What is Text Mining?

Sophia Ananiadou
National Centre for Text Mining
www.nactem.ac.uk
University of Manchester

Outline

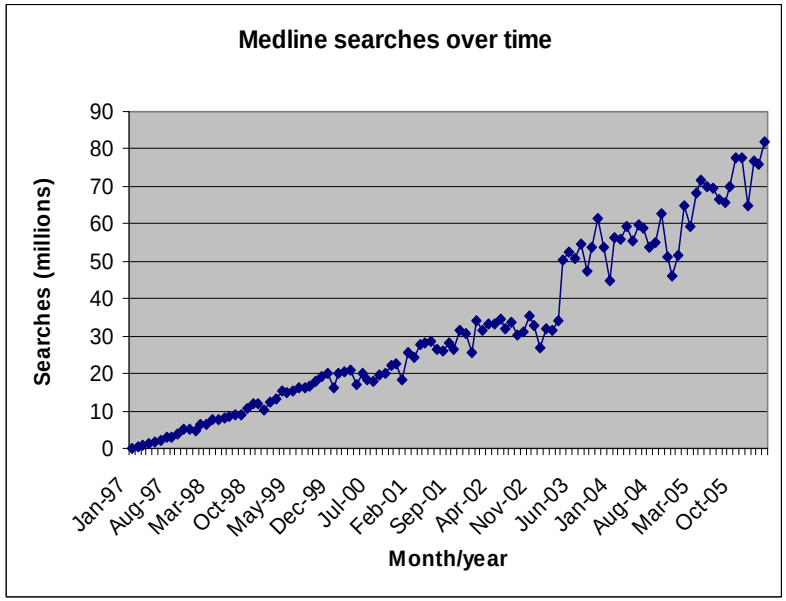
- Aims of text mining
- Text Mining steps
- Text Mining uses
- Applications

Aims

- Extract and discover knowledge hidden in text automatically
- Aid domain experts by automatically:
  - identifying concepts
  - extracting facts/relations
  - discovering implicit links
  - generating hypotheses

Why do we need text mining? Too much information!

- Information overload
- Growth of searches



- Information overlook
- Many heterogeneous resources
  - Text (blogs, papers, surveys, news..)
  - Databases
  - Web
  - Ontologies

The role of text in information management

- We are inundated with huge amounts of data
  - Unstructured information (text)
  - Different text types, genres, domains..
  - Semi-structured information (XML + text)
  - Structured information (databases)
- We need to make sense of data
- We need to manage information and knowledge effectively

UK National Centre for Text Mining (NaCTeM)

- The 1st national text mining centre in the world www.nactem.ac.uk
- **Remit**: **Provision of text mining services to support UK research**
- **Funded by**: the JISC, BBSRC, EPSRC
- **Phase I (2005-2008)**: biology
- **Phase II (2008-2011)** : bio-medicine, social sciences
  - humanities

Why is there a need in the UK for a national centre for text mining?

- Some researchers knew they wanted TM
- TM key component of **e-Science**
- Involve more researchers (from all domains) in doing e-science and e-research
  - TM seen as **key technology** for researchers
  - And one **applicable in every domain** (broad interest/support from major funding bodies)

Embedding Text Mining within e-Science in the UK

e-Science [...] enables new research and increases productivity through *shared e-Infrastructure*, the development of computational and logical models and new ways to discover and use the growing range of *distributed* and *interoperable* resources. It supports **multidisciplinary** and **collaborative working** and a culture that adopts the emerging methods.

M. Atkinson (2007) ***Beyond e-Science***

What the users want to do with their data (minimally)

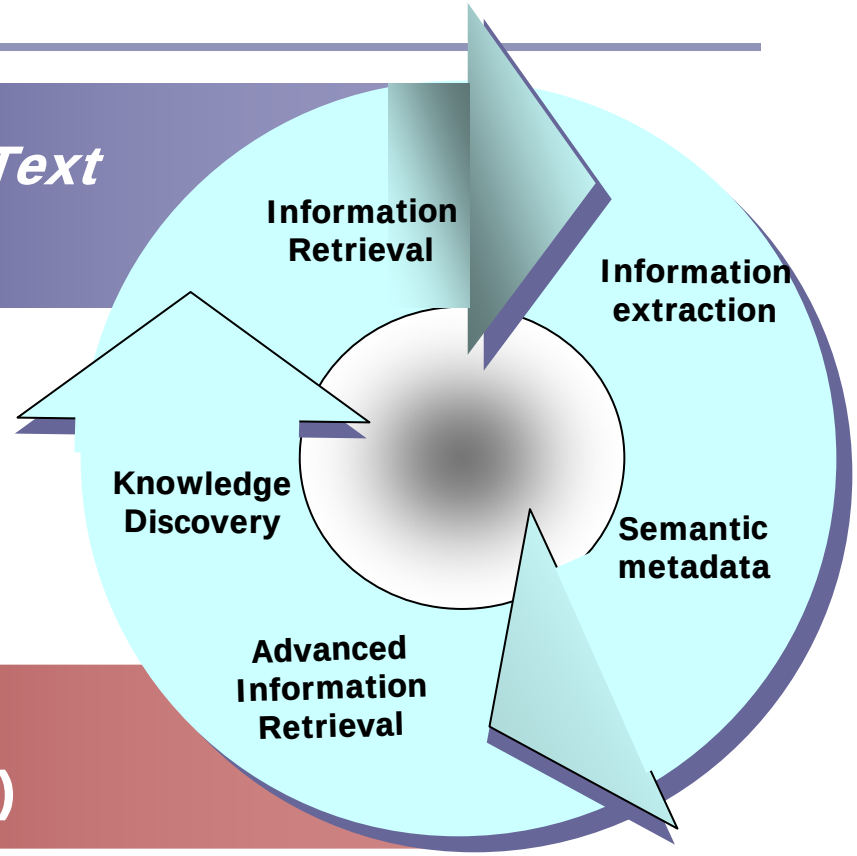
- Easier access to data
- Sharing data with their peers
- Annotating data with metadata
- Managing data across locations
- Integrating data within workflows, Web Services
- Aids for semantic metadata creation; enriching data with related metadata e.g. experimental results

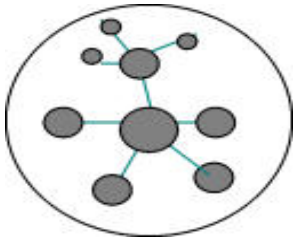
TEXT MINING RIDES TO THE RESCUE

From Text to Knowledge:

tackling the data deluge through text mining

Unstructured Text
(implicit knowledge)





Structured content
(explicit knowledge)

Text Mining Steps (1/3)

- Information Retrieval (Google)
  - Finding within a large document collection, a subset of documents, relevant to a user's query
    - Query term "blood" or Boolean query "blood pressure"
  - Too many documents are retrieved, prohibitively large for humans to read
  - Many retrieved documents are irrelevant to our query
  - Many relevant documents are missing

Text Mining Steps (2/3)

- Information Extraction, nuggets of text
  - Identify information nuggets from text, fill existing templates, create structured information, populate text databases

Slot	Information
Date	7/10/96 (today)
Location	SanSalvador
Victim injured	policeman
Victim attacked	guards
Perpetrator	urban guerrillas

San Salvador, 7/10/96

It has been officially reported that a policeman was wounded today when urban guerrillas attacked the guards at a power substation located downtown San Salvador.

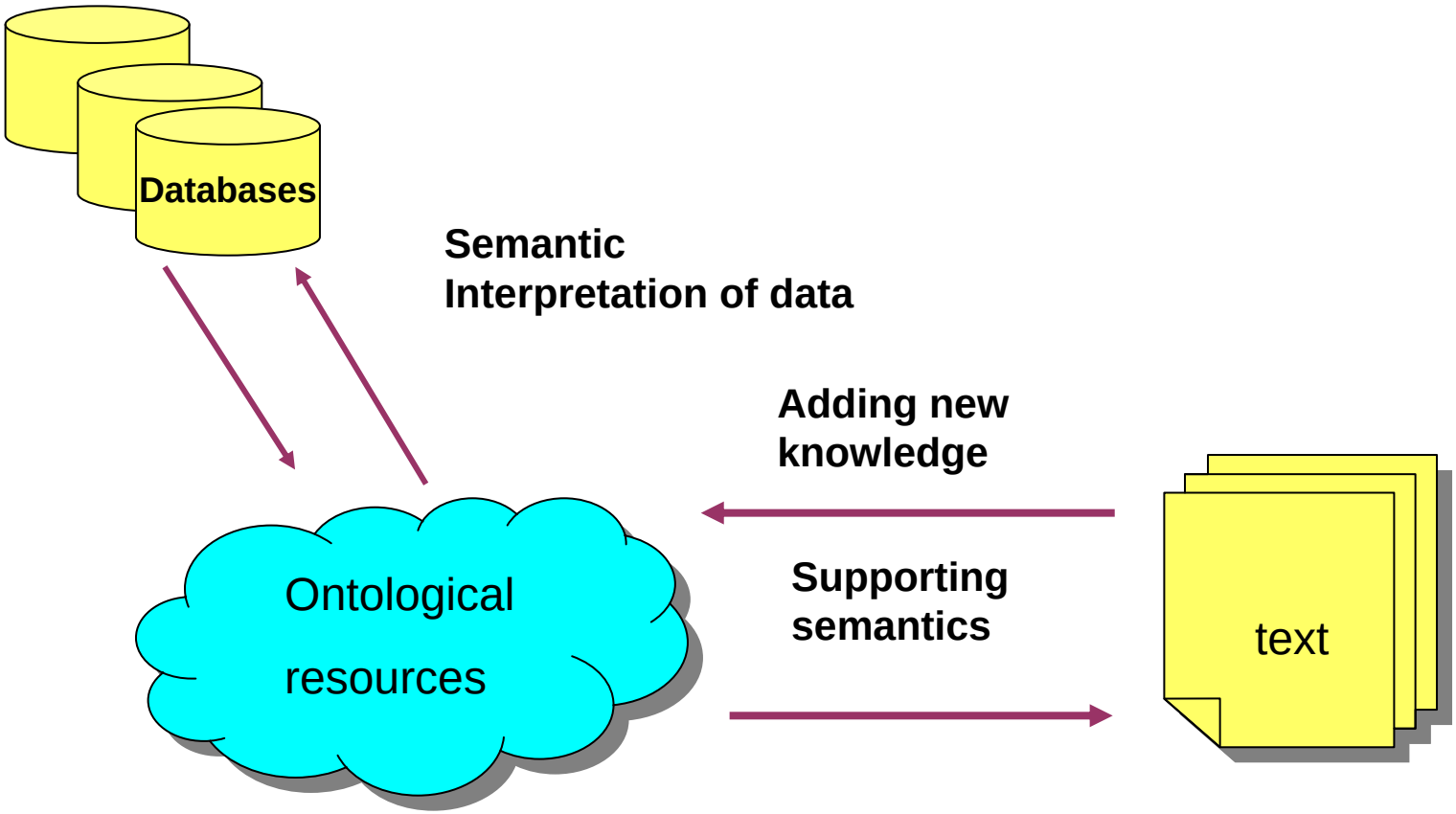
Information Extraction

- recognise specific relations and events (typically expressed by verbs – *attacked*)
- domain restricted (newswire, biology…)
- more complex NLP techniques
  - more than 'bag of words' approach
  - deep parsing
- Output: filled templates of entities and facts
- IE extracts only what we are looking for, i.e. what has been defined by patterns

Text Mining Steps (3/3)

- Data Mining: finds associations among pieces of information extracted from many different texts, implicit links
- Integration with databases, ontologies

Integrating Text with DBs and Ontologies



Uses (1/

Business applications

- Business intelligence (market analysis), competitors, identify risks, make predictions
- Customer views and opinions from blogs (opinion mining, trends analysis)
- Find nuggets of relevant information immediately, systematically
- Remove tedious process of finding information

Uses (2/3)

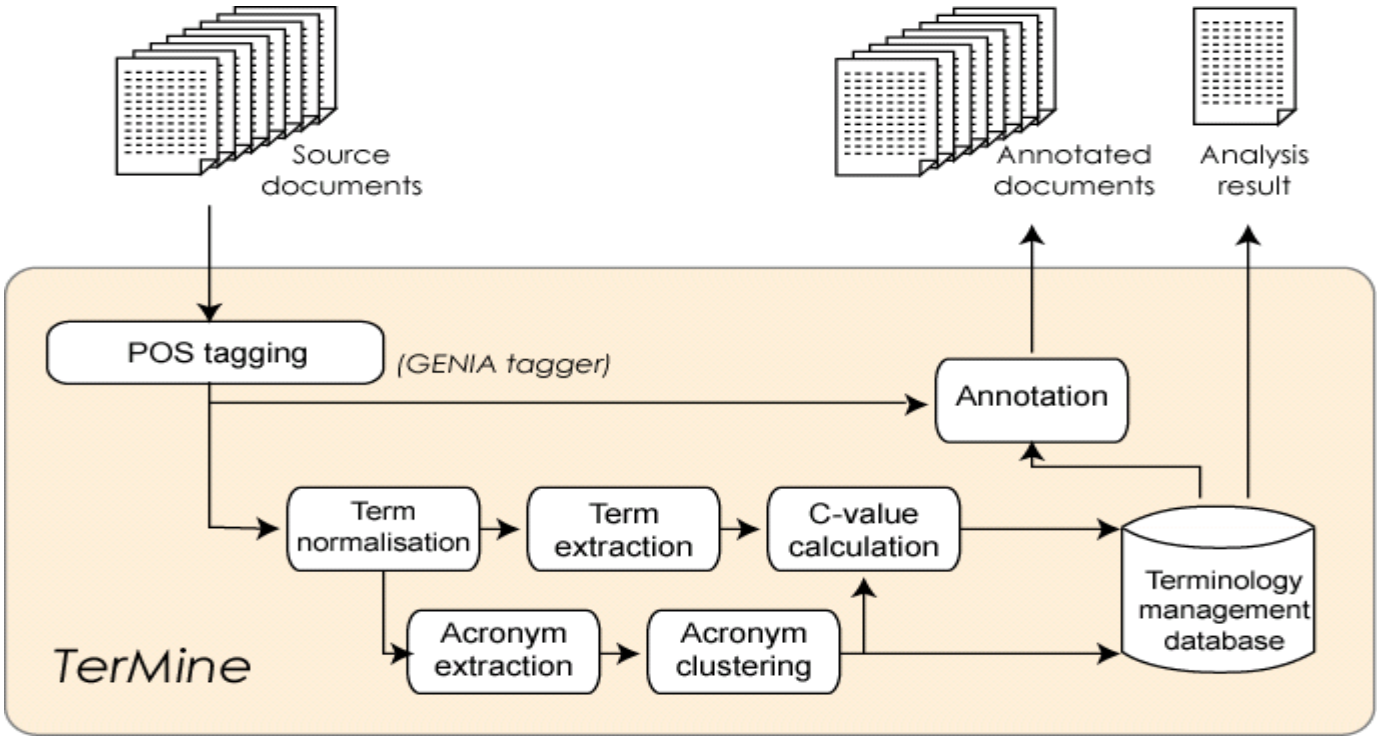
Media

- Semantic searching needed for new models of information access, linking and extraction
- Personalisation of searching
  - Document classification and clustering based on personalised queries
  - Social networking + text mining
  - Topic clusters of news
  - Frame analysis of news

Uses (3/3)

- Text Mining enriches text with semantic annotations
  - **For authors tools for semantic annotation**
    - **Intelligent information management**
  - **For publishers enrichment of digital libraries**
  - **For scientists intelligent searching, linking and integration of text, databases**

Applications – extracting terms



TerMine [GUI]

File Edit View Help

Source documents

Beta-arrestin binding to the beta2-adrenergic receptor requires both receptor phosphorylation and receptor activation.

Krasel C, Bunemann M, Lorenz K, Lohse MJ.
Institute for Pharmacology and Toxicology, Versbacher Strasse 9, D-97078 Wurzburg, Germany.

Homologous desensitization of beta2-adrenergic receptors has been shown to be mediated by phosphorylation of the agonist-stimulated receptor by G-protein-coupled receptor kinase 2 (GRK2) followed by binding of beta-arrestins to the phosphorylated receptor. Binding of beta-arrestin to the receptor is a prerequisite for subsequent receptor desensitization , internalization via clathrin-coated pits , and the initiation of alternative signaling pathways. In this study we have investigated the interactions between receptors and beta-arrestin2 in living cells using fluorescence resonance energy transfer. We show that (a) the initial kinetics of beta-arrestin2 binding to the receptor is limited by the kinetics of GRK2-mediated receptor phosphorylation ; (b) repeated stimulation leads to the accumulation of GRK2-phosphorylated receptor , which can bind beta-arrestin2 very rapidly ; and (c) the interaction of beta-arrestin2 with the receptor depends on the activation of the receptor by agonist because agonist withdrawal leads to swift dissociation of the receptor-beta-arrestin2 complex. This fast agonist-controlled association and dissociation of beta-arrestins from prephosphorylated receptors should permit r
receptor sensitivity in repeatedly stimulated cells such as neurons.

beta2-adrenergic receptor gene single-nucleotide polymorphisms are
with rheumatoid arthritis in northern Sweden.

Xu B, Arlehag L, Rantapaa-Dahlquist SB, Lefvert AK.
Department of Immunology, American Red Cross Biomedical Research and Development, MD 20855, USA. xub

The beta2-adrenergic receptor (beta2-AR) belongs to the group of G-protein-coupled receptors and is present on skeletal and cardiac muscle cells and on lymphocytes. The gene encoding beta2-AR (ADRB2) displays a moderate degree of heterogeneity in the human population and the distributions of single-nucleotide polymorphisms (SNPs) at amino acid positions 16 , 27 , and 164 are changed in asthma , obesity , and hypertension and in the autoimmune disease myasthenia gravis. An involvement of the beta2-AR has also been suggested in human rheumatoid arthritis (RA) and its animal model. We describe here an increased prevalence of the alleles Arg16 and Gln27 and a lower prevalence of homozygosis for Gly16 and Glu27 in patients with RA. Patients having the genotype combination GlyGly16-GlnGlu27 had higher levels of rheumatoid factor (RF) and a more active disease than other patients. Patients having the genotype Arg16-Gln27+ had higher levels of RF when compared to those having Arg16+Gln27+ , and patients who were carriers of Gln27 had a more active disease than non-carriers of Gln27. Our results show an association of beta2-AR SNPs with RA in a population from the northern part of Sweden. Our study also confirms the strong linkage disequilibrium of genotypes at amino acid

Result 1 - 50 of about **1131** terms

Rank	Term	Score
1	beta2-adrenergic receptor	65.7778
2	blood pressure	16.8
3	beta2-adrenergic receptor gene	14.8496
4	single nucleotide polymorphisms	9.50977
5	adrenergic receptor	9.14286
6	Glu16 allele	8
[illegible]	[illegible]	8
[illegible]	[illegible]ex	7.92481
[illegible]	[illegible]	7.92481
[illegible]	[illegible]	7.625
11	cystic fibrosis	7.33333
12	confidence interval	7
13	metabolic syndrome	7
14	allelic frequency	7
15	bioluminescence resonance energy transfer	6.8

Terms derived from text used as Index terms

Ready

Problems – term variation & ambiguity

- Acronyms

ER estrogen receptor
emergency room
endoplasmic reticulum

- Spelling

Tumour – tumor
Oestrogen - estrogen
NF-kB, NF-KB,
NF-kappa B,
nuclear factor kappa B

- Gene terms may be also common English words
- *BAD* human gene encoding BCL-2 family of proteins
- *bad news, bad prediction*

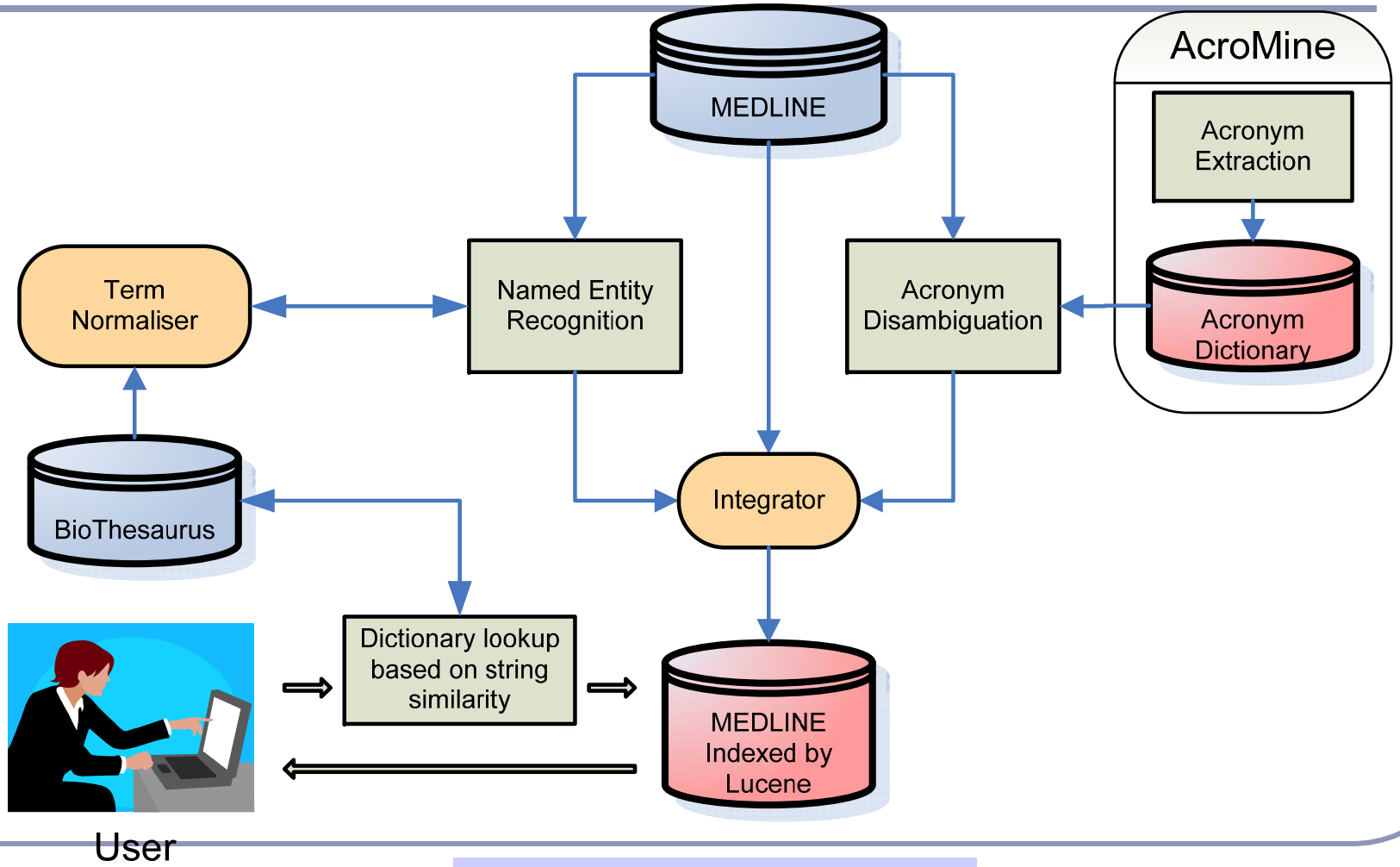
Semantic searching based on named-entity annotation

The peri-kappa B site mediates human immunodeficiency virus type 2 enhancer activation in monocytes …

DNA virus cell_type

- Entity types (defined by Ontologies)
  - Genes/protein names
  - Enzymes, substances, etc.
  - Names (people, organisations…)

KLEIO architecture



Kleio @ NaCTeM - Mozilla Firefox (nactem4.mc.man.ac.uk)

File Edit View Go Bookmarks Tools Help

http://nactem4.mc.man.ac.uk:8080/Kleio/SimpleSearch.c Go

Red Hat, Inc. Red Hat Network Support Shop Products Training

NaCTeM
The National Centre for Text Mining

New Query

NaCTeM Services
Termine
Acromine
Cheshire/Termine
Medie
Info-Pubmed

Query: cat

Results 1 - 10 of 60711

First Previous Next Last

1 2 3 4 5 6 7 8 9 10 11 ...

Axillary lymphadenopathy secondary to cat-scratch disease.
Journal: Irish medical journal. 2005 Sep;98(8):243-4
[No abstract listed]
PubMedID 16255119 - View Abstract

Early relief of osteoarthritis symptoms with a natural mineral supplement and a herbomineral combination: a randomized controlled trial [ISRCTN38432711].
Journal: Journal of inflammation (London, England) 2005 Oct;2:11
... , sierrasil, alone and in combination with a cat's claw extract (Uncaria guianensis), vincaria, has ... (2 g/day) + cat's claw extract (100 mg/day) or placebo, administered for 8 weeks ... with a cat's claw extract, improved joint health and function within 1-2 weeks ...
PubMedID 16242032 - View Abstract

The cover. Black Cat on a Chair.
Journal: JAMA 2005 Oct;294(15):1867
[No abstract listed]
PubMedID 16234486 - View Abstract

Secondary calcium-binding parameter of Bacillus amyloliquefaciens alpha-amylase obtained from inhibition kinetics.
Journal: J. Biosci. Bioeng. 2003;96(3):262-7
... the equilibrium dissociation constant (K(m)) and k(cat) for the hydrolytic catalysis. The enzymatic ... concentration. Because k(cat) was practically constant at the high calcium concentration range ...
PubMedID 16233519 - View Abstract

The action modes of an extracellular beta-1,3-glucanase isolated from Bacillus clausii NM-1 on beta-1,3-glucooligosaccharides.
Journal: J. Biosci. Bioeng. 2003;96(1):32-7
... laminarihexaose were rapidly hydrolyzed, while laminaritetraose was slowly hydrolyzed. The k(cat)/K(m ... terminal. The value of k(cat)/K(m) also suggested that the sixth binding position ...
PubMedID 16233479 - View Abstract

Purification and characterization of malate dehydrogenase from Corynebacterium glutamicum.
Journal: J. Biosci. Bioeng. 2003;95(6):562-6
... both for NADU and NADPH as coenzyme on the bases of the k(cat) values at pH 6.5 which is the optimum pH for the both coenzymes. Plotting of the logarithms of the 1/Km, k(cat), and k(cat)/K(m) values with respect to oxalacetate against pH lead to speculation ...
PubMedID 16233457 - View Abstract

Employing chimeric xylanases to identify regions of an alkaline xylanase participating in enzyme activity at basic pH.
Journal: J. Biosci. Bioeng. 2002;94(5):395-400
... in determinations of the k(cat) values. The pK(a1) values for the APnc and Apnc chimeric enzymes ...
PubMedID 16233324 - View Abstract

Replacement of His12 or His119 of bovine pancreatic ribonuclease A with acidic amino acid residues for the modification of activity and stability.

Done

Kleio @ NaCTeM - Mozilla Firefox (nactem4.mc.man.ac.uk)

File Edit View Go Bookmarks Tools Help

http://nactem4.mc.man.ac.uk:8080/Kleio/SimpleSearch.c Go

Red Hat, Inc. Red Hat Network Support Shop Products Training

NaCTeM
The National Centre for Text Mining

New Query

NaCTeM Services
Termine
Acromine
Cheshire/Termine
Medie
Info-Pubmed

Query: PROTEIN:cat

Results 1 - 10 of 237

First Previous Next Last

1 2 3 4 5 6 7 8 9 10 11 ...

Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage.
Journal: Nucleic Acids Res. 2005;33(6):e58
... (cat) was synthesized using ...
PubMedID 15800209 - View Abstract

[Site-directed mutagenesis and promoter functional analysis of RM07 DNA fragment from Halobacterium halobium in Escherichia coli]
Journal: Yi Chuan Xue Bao 2004 May;31(5):525-32
... acetyltransferase (cat) reporter gene in pKK232-8 in ...
PubMedID 15478616 - View Abstract

[Experimental study on phenotypic conversion of clinical chloromycetin-resistant strains of E. coli to drug-sensitive strains by using EGS technique in vitro]
Journal: Zhonghua Yi Xue Za Zhi 2004 Aug;84(15):1294-8
... Cm acetyl transferase (cat) and containing kanamycin ...
PubMedID 15387969 - View Abstract

Characterization of a baculovirus lacking the alkaline nuclease gene.
Journal: J. Virol. 2004 Oct;78(19):10650-6
... acetyltransferase gene (cat) and a bacmid containing the ...
PubMedID 15367632 - View Abstract

Effect of 3' terminal codon pairs with different frequency of occurrence on the expression of cat gene in Escherichia coli.
Journal: Curr. Microbiol. 2004 Feb;48(2):97-101
... acetyltransferase (cat) gene expression in E. coli ... opposite effect on the yield of CAT protein in comparison with ...
PubMedID 15057475 - View Abstract

New chiral ruthenium(II) catalysts containing 2,6-bis(4'-(R)-phenyloxazolin-2'-yl)pyridine (Ph-pybox) ligands for highly enantioselective transfer hydrogenation of ketones.
Journal: Chemistry (Weinheim an der Bergstrasse, Germany) 2004 Jan;10(2):425-32
... the presence of NaOH (ketone/cat/NaOH 500:1:6). cis-Ph-pybox ...
PubMedID 14735511 - View Abstract

The acrAB locus is involved in modulating intracellular acetyl coenzyme A levels in a strain of Escherichia coli CM2555 expressing the chloramphenicol acetyltransferase (cat) gene.
Journal: Arch. Microbiol. 2003 Nov;180(5):362-6
... resistance gene (cat) from a multicopy plasmid. ...
PubMedID 14614545 - View Abstract

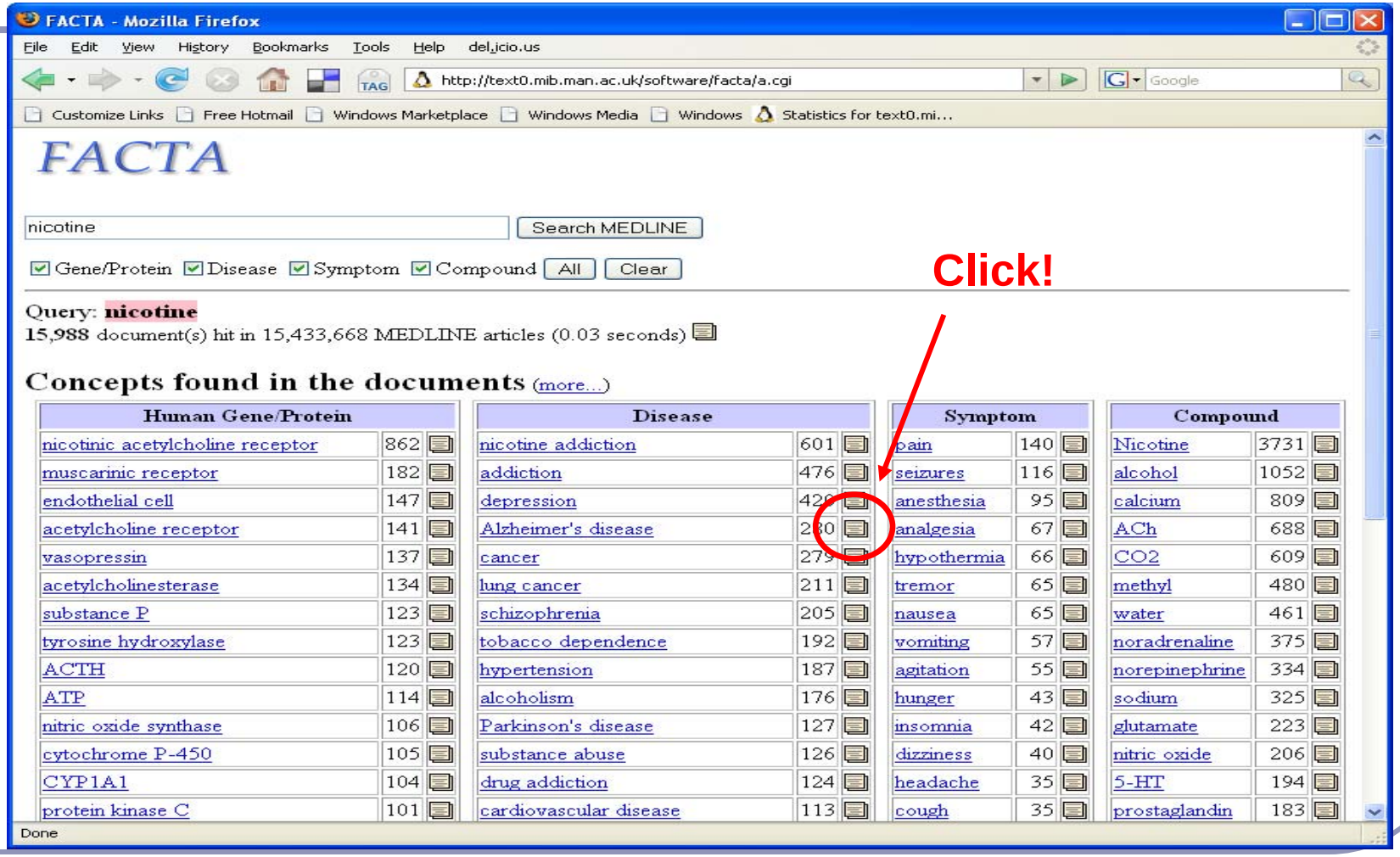
Evidence supporting a major promoter in the Trypanosoma cruzi rRNA gene.
Journal: FEMS Microbiol. Lett. 2003 Aug;225(2):221-5
... acetyl transferase gene (cat) as a reporter. The data ...
PubMedID 12951245 - View Abstract

Enhanced secretion of heterologous cyclodextrin glycosyltransferase by a mutant of Bacillus licheniformis defective in the D-alanylation of teichoic acids.
Journal: Lett. Appl. Microbiol. 2003;37(1):75-80

Done

Fewer documents with more precise query

Extracting associations between entities



FACTA - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

http://text0.mib.man.ac.uk/software/facta/a.cgi?query=nicotine|1111|0|15066

Customize Links | Free Hotmail | Windows Marketplace | Windows Media | Windows | Statistics for text0.mi...

other disorders. We report the development and evaluation of a putative antagonist, 5-(3'-fluoropropyl)-3-(2-(S)-pyrrolidinylmethoxy)pyridine (nifrolidine) as a PET agent for **nicotine** alpha(4)beta(2) receptors. ... Blocking studies were performed by subcutaneous injection of **nicotine** (10 mg/kg). ... This specific binding was completely abolished by 300 mumol/L **nicotine**. ...
PMID:15632043 J. Nucl. Med. 2005 Jan

Repeated **nicotine** exposure in rats: effects on memory function, cholinergic markers and nerve growth factor.
A decrease in the number of nicotinic-acetylcholine receptors (nAChRs) in the brain is thought to contribute to the cognitive dysfunction associated with diseases as diverse as **Alzheimer's disease** and schizophrenia. Interestingly, **nicotine** and similar compounds have been shown to enhance memory function and increase the expression of nAChRs and therefore, could have a therapeutic role in the aforementioned diseases. **Nicotine** has also been shown to exert positive effects on certain neurotro ch as nerve growth factor (NGF), and therefore could play a role beyond mere symptomatic therapy. ... Studies to further investi cotine on NGF especially its high- and low-affinity receptors are also needed. In the present study, male Wistar rat of **nicotine** (0.35 mg/kg every 12 h) for 14 days demonstrated improved memory perform ated that **nicotine** learning- a phospho-T positive eff
PMID:156

Current s
... This obs subjects w representa
PMID:153

Galantan
Galantamin protected ne
PMID:15541385

... Alzheimer's disease and schizophrenia. Interestingly, **nicotine** and similar compounds have been shown to enhance memory function and increase the expression of nAChRs and therefore, could have a therapeutic role in the aforementioned diseases.

Nicotinic acetylcholine receptor system and neuropsychiatric disorders.
... There is a growing body of evidence linking alterations in nicotinic receptor number and/or function to conditions such as schizophrenia,

Done

Text mining for social sciences

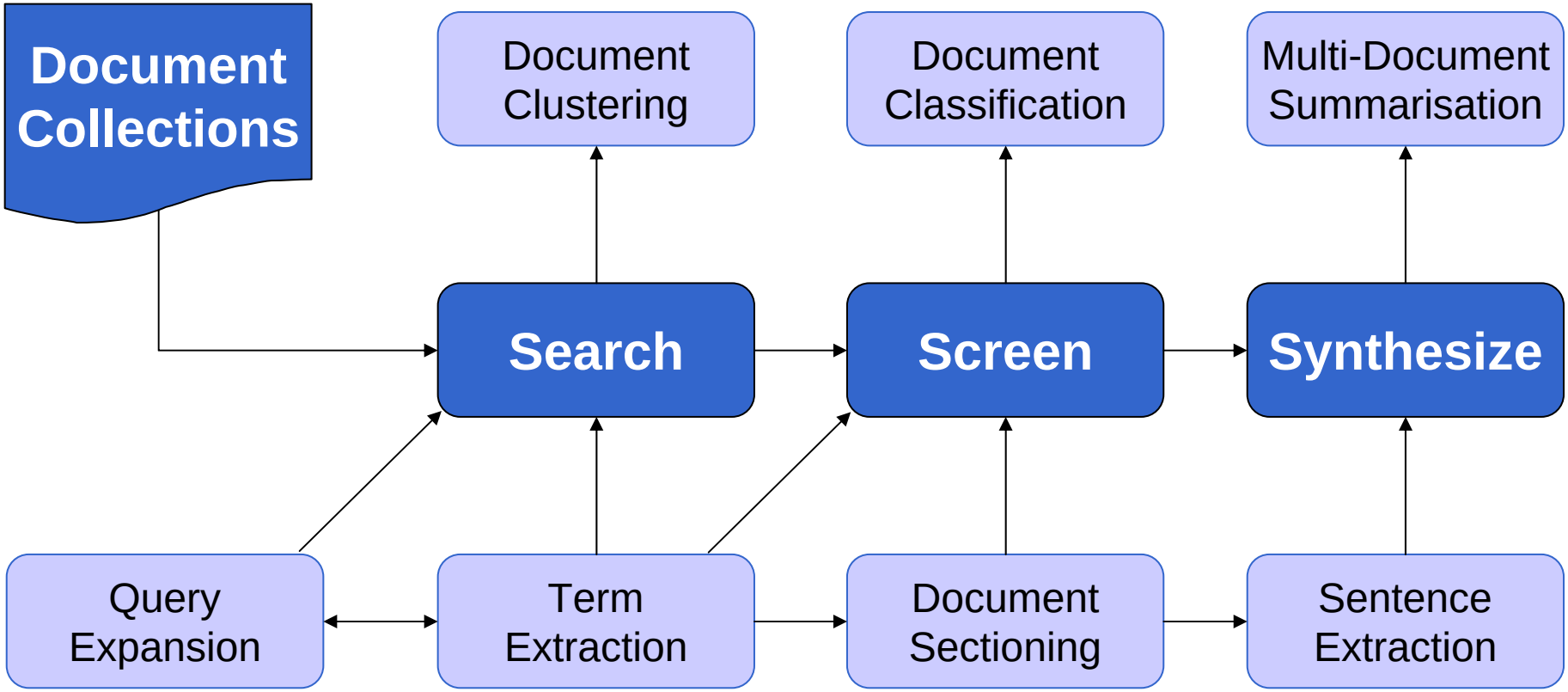
- ASSERT project (NaCTeM-NCeSS-EPPI)
- Assisting the process of systematic reviewing in social sciences
- Engaging the user community: EPPI

  (Evidence for Policy and Practice Information and Co-ordinating Centre)
  - Document classification
  - Information extraction
  - Summarisation

The process of Systematic Reviewing

- *Searching:* extensive searches to locate as much relevant research as possible according to a query.
- *Screening*: narrows the scope of search to only the relevant documents to a specific review.
  - Highlights key evidence and results that may impact on the policy.
- *Synthesizing*: correlates evidence from a plethora of resources and summarises the results.
- The process is mainly manual

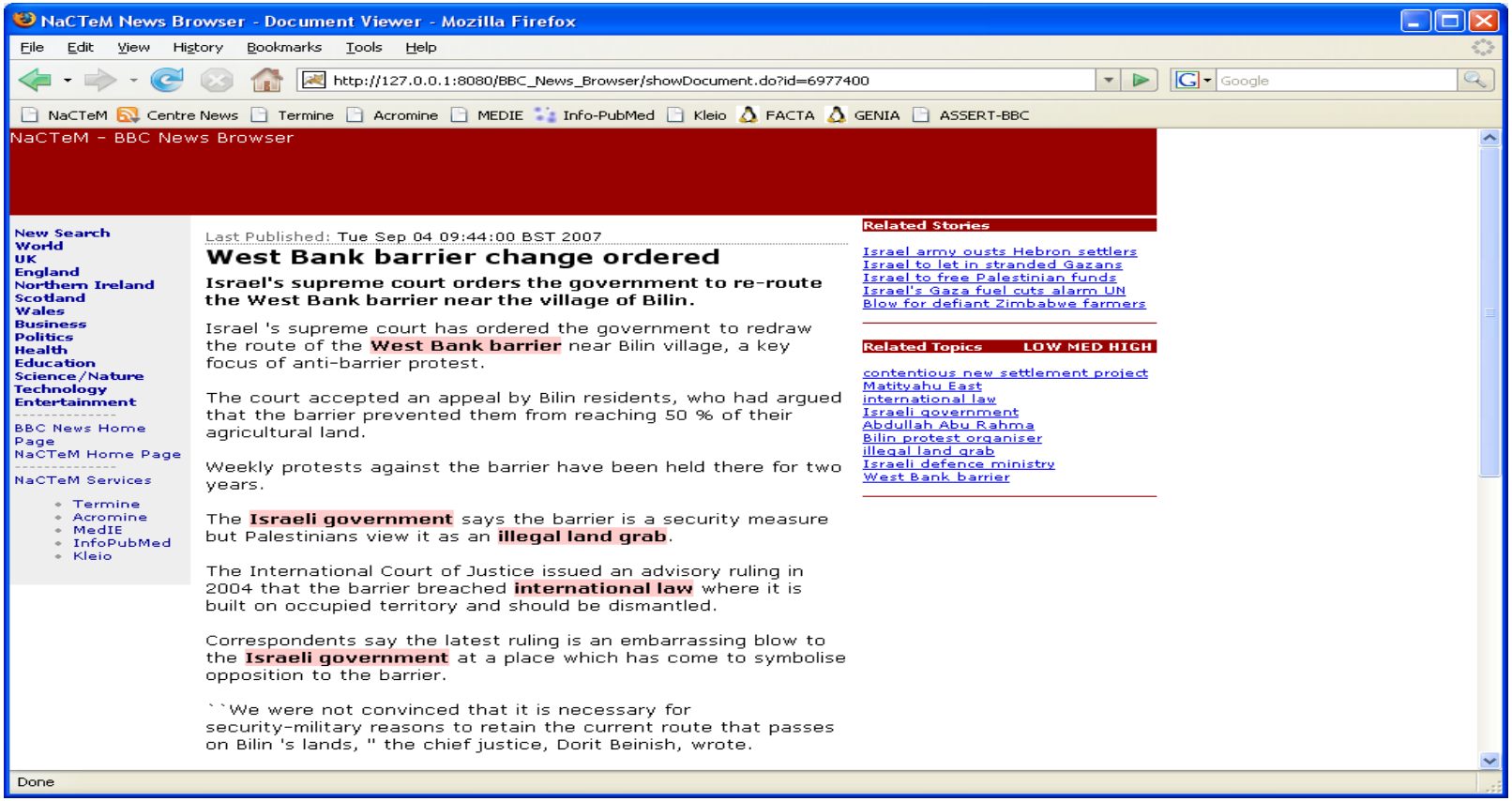
Overview of ASSERT



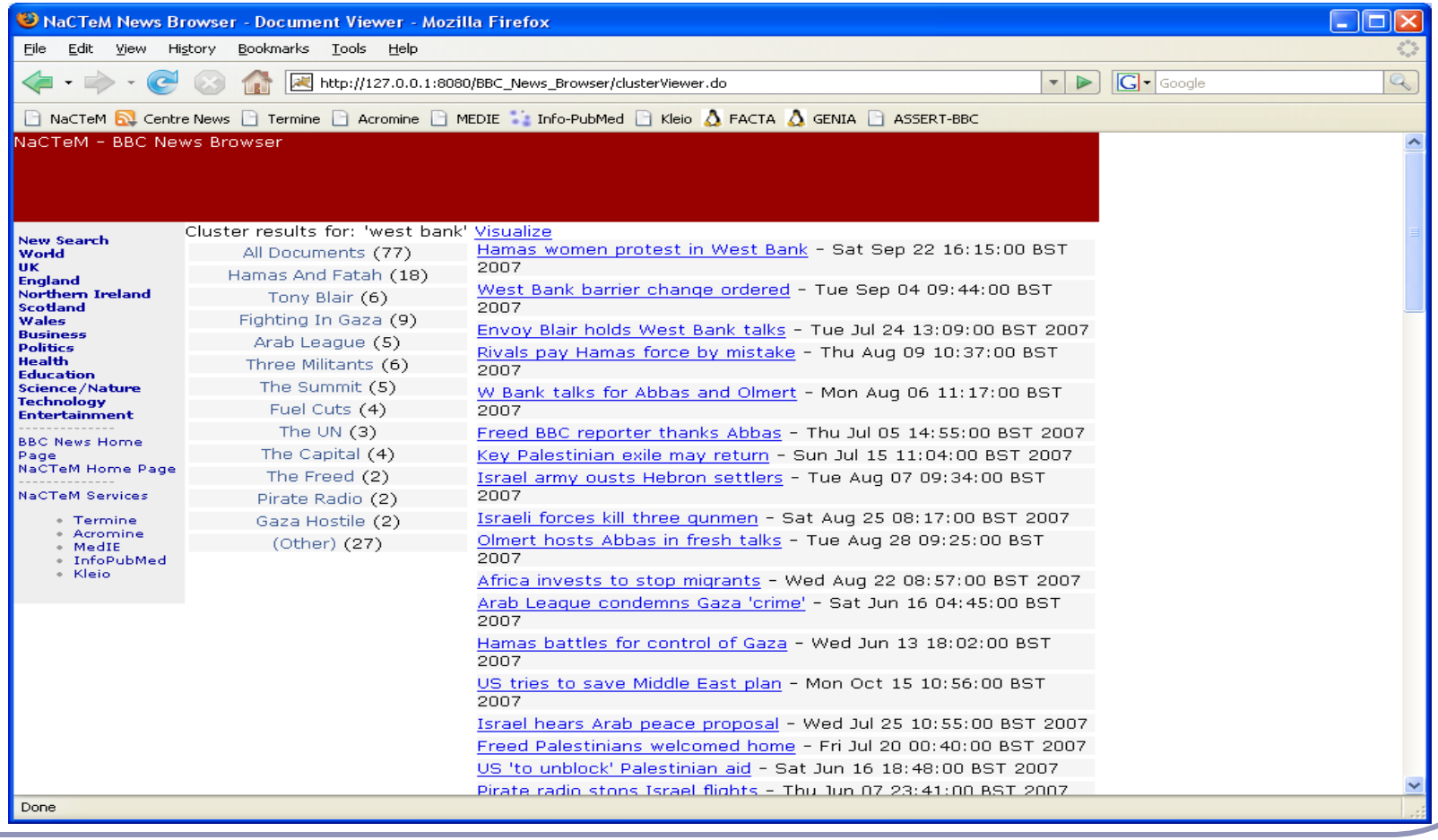
BBC Pilot project

- Analyse, structure and visualise BBC news online, according to a user's query using advanced text mining techniques
- Concept discovery and retrieval
  - interface allows a user to enter a query across the document collection and automatically calculate a list of concepts specific to the query and ranked by perceived importance.

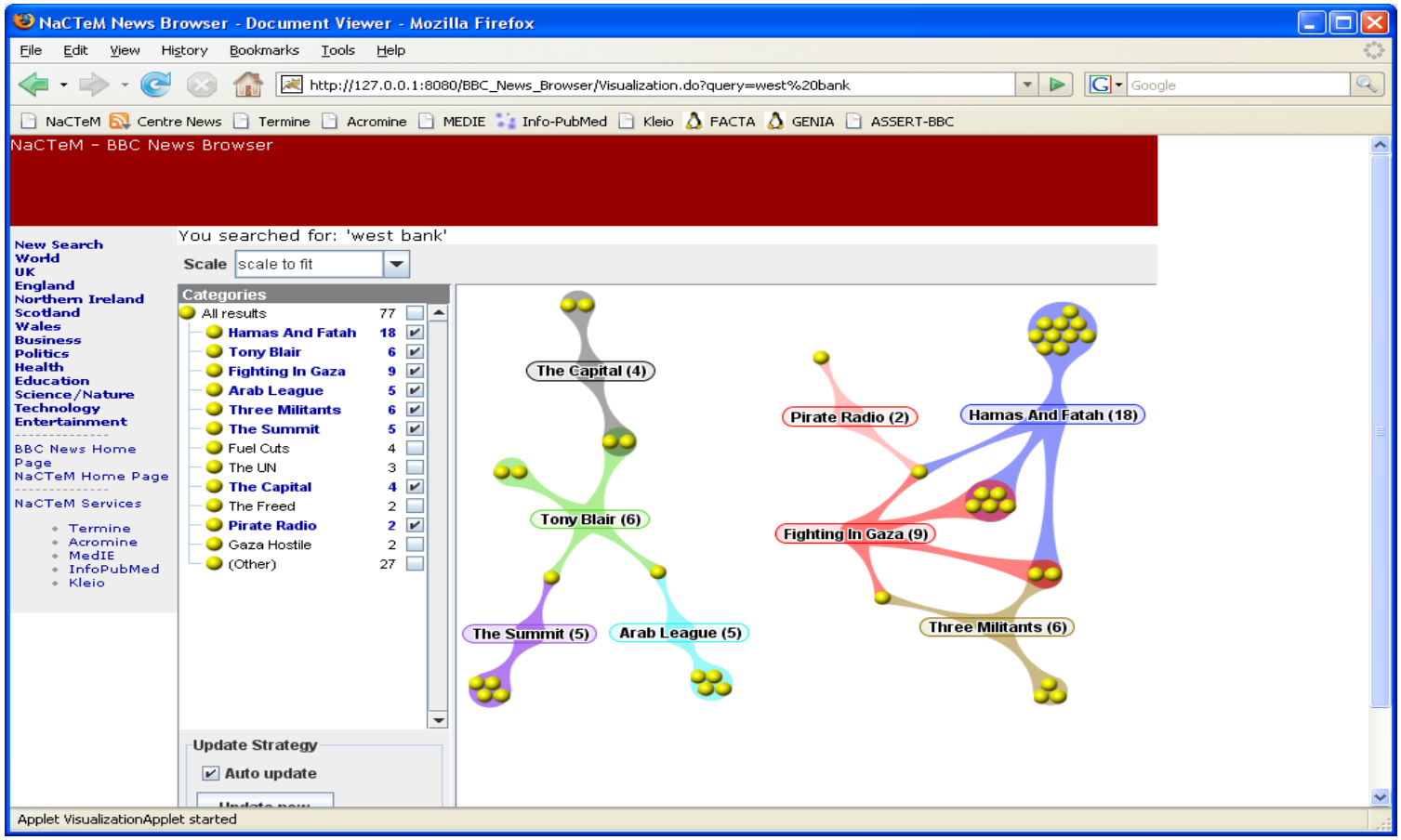
Finding news using text mining



Clustering the documents



Visualising the results



Benefits to Users

- Provision of a focused search with goal based results
- Allows expansion beyond known keywords for a more complete search
- Visualization of a result set creates an overview of the research in the domain
- Saves time and effort

What do our users/clients use our services for?

- Creation of controlled vocabularies, extraction of interactions, creation of models and networks, database curation (BOOTStrep)
- Bibliographic searching, automatic classification and semantic extraction in support of subject searching (ASSERT, INTUTE)
- Ontology building
- Media frame analysis (ASSIST)
- Semantic extraction to support indexing and linking across repositories (INTUTE)
- Extracting bio-processes, gene-disease mining (Pfizer)
- Maintaining and constructing pathways (REFINE)
- Classification of on-line news feeds, document classification
- Topic detection (BBC)

Text Mining for Knowledge Discovery

- Semantic enrichment relies on NLP based TM technologies
- Applications based on semantically annotated texts enable knowledge discovery
- Linking text with domain knowledge
  - Integrate other knowledge sources, ontologies, terminologies
  - Integration of distributed TM software (workflows)