes several pieces of NLP software available to the public. These are statistical NLP (Natural Language Processing) toolkits for various major computational linguistics problems.

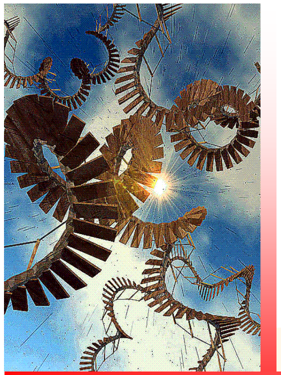
All the software we distribute is written in Java. Recent distributions require Sun JDK 1.5+ (some of the older ones run on JDK 1.4). Distribution packages include components for command-line invocation, jar files, a Java API, and source code.

Supported software distributions

This code is being developed, and we try to answer questions and fix bugs on a best-effort basis.

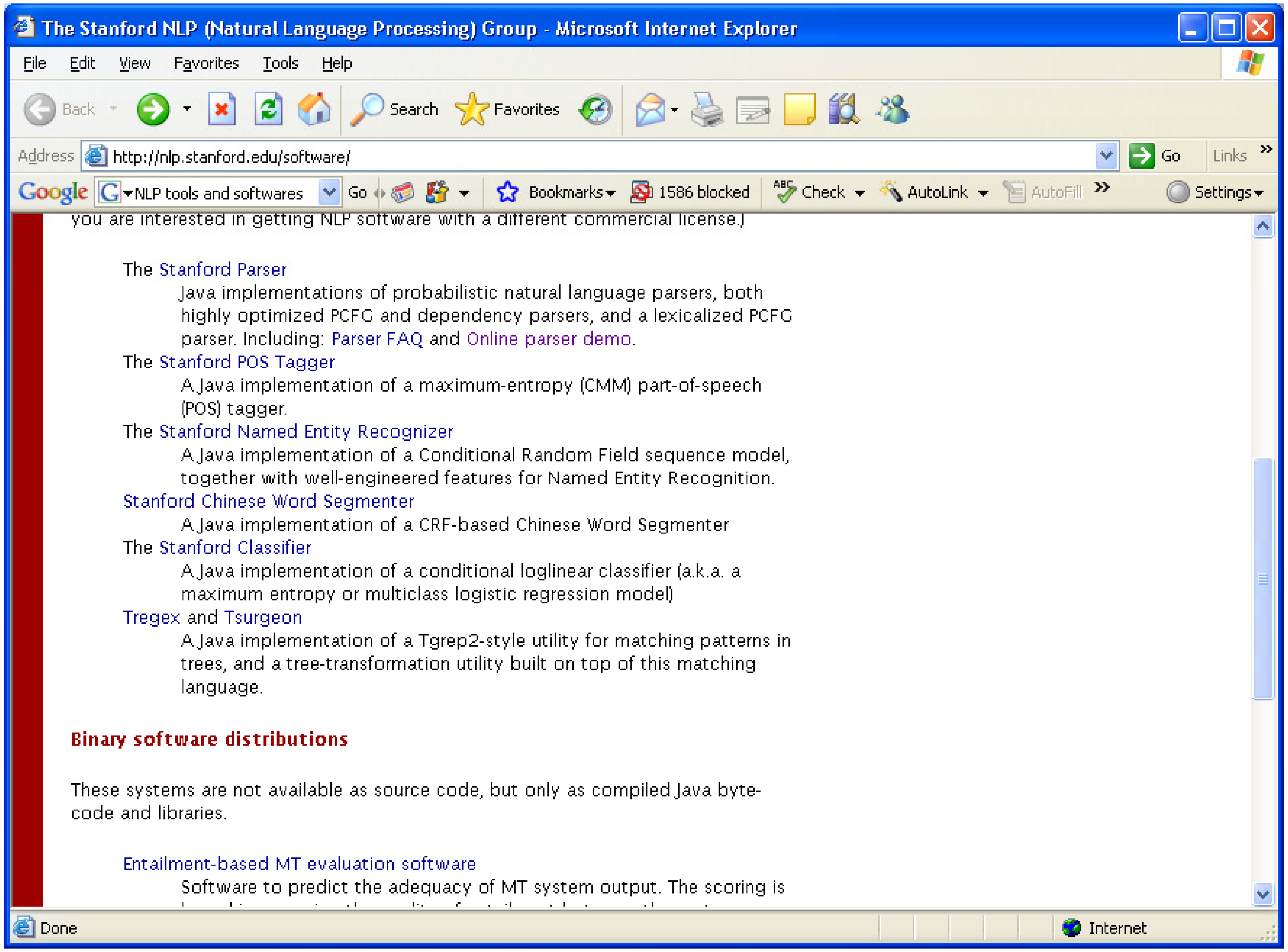
All these software distributions are licensed under the GNU Public License (currently v2). (Note that this is the *full* GPL, which allows its use for research purposes or other free software projects but *does not* allow its incorporation into any type of distributed proprietary software, even in part or in translation. Please contact us if you are interested in getting NLP software with a different commercial license.)

The Stanford Parser



Text Mining:

Resources Example



The Stanford NLP (Natural Language Processing) Group - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://nlp.stanford.edu/software/ Go Links
Google NLP tools and softwares Go Bookmarks 1586 blocked Check AutoLink AutoFill Settings
Done
Internet

you are interested in getting NLP software with a different commercial license.)

The Stanford Parser
Java implementations of probabilistic natural language parsers, both highly optimized PCFG and dependency parsers, and a lexicalized PCFG parser. Including: Parser FAQ and Online parser demo.

The Stanford POS Tagger
A Java implementation of a maximum-entropy (CMM) part-of-speech (POS) tagger.

The Stanford Named Entity Recognizer
A Java implementation of a Conditional Random Field sequence model, together with well-engineered features for Named Entity Recognition.

Stanford Chinese Word Segmenter
A Java implementation of a CRF-based Chinese Word Segmenter

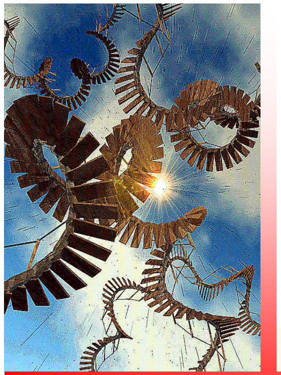
The Stanford Classifier
A Java implementation of a conditional loglinear classifier (a.k.a. a maximum entropy or multiclass logistic regression model)

Tregex and Tsurgeon
A Java implementation of a Tgrep2-style utility for matching patterns in trees, and a tree-transformation utility built on top of this matching language.

Binary software distributions

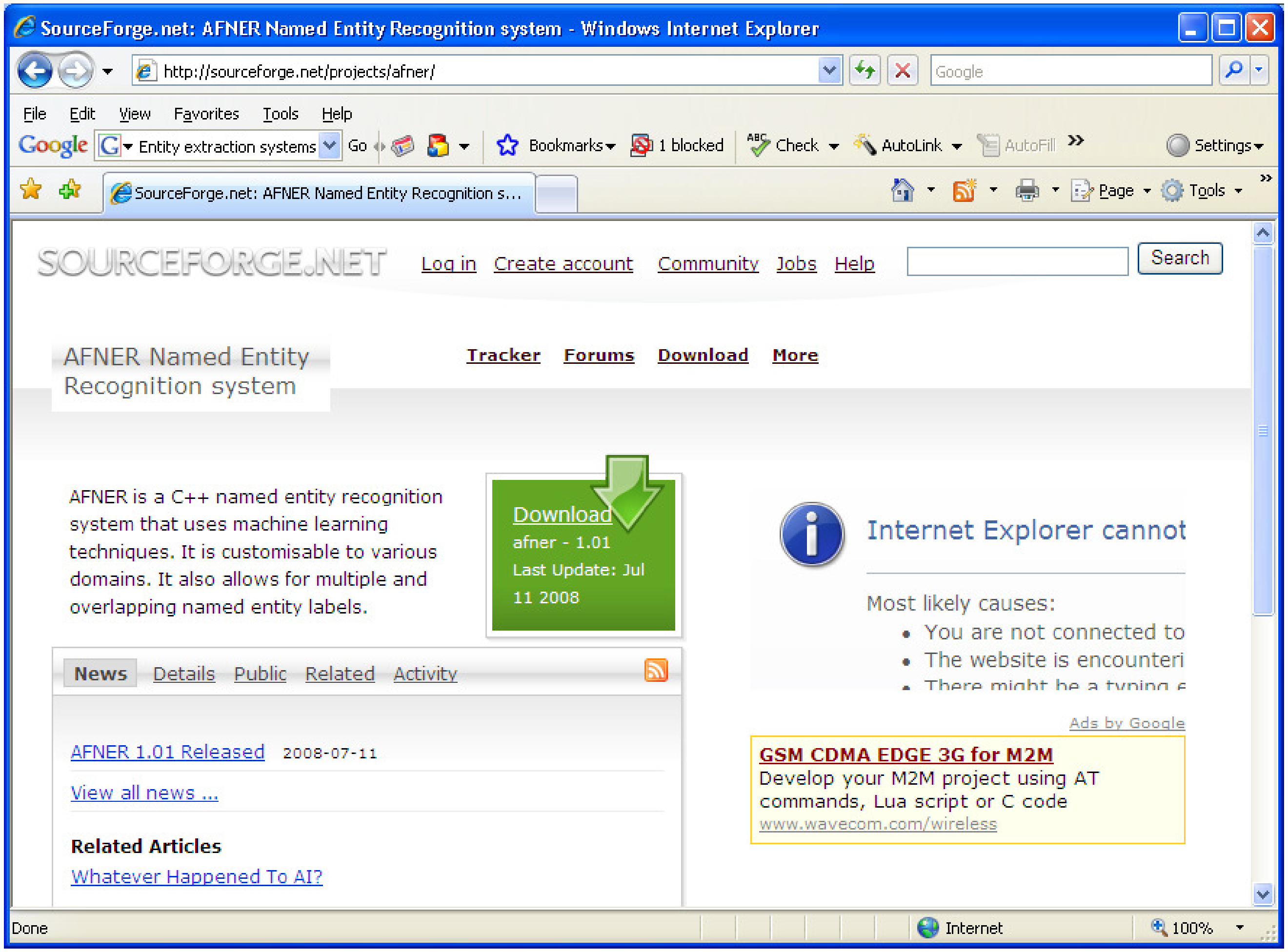
These systems are not available as source code, but only as compiled Java byte-code and libraries.

Entailment-based MT evaluation software
Software to predict the adequacy of MT system output. The scoring is



Text Mining:

Resources Example



SourceForge.net: AFNER Named Entity Recognition system - Windows Internet Explorer

http://sourceforge.net/projects/afner/

File Edit View Favorites Tools Help
Google Entity extraction systems Go Bookmarks 1 blocked Check AutoLink AutoFill Settings
SourceForge.net: AFNER Named Entity Recognition s...
Page Tools
SOURCEFORGE.NET
Log in Create account Community Jobs Help Search
Tracker Forums Download More

AFNER Named Entity Recognition system

AFNER is a C++ named entity recognition system that uses machine learning techniques. It is customisable to various domains. It also allows for multiple and overlapping named entity labels.

Download
afner - 1.01
Last Update: Jul 11 2008

AFNER 1.01 Released 2008-07-11

View all news ...

Related Articles

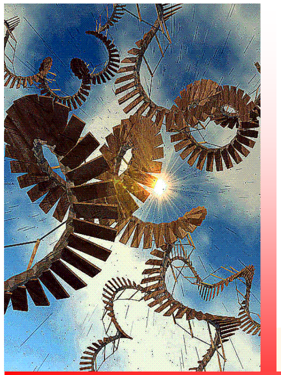
Whatever Happened To AI?

Internet Explorer cannot

Most likely causes:

- You are not connected to
- The website is encounteri
- There might be a typing e

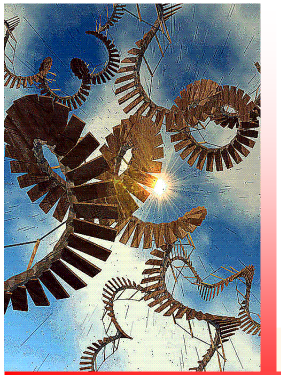
Part II: **Text Mining and Biomedical Literature**



Text Mining:

Biology – why?

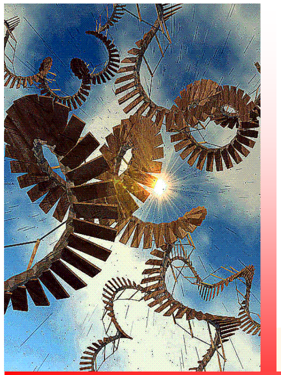
- Rich sources of text in the form of
  - Abstracts
  - Full text
  - Patients' records
  - Annotations in data sources (sequence and structure databases)
- For example abstract database Medline contains
  - 18 million records (abstracts)
  - ~50,000 records are added every month
- Novel biomedical information are hidden across the text
  - such as protein interactions, protein localization, gene annotations, molecular pathways etc



Text Mining:

Why Text About Biology is Special

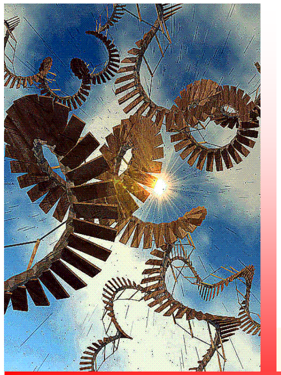
- Large number of Entities/concepts (gene, proteins etc)
- Evolving field, no wild followed standards for terminology ->Rapid change and inconsistency
- Ambiguity (many proteins and genes have same name)
- Synonymy (many proteins and genes have many names)
- Abbreviations (large use of abbrevations in text)



Text Mining:

What are concepts/relations of interest

- Genes (T-Gene)
- Proteins (P53)
- Compounds
- Biological Functions (lipid metabolism)
- Biological Process (cell death, apoptosis)
- Pathways (cell metabolism, Urea Cycle)
- Dieses (Cancer, Alzheimer's, etc.)



Text Mining:

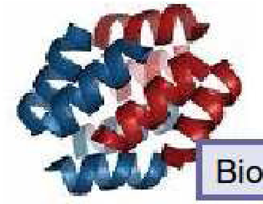
Curation of Biological Literature

- Classical Method: Manual Curation
  - Trained human experts reads scientific literature and extracts information of interest
  - Manual time consuming and labor intensive process
  - Accurate through human inference and background knowledge
  - (E.g.) MeSH Uniprot, GOA, SGD, MGI etc.
- Text Mining assisted Curation
  - Retrieval of relevant literature from literature repositories
  - Textual evidence and entity detection
  - Revision and editing of manual records
  - E.g. TextPresso, Rodriguez-Penagos et al (gene regulation), Grover el at (PPI), Chang et al (Pathways), Ongenaert et al (methylation)

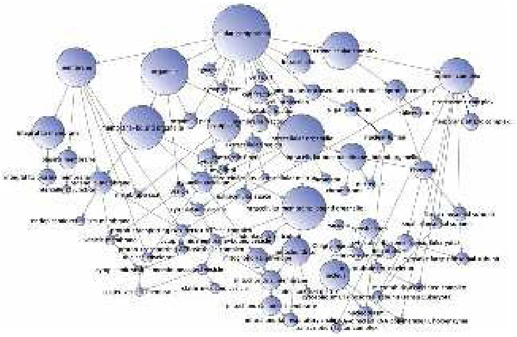
Text Mining:

Curation of Literature in Biology – Pictorial summary

Scientific Literature

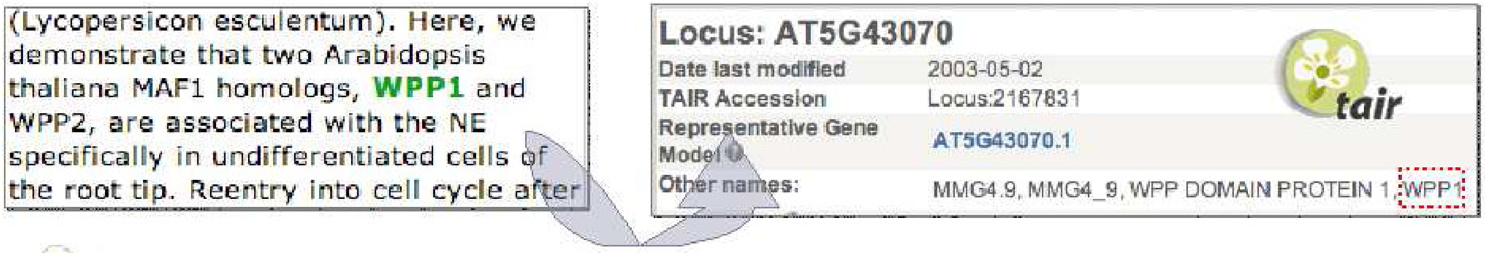


Bio-entities



Controlled vocabularies

Database curator



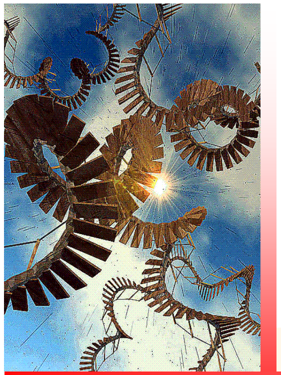
(Lycopersicon esculentum). Here, we demonstrate that two Arabidopsis thaliana MAF1 homologs, **WPP1** and WPP2, are associated with the NE specifically in undifferentiated cells of the root tip. Reentry into cell cycle after

Locus: AT5G43070

Date last modified	2003-05-02
TAIR Accession	Locus:2167831
Representative Gene Model	AT5G43070.1
Other names:	MMG4.9, MMG4_9, WPP DOMAIN PROTEIN 1, WPP1

tair FlyBase MGI GDB SGD swissprot

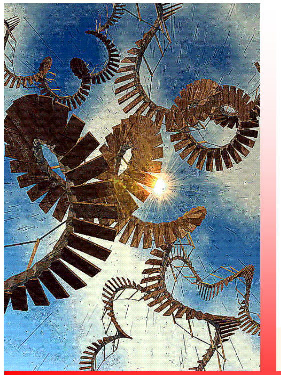
MaizeGDB Maize Genetics and Genomics Database RGD EcoCyc GO MINT



Text Mining:

Current Literature Repositories

- e-Books: NCBI Bookshelf
- Citation of Biomedical Research Articles + Abstract: PubMed (http://www,ncbi.nlm.nih.gov/pubmed)
- Full text research articles:
  - PubMed Central (PMC)
  - Highwire Press
  - BioMed Central
- Google Scholar



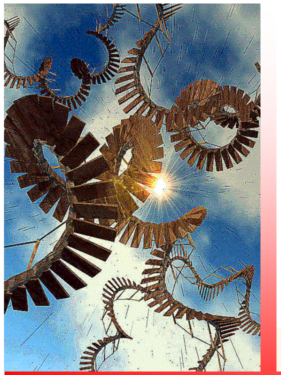
Text Mining:

PUBMED

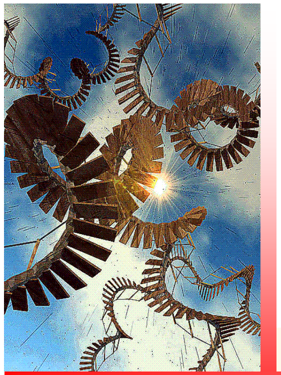
- Overview
  - Developed by NCBI
  - Citation entries of scientific articles of all biomedical sciences
  - Each entry is characterized by a unique identifier, the PubMed identifier: PMID
  - Often links to the full text articles are displayed
- Statistics
  - No. of Citations 16 million
  - No. of Indexed Journals approx. 5000
  - No. of English Articles 12 million
  - No. of Articles with Abstracts 7,000,000

Text Mining:

PUBMED

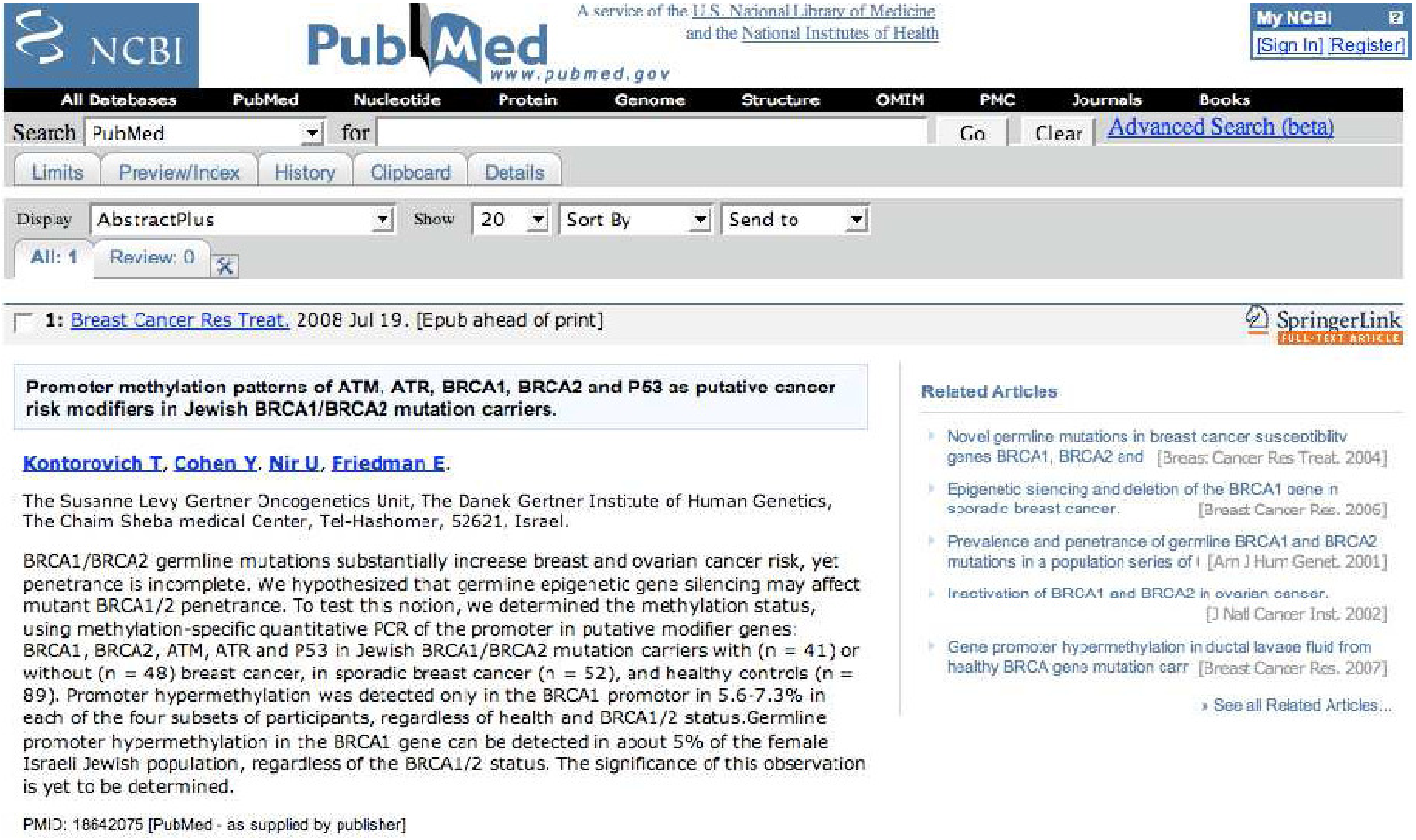


- Approximately 1 million entries refer to gene descriptions
- Author, journal and title information of the publication
- Some records with gene symbols and molecular sequence databank numbers
- Indexed with Medical Subject Headings (MeSH)
- Accessed online through a text-based search query system called Entrez
- Offers additional programming utilities, the Entrez Programming Utilities (eUtils)
- Majority of (apprx 80%) current biomedical text mining is based on PubMed



Text Mining:

PUBMED – web page



1: Breast Cancer Res Treat. 2008 Jul 19. [Epub ahead of print]

Promoter methylation patterns of ATM, ATR, BRCA1, BRCA2 and P53 as putative cancer risk modifiers in Jewish BRCA1/BRCA2 mutation carriers.

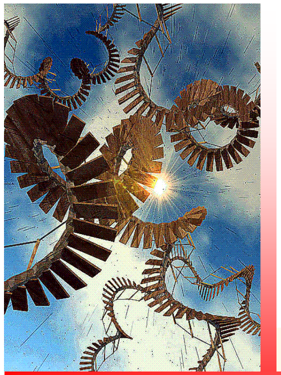
Kontorovich T, Cohen Y, Nir U, Friedman E.

The Susanne Levy Gertner Oncogenetics Unit, The Danek Gertner Institute of Human Genetics, The Chaim Sheba medical Center, Tel-Hashomer, 52621, Israel.

BRCA1/BRCA2 germline mutations substantially increase breast and ovarian cancer risk, yet penetrance is incomplete. We hypothesized that germline epigenetic gene silencing may affect mutant BRCA1/2 penetrance. To test this notion, we determined the methylation status, using methylation-specific quantitative PCR of the promoter in putative modifier genes: BRCA1, BRCA2, ATM, ATR and P53 in Jewish BRCA1/BRCA2 mutation carriers with (n = 41) or without (n = 48) breast cancer, in sporadic breast cancer (n = 52), and healthy controls (n = 89). Promoter hypermethylation was detected only in the BRCA1 promotor in 5.6-7.3% in each of the four subsets of participants, regardless of health and BRCA1/2 status.Germline promoter hypermethylation in the BRCA1 gene can be detected in about 5% of the female Israeli Jewish population, regardless of the BRCA1/2 status. The significance of this observation is yet to be determined.

PMID: 18642075 [PubMed - as supplied by publisher]

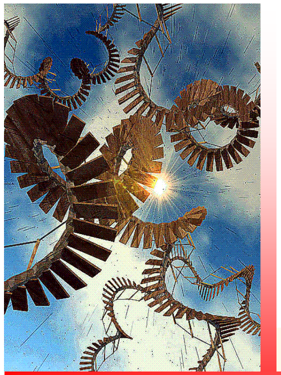
- Novel germline mutations in breast cancer susceptibility genes BRCA1, BRCA2 and [Breast Cancer Res Treat. 2004]
- Epigenetic silencing and deletion of the BRCA1 gene in sporadic breast cancer. [Breast Cancer Res. 2006]
- Prevalence and penetrance of germline BRCA1 and BRCA2 mutations in a population series of ([Am J Hum Genet. 2001]
- Inactivation of BRCA1 and BRCA2 in ovarian cancer. [J Natl Cancer Inst. 2002]
- Gene promoter hypermethylation in ductal lavage fluid from healthy BRCA gene mutation carr [Breast Cancer Res. 2007]



Text Mining:

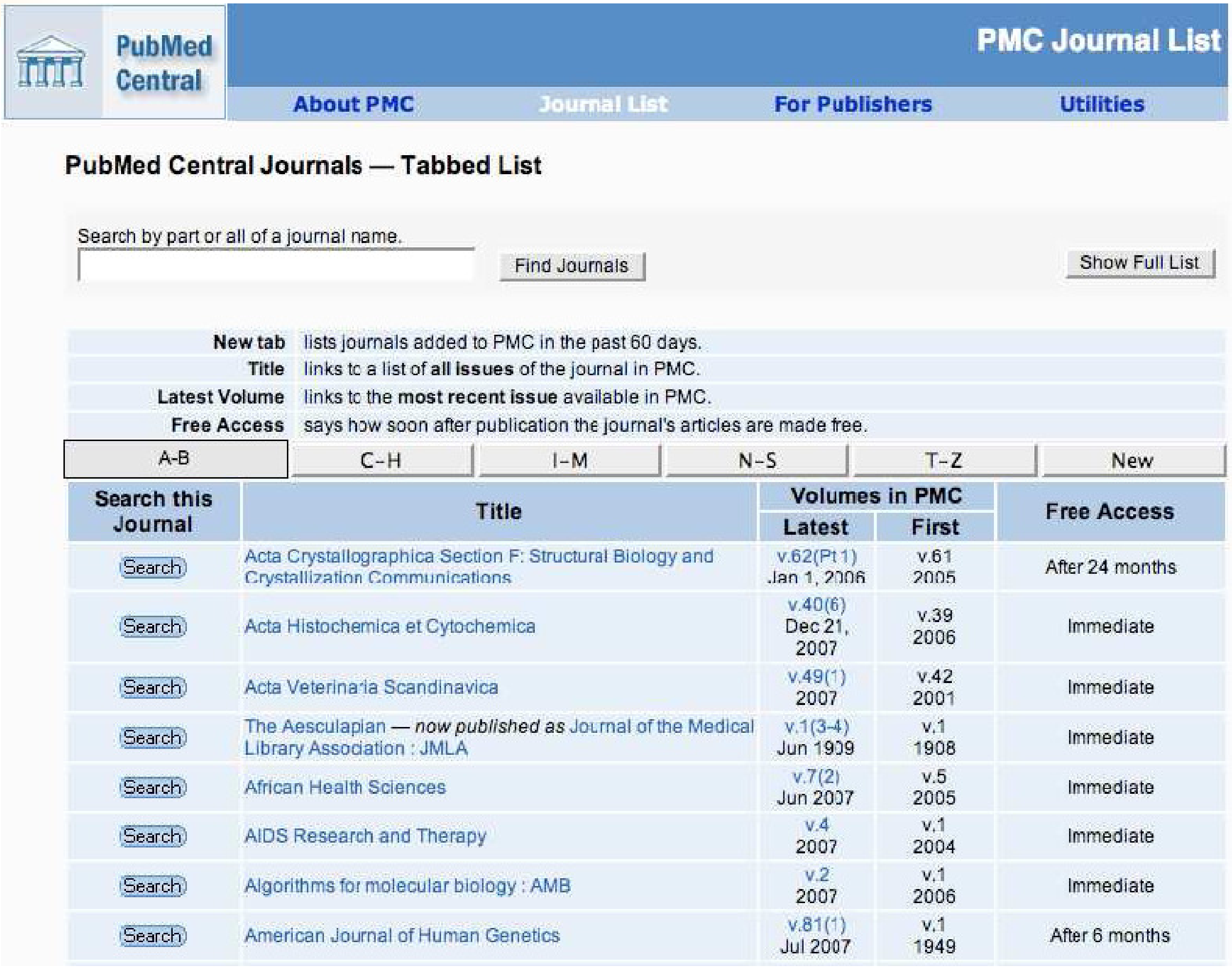
PUBMED Central

- Digital archive of full text life science journals
- Articles have a unique PMCID
- Allows Boolean query search
- Offers free full text articles
- Journal Publishing XML DTD, but also other widely used DTD in life science



Text Mining:

PUBMED Central – web page

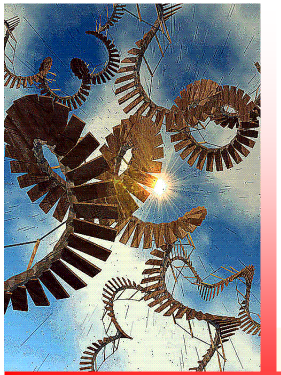


PubMed Central Journals — Tabbed List

Search by part or all of a journal name.

New tab	lists journals added to PMC in the past 60 days.
Title	links to a list of **all issues** of the journal in PMC.
Latest Volume	links to the **most recent issue** available in PMC.
Free Access	says how soon after publication the journal's articles are made free.

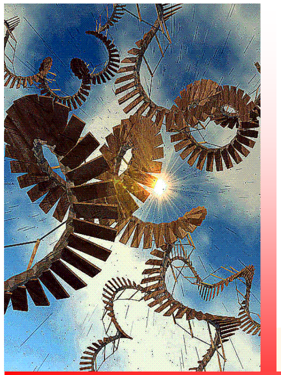
Search this Journal	Title	Volumes in PMC		Free Access
		Latest	First	
Search	Acta Crystallographica Section F: Structural Biology and Crystallization Communications	v.62(Pt 1) Jan 1, 2006	v.61 2005	After 24 months
Search	Acta Histochemica et Cytochemica	v.40(6) Dec 21, 2007	v.39 2006	Immediate
Search	Acta Veterinaria Scandinavica	v.49(1) 2007	v.42 2001	Immediate
Search	The Aesculapian — *now published as* Journal of the Medical Library Association : JMLA	v.1(3-4) Jun 1909	v.1 1908	Immediate
Search	African Health Sciences	v.7(2) Jun 2007	v.5 2005	Immediate
Search	AIDS Research and Therapy	v.4 2007	v.1 2004	Immediate
Search	Algorithms for molecular biology : AMB	v.2 2007	v.1 2006	Immediate
Search	American Journal of Human Genetics	v.81(1) Jul 2007	v.1 1949	After 6 months



Text Mining:

NCBI Book self

- Collection of biomedical text books
- Allows boolean query searches
- Offers free full text articles
- Direct searching the books or from PubMed abstract



Text Mining:

Google Scholar

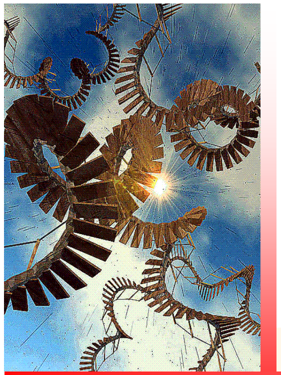
- Google Scholar is a freely accessible Web search engine that indexes the full text of scholarly literature across an array of publishing formats and disciplines. Released in beta in November 2004

- Serves as one full-text biomedical resource for text mining

Text Mining:

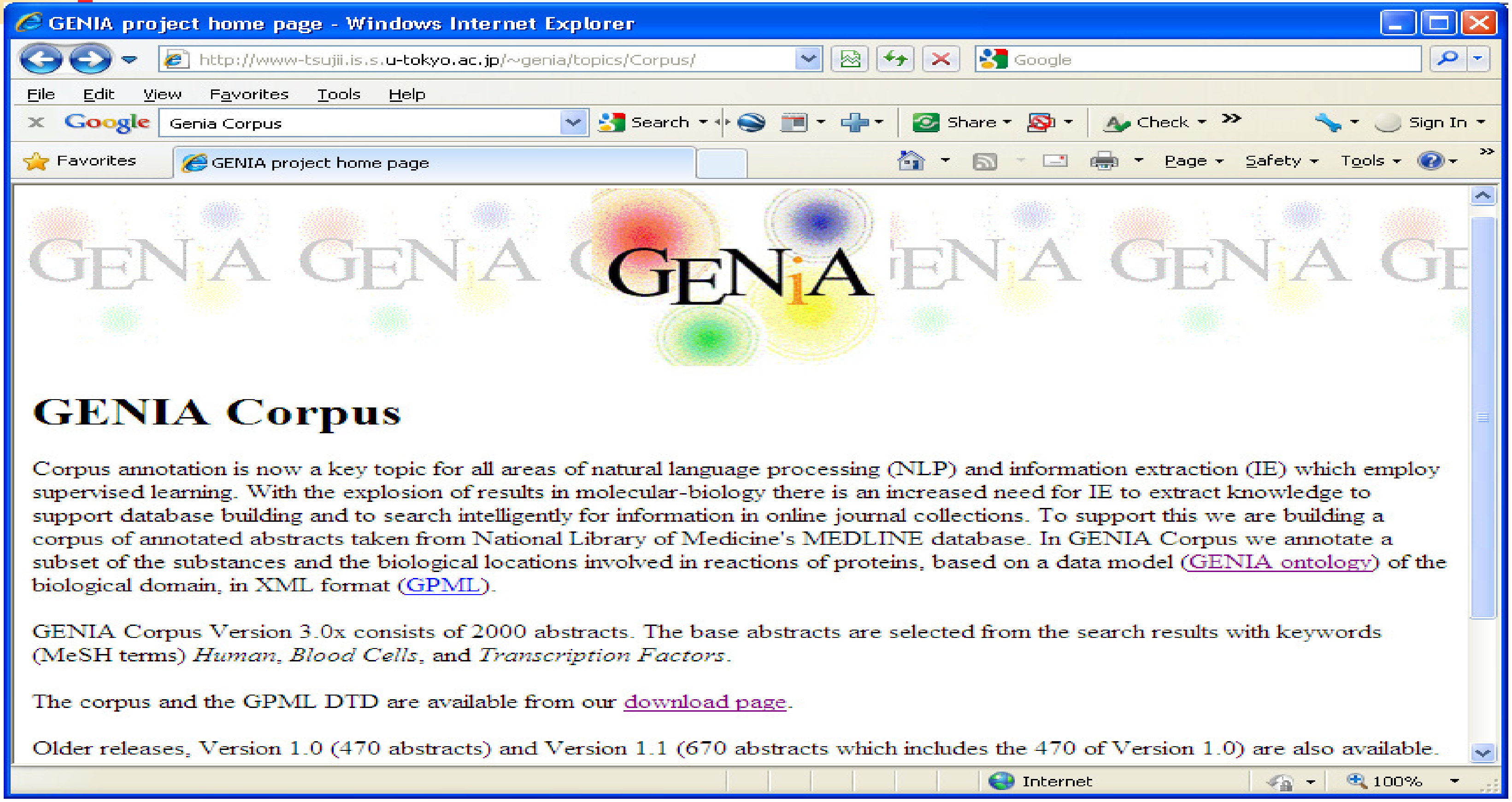
Other Biomedical Corpus

- BioCreative corpus
- GENIA corpus
- Yapex corpus



Text Mining:

GENIA Corpus


GENIA project home page - Windows Internet Explorer
http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/
Google
File Edit View Favorites Tools Help
Google Genia Corpus
Search
Share
Check
Sign In
Favorites
GENIA project home page
Page
Safety
Tools
GENIA
Internet
100%

GENIA Corpus

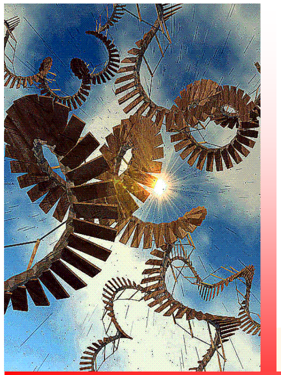
Corpus annotation is now a key topic for all areas of natural language processing (NLP) and information extraction (IE) which employ supervised learning. With the explosion of results in molecular-biology there is an increased need for IE to extract knowledge to support database building and to search intelligently for information in online journal collections. To support this we are building a corpus of annotated abstracts taken from National Library of Medicine's MEDLINE database. In GENIA Corpus we annotate a subset of the substances and the biological locations involved in reactions of proteins, based on a data model (GENIA ontology) of the biological domain, in XML format (GPML).

GENIA Corpus Version 3.0x consists of 2000 abstracts. The base abstracts are selected from the search results with keywords (MeSH terms) *Human*, *Blood Cells*, and *Transcription Factors*.

The corpus and the GPML DTD are available from our download page.

Older releases, Version 1.0 (470 abstracts) and Version 1.1 (670 abstracts which includes the 470 of Version 1.0) are also available.

http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/



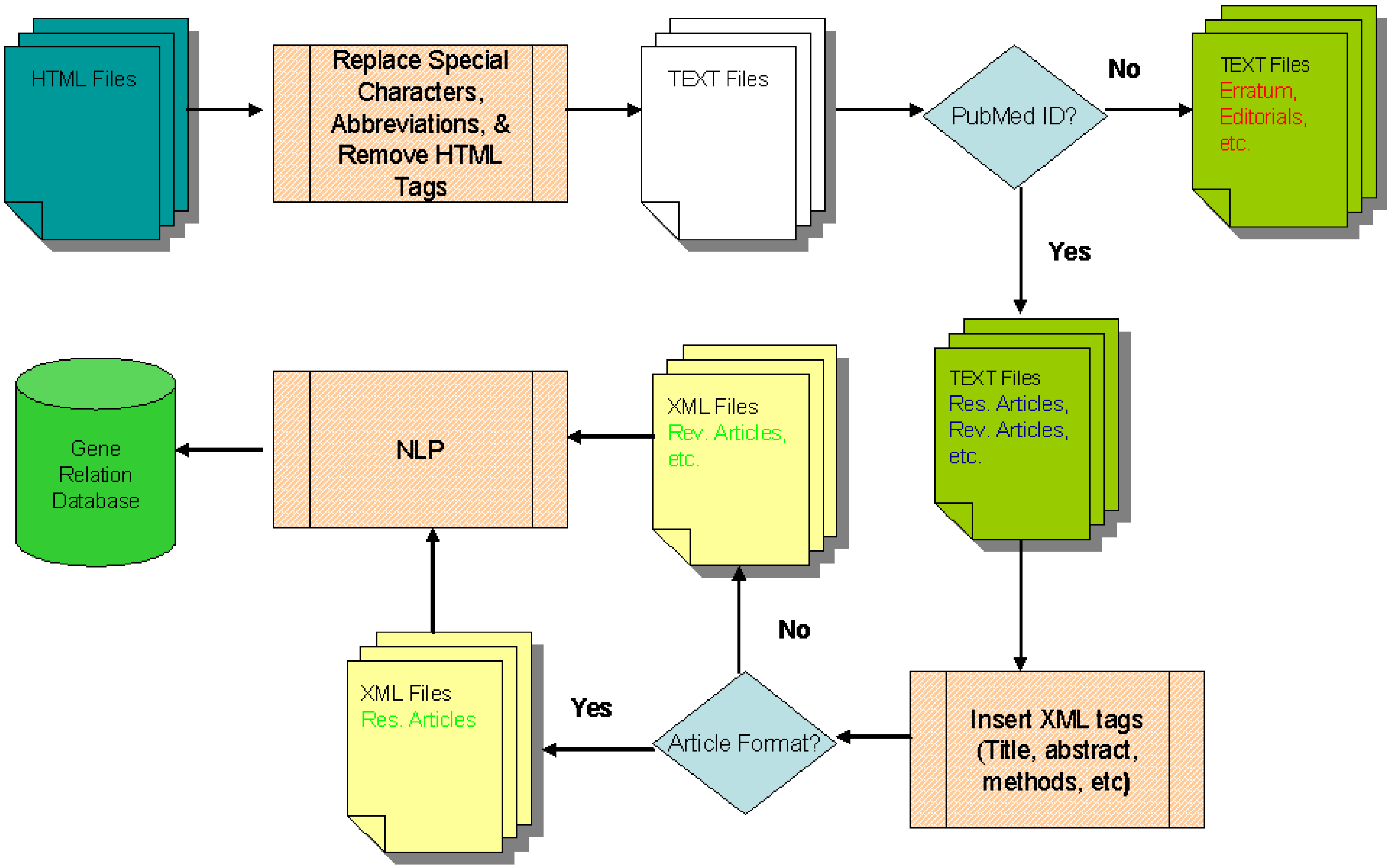
Text Mining:

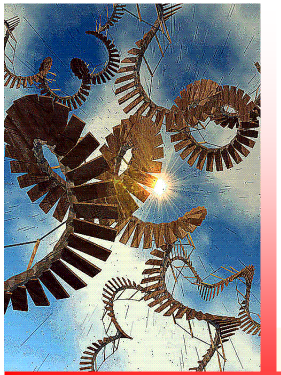
Applications Areas in Biology

- Help to address the following problems:
  - Finding biological named entities (e.g. protein, gene, chemical names etc.) in context to particular study
  - Finding molecule interactions (e.g. protein-protein interactions, protein-gene interactions etc.)
  - Finding relations between bio-concepts (e.g. relations between genes-disease, disease-drug)
  - Finding bio-chemical pathways
  - Finding sub-cellular localization information of proteins
  - Constructing biological vocabulary/ontology from text
  - Automatically Curating biological databases
  - Assisting gene expression data mining process
  - Knowledge-based information retrieval in context to biological repositories (e.g. MEDLINE etc.)

Text Mining

Sample Data Processing – Biomedical Text

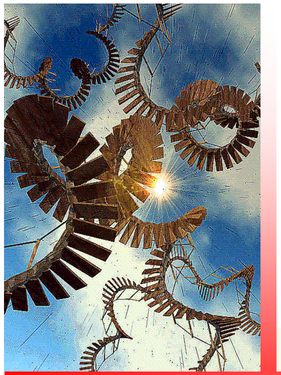

HTML Files
Replace Special Characters, Abbreviations, & Remove HTML Tags
TEXT Files
PubMed ID?
No
TEXT Files Erratum, Editorials, etc.
Yes
TEXT Files Res. Articles, Rev. Articles, etc.
Insert XML tags (Title, abstract, methods, etc)
Article Format?
Yes
XML Files Res. Articles
No
XML Files Rev. Articles, etc.
NLP
Gene Relation Database



Text Mining:

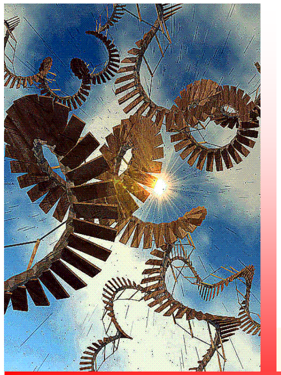
BioMedical Text Mining Systems - Examples

- iHOP
  - http://www.ihop-net.org/UniPub/iHOP/
  - Gene centric search Engine
- EBIMed
  - http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp
  - Concept based search linked to Uniprot
- GoPubMed
  - http://www.gopubmed.org/
  - Clusters documents based on Gene/MesH Ontology
- BioMinT
  - http://biomint.pharmadm.com/
  - An easy to use information retrieval and extraction tool
- Textpresso
  - http://www.textpresso.org/
  - Text categorization genome search engine



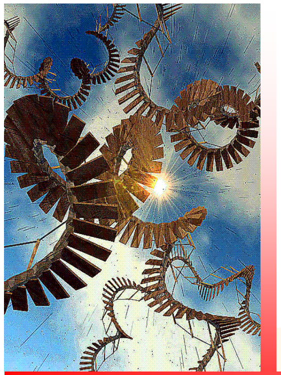
Reference

- Shatkay H., “Hairpins in bookstacks: Information retrieval from biomedical text”, *Briefings in Bioinformatics,* Vol. 6(3), 222-238, (2005).
- Natarajan J., Berrar D., Hack C.J., Dubitzky W., “Knowledge discovery in biology and biotechnology texts: A review of techniques, evaluation strategies, and applications”, *Critical Reviews in Biotechnology*, Vol. 25, 31-52, (2005).
- Krallinger M., Valencia A., “Text-Mining and Information-Retrieval Services for Molecular Biology”, *Genome Biology*, Vol 6, 224 (2005).



Acknowledgement

- Prof. Werner Dubitzky – Univeristy of Ulster
- Dr. Daniel Berrar – Unveristy of Ulster
- Martin Krallinger and Ashish V Tendulkar – APBIO Text Mining Tools in Biology
- Dr. Hagit Shatkay http://www.shatkay.org/



Thank You

Contact:

N. JEYAKUMAR: n.jeyakumar@yahoo.co.in