Text Mining Techniques

John McNaught
Deputy Director
National Centre for Text Mining
John.McNaught@manchester.ac.uk

Overview

- Gentle introduction to TM
- Applications/case studies
  - To help understand what TM could do for you

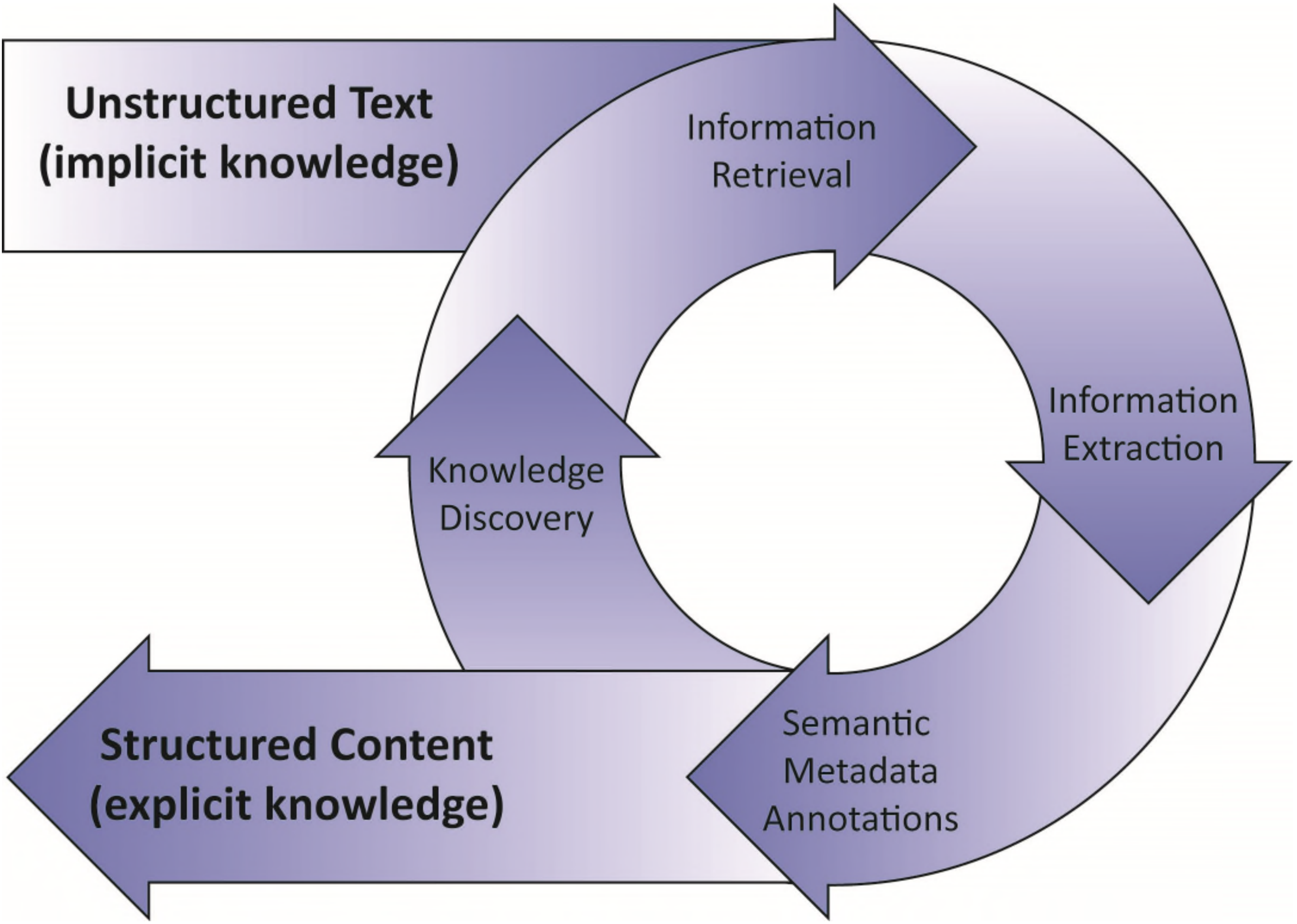
NaCTeM

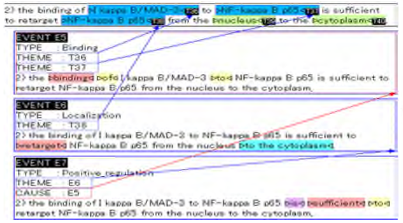
- First publicly-funded (JISC) national text mining centre in the world
  - Provide services to research community
- Initial focus on biology, then social sciences, medicine, chemistry, …
- Processing on a large scale, e.g. for UKPMC (Wellcome T.+17 other funders)
- www.nactem.ac.uk

What is text mining?

- Goal: Discover new knowledge from old
- How:
  - Process (typically) very large amounts of text
    - Millions of documents not unusual
  - Identify and extract information
  - (Link extracted information to already curated knowledge)
  - Mine to discover implicit significant associations
  - Flag (unknown) associations for researcher to investigate further
  - Spin-off on the way: render information explicit

From text to new knowledge





2) the binding of I kappa B/MAD-3 to NF-kappa B p65 is sufficient to retarget NF-kappa B p65 from the nucleus to the cytoplasm

EVENT E5
TYPE : Binding
THEME : T36
THEME : T37
2) the binding of I kappa B/MAD-3 to NF-kappa B p65 is sufficient to retarget NF-kappa B p65 from the nucleus to the cytoplasm.

EVENT E6
TYPE : Localization
THEME : T38
2) the binding of I kappa B/MAD-3 to NF-kappa B p65 is sufficient to retarget NF-kappa B p65 from the nucleus to the cytoplasm

EVENT E7
TYPE : Positive_regulation
THEME : E6
CAUSE : E5
2) the binding of I kappa B/MAD-3 to NF-kappa B p65 is sufficient to retarget NF-kappa B p65 from the nucleus to the cytoplasm.

MANCHESTER
1824
The University of Manchester

NaCTeM
The National Centre for Text Mining

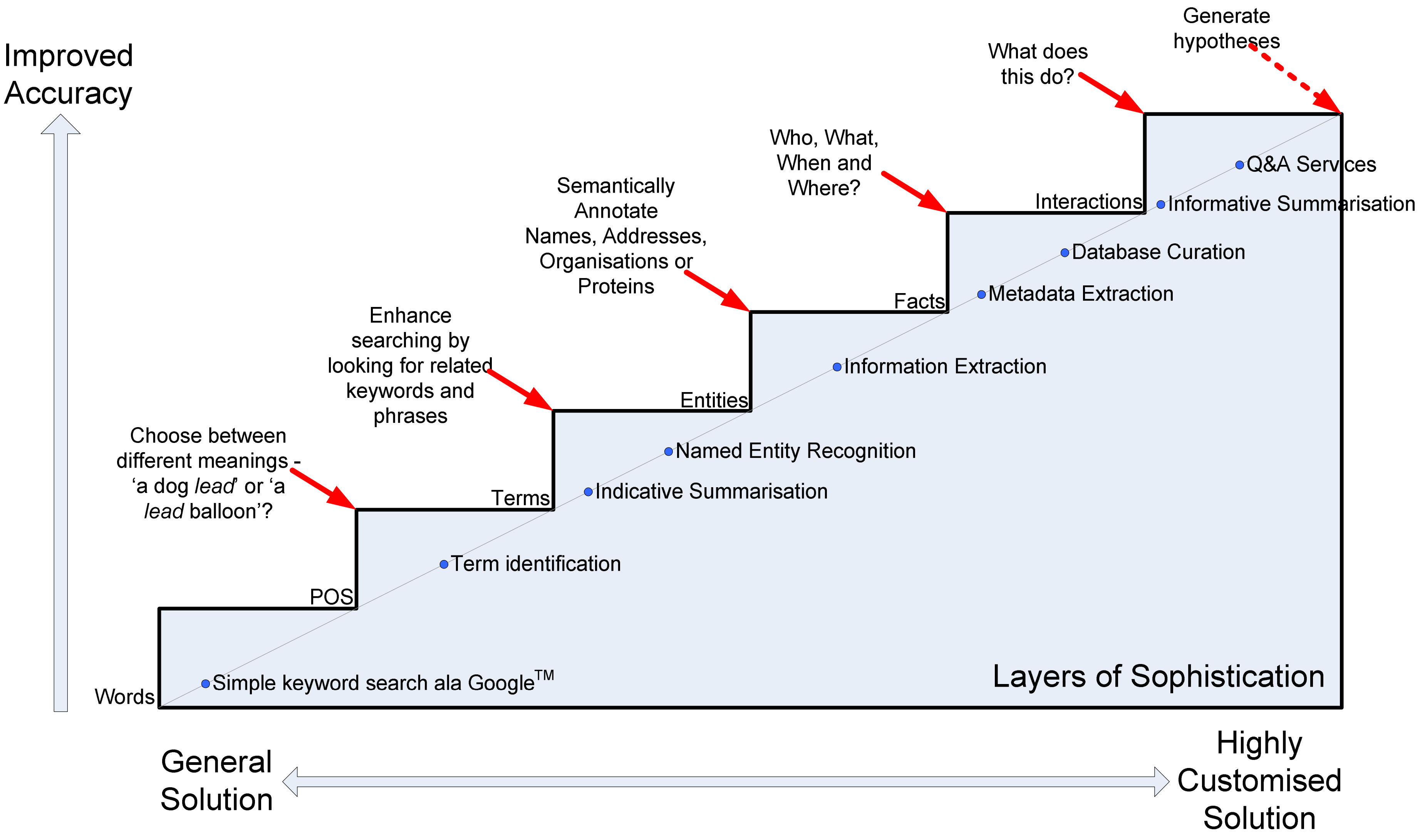
What does TM offer?

- Finds unsuspected knowledge
  - E.g. Disease-gene associations
- Enables discoveries human effort could not achieve (information overload/overlook)
- Enables better search/navigation of literature
  - Semantic search via extracted semantic metadata
- Reduces time spent searching
  - 15-48% of researcher time spent on classic search, 20-50% of classic searches unsatisfied

A complex problem

- TM involves
  - Many components (converters, analysers, miners, visualisers, ...)
  - Many resources (grammars, ontologies, lexicons, terminologies, thesauri, CVs)
  - Many combinations of components and resources for different applications
  - Many different user requirements and scenarios, training needs

Accurate mining of sophisticated knowledge demands customisation


Improved Accuracy
Generate hypotheses
What does this do?
Who, What, When and Where?
Semantically Annotate Names, Addresses, Organisations or Proteins
Enhance searching by looking for related keywords and phrases
Choose between different meanings - ‘a dog *lead*’ or ‘a *lead* balloon’?
Words
POS
Terms
Entities
Facts
Interactions
Q&A Services
Informative Summarisation
Database Curation
Metadata Extraction
Information Extraction
Named Entity Recognition
Indicative Summarisation
Term identification
Simple keyword search ala Google™
Layers of Sophistication
General Solution
Highly Customised Solution

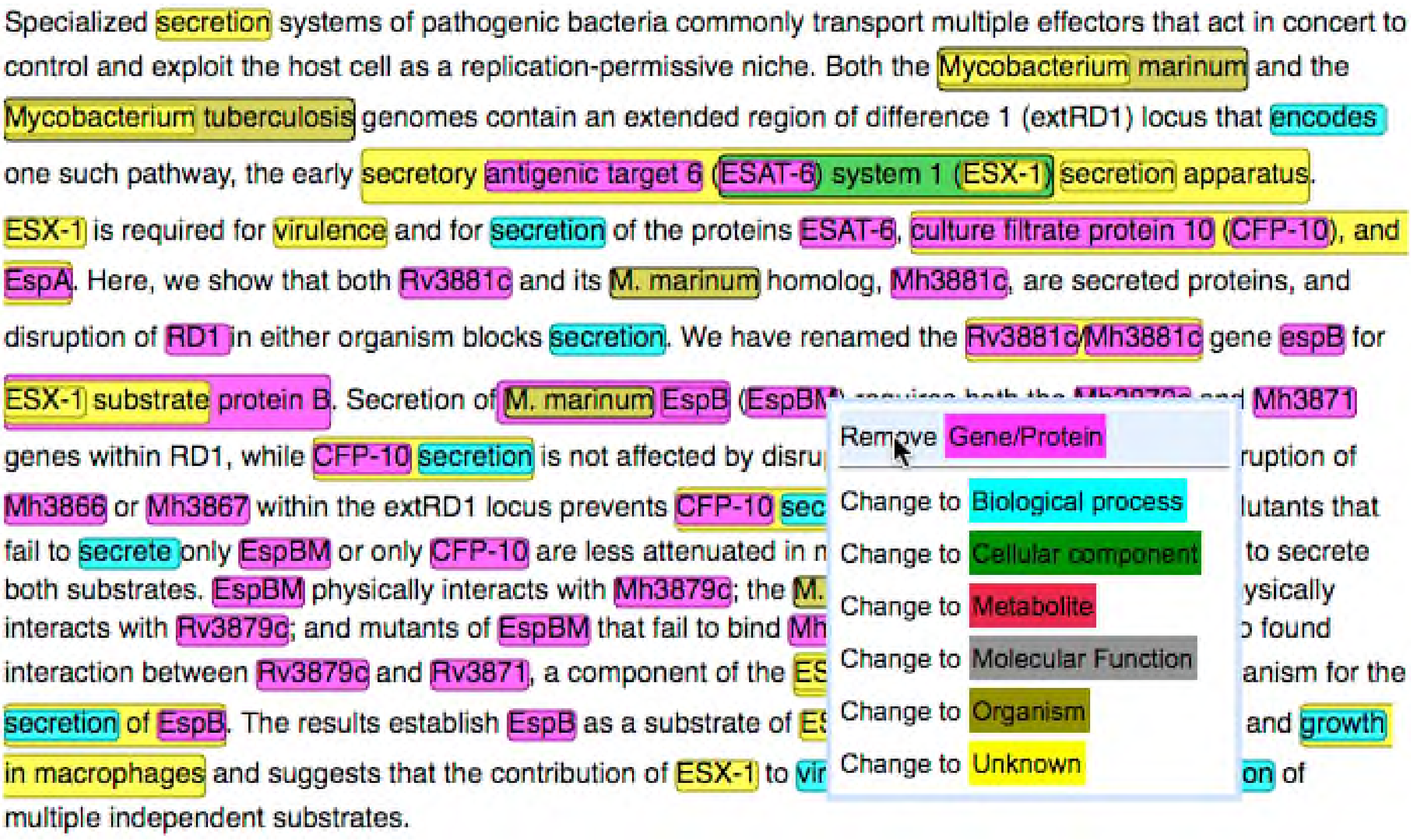
Customisation

- Rule writing
  - – Expert expresses knowledge in formalism
    - Issues of exhaustivity, specificity
- Machine learning (supervised)
  - – Expert annotates large amounts of text for desired phenomena
    - Accelerated annotation editor (e.g. NaCTeM’s Argo)
      - – System ‘completes’ annotation as it learns from your decisions
  - – Train machine learning algorithm
    - Apply trained model to find similar patterns of phenomena in new texts
- Hybrid approaches
- Knowledge poor vs. knowledge rich approaches

Brief aside

- Many of the systems and applications that follow are based on biomedicine
  - A small intellectual leap required…
- Techniques are portable to other domains and text types
  - Modulo customisation…
  - Modulo understanding user requirements
    - What’s a frame (for you, for your colleague)?

Producing multilevel annotated data: NaCTeM’s Argo editor

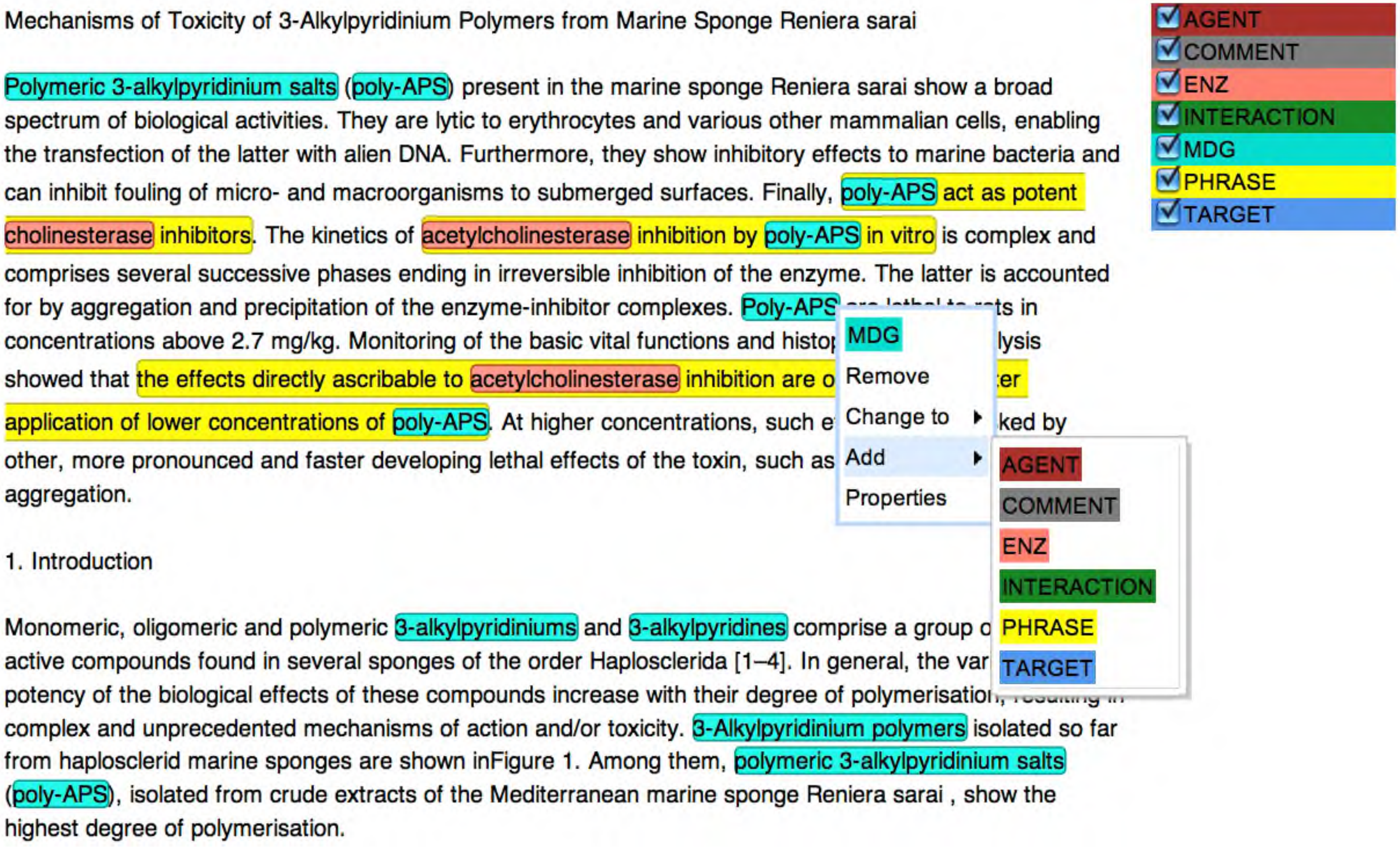


Specialized secretion systems of pathogenic bacteria commonly transport multiple effectors that act in concert to
control and exploit the host cell as a replication-permissive niche. Both the Mycobacterium marinum and the
Mycobacterium tuberculosis genomes contain an extended region of difference 1 (extRD1) locus that encodes
one such pathway, the early secretory antigenic target 6 (ESAT-6) system 1 (ESX-1) secretion apparatus.
ESX-1 is required for virulence and for secretion of the proteins ESAT-6, culture filtrate protein 10 (CFP-10), and
EspA. Here, we show that both Rv3881c and its M. marinum homolog, Mh3881c, are secreted proteins, and
disruption of RD1 in either organism blocks secretion. We have renamed the Rv3881c/Mh3881c gene espB for
ESX-1 substrate protein B. Secretion of M. marinum EspB (EspBM ... Mh3871
genes within RD1, while CFP-10 secretion is not affected by disru ... ruption of
Mh3866 or Mh3867 within the extRD1 locus prevents CFP-10 sec ... lutants that
fail to secrete only EspBM or only CFP-10 are less attenuated in m ... to secrete
both substrates. EspBM physically interacts with Mh3879c; the M. ... ysically
interacts with Rv3879c; and mutants of EspBM that fail to bind Mh ... o found
interaction between Rv3879c and Rv3871, a component of the ES ... anism for the
secretion of EspB. The results establish EspB as a substrate of E ... and growth
in macrophages and suggests that the contribution of ESX-1 to vir ... on of
multiple independent substrates.

Remove Gene/Protein
Change to Biological process
Change to Cellular component
Change to Metabolite
Change to Molecular Function
Change to Organism
Change to Unknown

Extract from McLaughlin et al. (2007) in PLoS Pathogens

Flexible annotation



Mechanisms of Toxicity of 3-Alkylpyridinium Polymers from Marine Sponge Reniera sarai

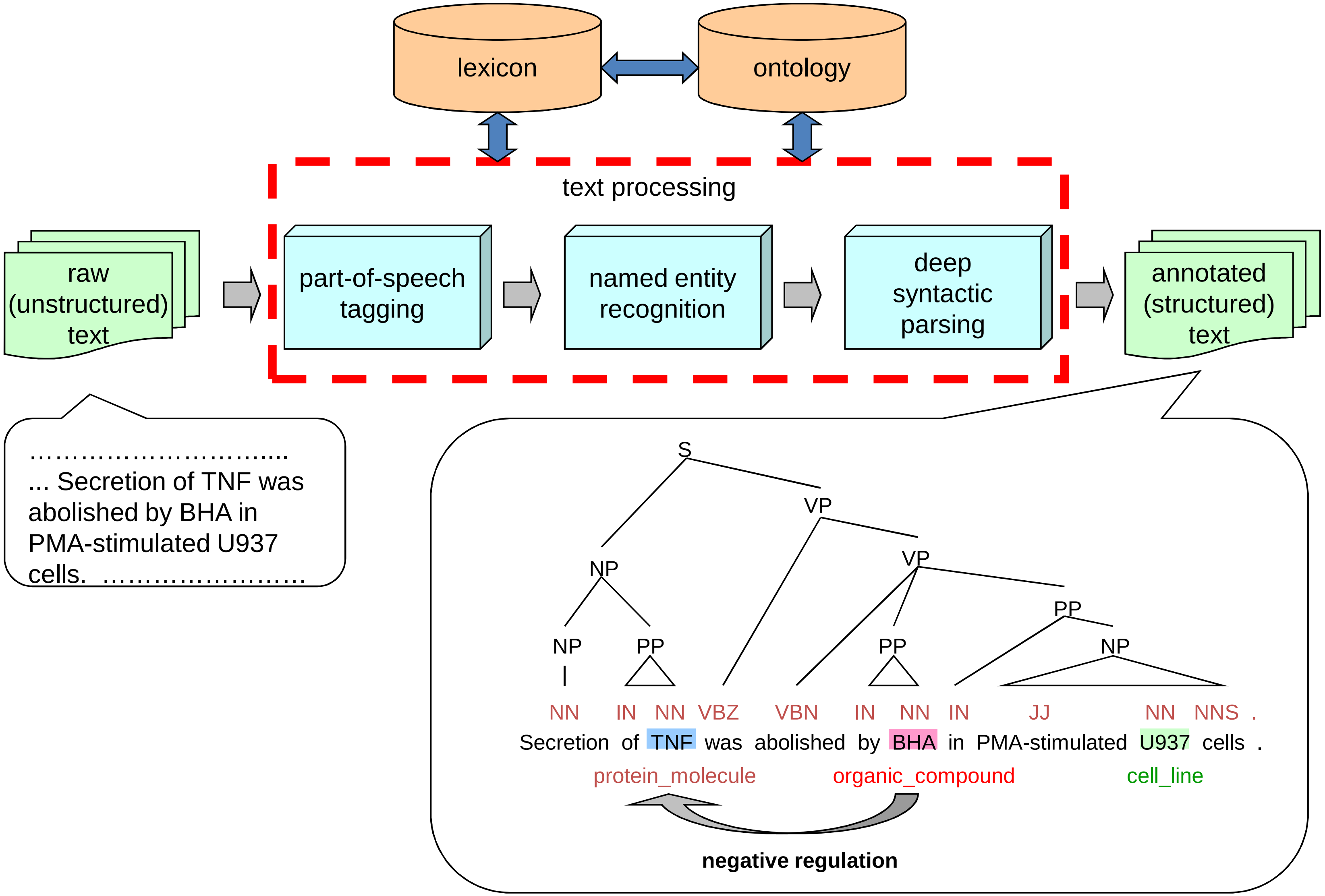
Polymeric 3-alkylpyridinium salts (poly-APS) present in the marine sponge Reniera sarai show a broad spectrum of biological activities. They are lytic to erythrocytes and various other mammalian cells, enabling the transfection of the latter with alien DNA. Furthermore, they show inhibitory effects to marine bacteria and can inhibit fouling of micro- and macroorganisms to submerged surfaces. Finally, poly-APS act as potent cholinesterase inhibitors. The kinetics of acetylcholinesterase inhibition by poly-APS in vitro is complex and comprises several successive phases ending in irreversible inhibition of the enzyme. The latter is accounted for by aggregation and precipitation of the enzyme-inhibitor complexes. Poly-APS are lethal to rats in concentrations above 2.7 mg/kg. Monitoring of the basic vital functions and histop lysis showed that the effects directly ascribable to acetylcholinesterase inhibition are o ter application of lower concentrations of poly-APS. At higher concentrations, such e ked by other, more pronounced and faster developing lethal effects of the toxin, such as aggregation.

1. Introduction

Monomeric, oligomeric and polymeric 3-alkylpyridiniums and 3-alkylpyridines comprise a group o active compounds found in several sponges of the order Haplosclerida [1–4]. In general, the var potency of the biological effects of these compounds increase with their degree of polymerisation, resulting in complex and unprecedented mechanisms of action and/or toxicity. 3-Alkylpyridinium polymers isolated so far from haplosclerid marine sponges are shown inFigure 1. Among them, polymeric 3-alkylpyridinium salts (poly-APS), isolated from crude extracts of the Mediterranean marine sponge Reniera sarai , show the highest degree of polymerisation.

Extract from Turk et al. (2007) in Marine Drugs

Information extraction: many techniques at many levels



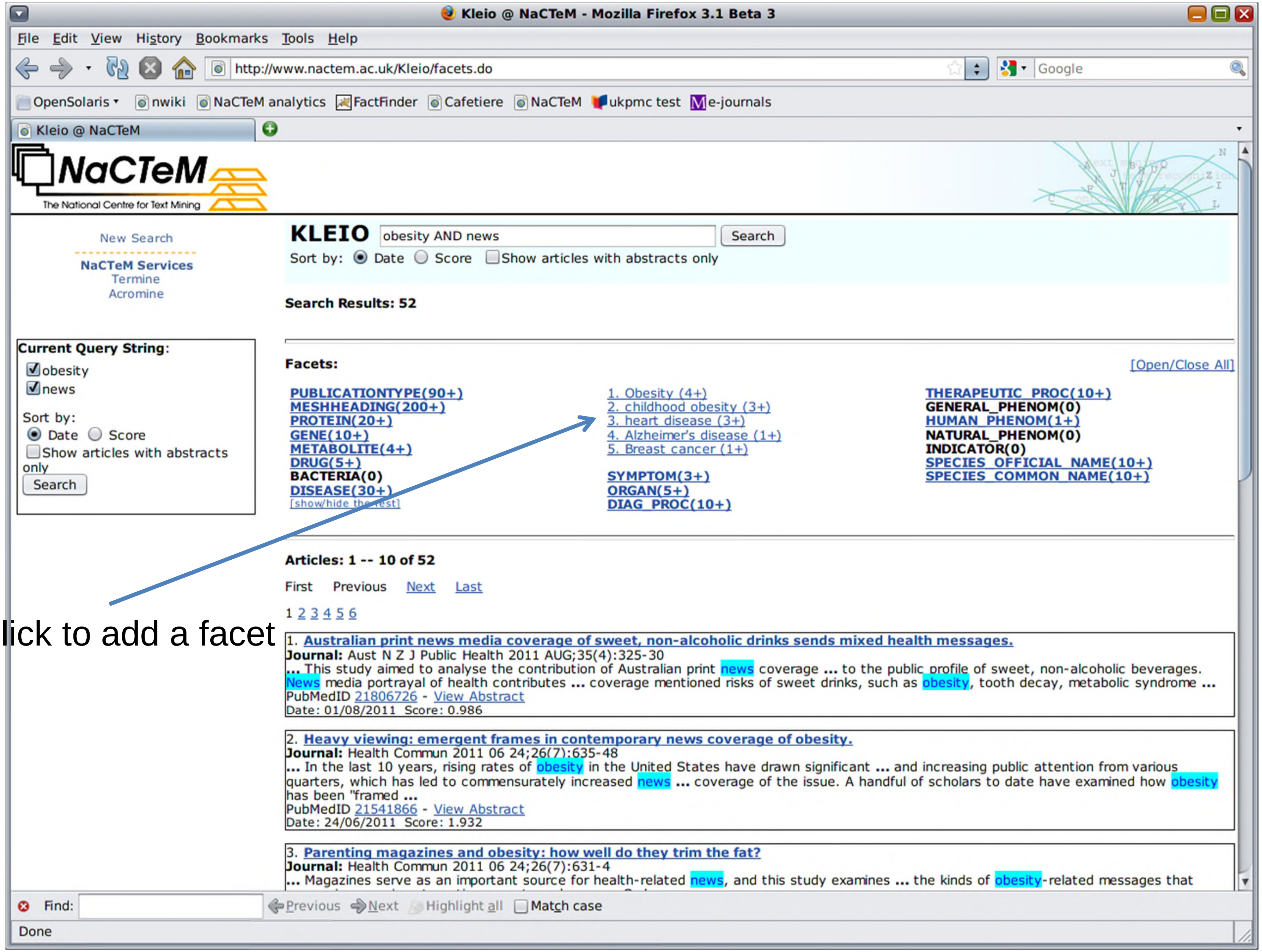
……………………….....
... Secretion of TNF was abolished by BHA in PMA-stimulated U937 cells. ……………………

Secretion of TNF was abolished by BHA in PMA-stimulated U937 cells .

Semantic search using named entity recognition

- Identify named entities in texts
  - – People, organisations, locations, genes, drugs, …
- Use entity types in searching
  - – Reduce hits to precise, relevant ones
- Make life easy by offering facetted search using the NEs
- NaCTeM’s Kleio
- www.nactem.ac.uk/software/Kleio/

General search


Kleio @ NaCTeM - Mozilla Firefox 3.1 Beta 3
File Edit View History Bookmarks Tools Help
http://www.nactem.ac.uk/Kleio/facets.do
Google
OpenSolaris nwiki NaCTeM analytics FactFinder Cafetiere NaCTeM ukpmc test e-journals
Kleio @ NaCTeM
NaCTeM
The National Centre for Text Mining
New Search
NaCTeM Services
Termine
Acromine

KLEIO obesity AND news [Search]

Sort by: ● Date ○ Score □ Show articles with abstracts only

Search Results: 52

Current Query String:

☑ obesity

☑ news

Sort by:

● Date ○ Score

□ Show articles with abstracts only

[Search]

Facets: [Open/Close All]

PUBLICATIONTYPE(90+)	1. Obesity (4+)	THERAPEUTIC_PROC(10+)
MESHHEADING(200+)	2. childhood obesity (3+)	GENERAL_PHENOM(0)
PROTEIN(20+)	3. heart disease (3+)	HUMAN_PHENOM(1+)
GENE(10+)	4. Alzheimer's disease (1+)	NATURAL_PHENOM(0)
METABOLITE(4+)	5. Breast cancer (1+)	INDICATOR(0)
DRUG(5+)		SPECIES_OFFICIAL_NAME(10+)
BACTERIA(0)	SYMPTOM(3+)	SPECIES_COMMON_NAME(10+)
DISEASE(30+)	ORGAN(5+)	
[show/hide the rest]	DIAG_PROC(10+)	

Articles: 1 -- 10 of 52

First Previous Next Last

1 2 3 4 5 6

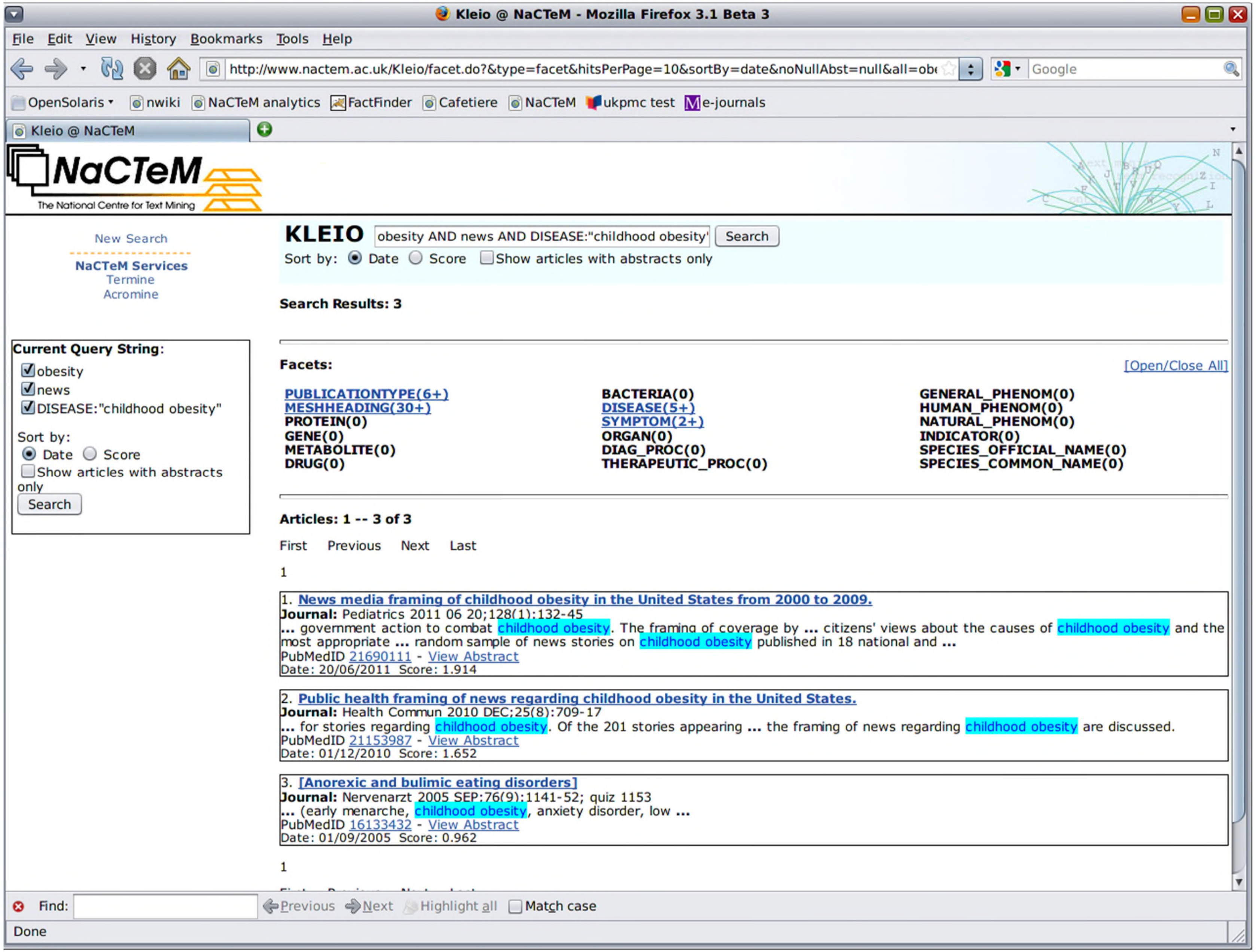
1. **Australian print news media coverage of sweet, non-alcoholic drinks sends mixed health messages.**
Journal: Aust N Z J Public Health 2011 AUG;35(4):325-30
... This study aimed to analyse the content of Australian print news coverage **...** to the public profile of sweet, non-alcoholic beverages. News media portrayal of health contributes **...** coverage mentioned risks of sweet drinks, such as obesity, tooth decay, metabolic syndrome **...**
PubMedID 21806726 - View Abstract
Date: 01/08/2011 Score: 0.986

2. **Heavy viewing: emergent frames in contemporary news coverage of obesity.**
Journal: Health Commun 2011 06 24;26(7):635-48
... In the last 10 years, rising rates of obesity in the United States have drawn significant **...** and increasing public attention from various quarters, which has led to commensurately increased news **...** coverage of the issue. A handful of scholars to date have examined how obesity has been "framed **...**
PubMedID 21541866 - View Abstract
Date: 24/06/2011 Score: 1.932

3. **Parenting magazines and obesity: how well do they trim the fat?**
Journal: Health Commun 2011 06 24;26(7):631-4
... Magazines serve as an important source for health-related news, and this study examines **...** the kinds of obesity-related messages that

Click to add a facet

Add a mined NE facet


Kleio @ NaCTeM - Mozilla Firefox 3.1 Beta 3
File Edit View History Bookmarks Tools Help
Google
OpenSolaris nwiki NaCTeM analytics FactFinder Cafetiere NaCTeM ukpmc test e-journals
Kleio @ NaCTeM
NaCTeM
The National Centre for Text Mining
New Search
NaCTeM Services
Termine
Acromine
Find:
Previous
Next
Highlight all
Match case
Done

http://www.nactem.ac.uk/Kleio/facet.do?&type=facet&hitsPerPage=10&sortBy=date&noNullAbst=null&all=obe

KLEIO obesity AND news AND DISEASE:"childhood obesity" Search

Sort by: ● Date ○ Score □ Show articles with abstracts only

Search Results: 3

Current Query String:

☑ obesity

☑ news

☑ DISEASE:"childhood obesity"

Sort by: ● Date ○ Score

□ Show articles with abstracts only

Search

Facets: [Open/Close All]

PUBLICATIONTYPE(6+)	BACTERIA(0)	GENERAL_PHENOM(0)
MESHHEADING(30+)	DISEASE(5+)	HUMAN_PHENOM(0)
PROTEIN(0)	SYMPTOM(2+)	NATURAL_PHENOM(0)
GENE(0)	ORGAN(0)	INDICATOR(0)
METABOLITE(0)	DIAG_PROC(0)	SPECIES_OFFICIAL_NAME(0)
DRUG(0)	THERAPEUTIC_PROC(0)	SPECIES_COMMON_NAME(0)

Articles: 1 -- 3 of 3

First Previous Next Last

1

1. **News media framing of childhood obesity in the United States from 2000 to 2009.**
Journal: Pediatrics 2011 06 20;128(1):132-45
... government action to combat childhood obesity. The framing of coverage by **...** citizens' views about the causes of childhood obesity and the most appropriate **...** random sample of news stories on childhood obesity published in 18 national and **...**
PubMedID 21690111 - View Abstract
Date: 20/06/2011 Score: 1.914

2. **Public health framing of news regarding childhood obesity in the United States.**
Journal: Health Commun 2010 DEC;25(8):709-17
... for stories regarding childhood obesity. Of the 201 stories appearing **...** the framing of news regarding childhood obesity are discussed.
PubMedID 21153987 - View Abstract
Date: 01/12/2010 Score: 1.652

3. **[Anorexic and bulimic eating disorders]**
Journal: Nervenarzt 2005 SEP;76(9):1141-52; quiz 1153
... (early menarche, childhood obesity, anxiety disorder, low **...**
PubMedID 16133432 - View Abstract
Date: 01/09/2005 Score: 0.962

1

Noisy data

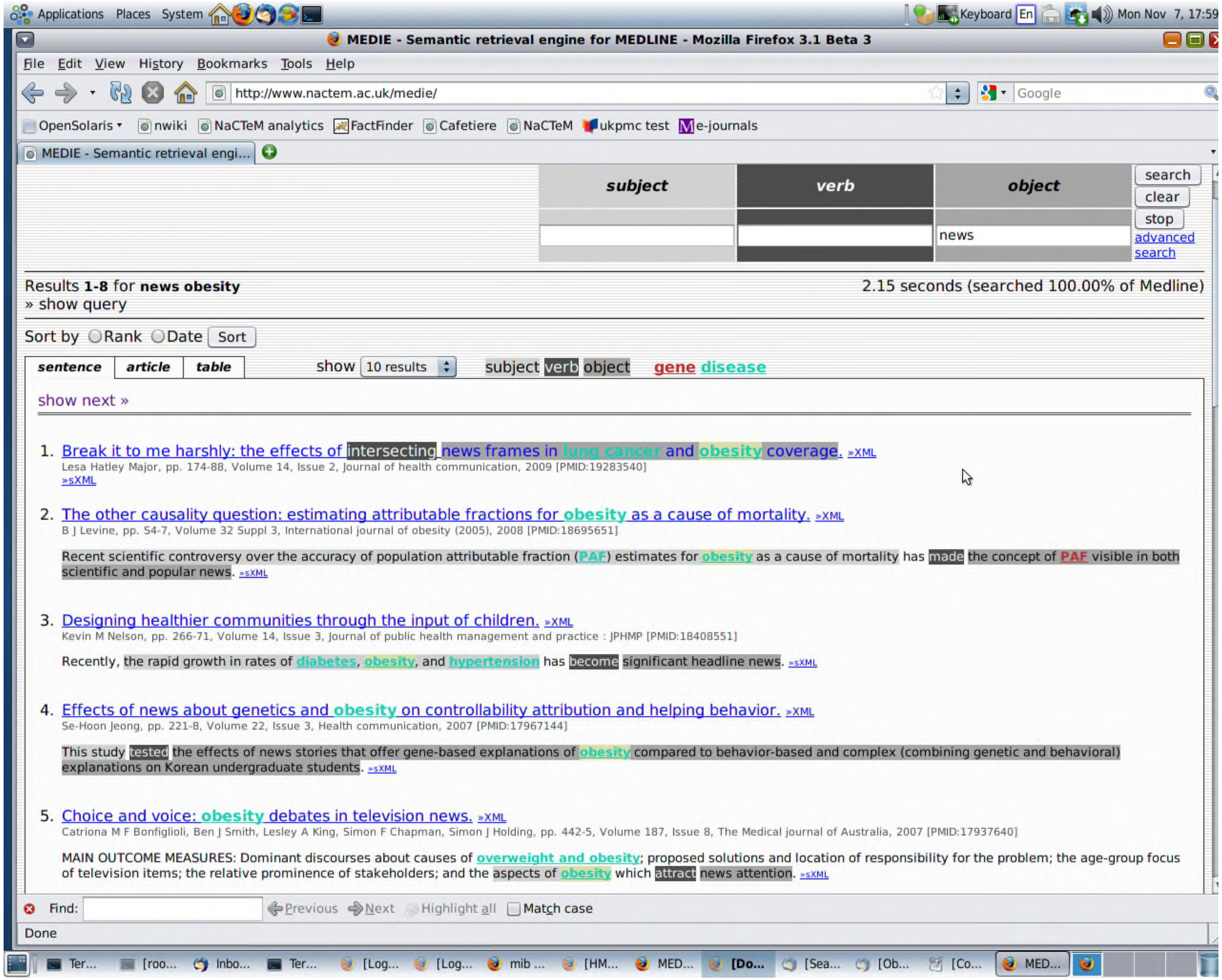
- Full-scale analysis techniques often not suitable for noisy data
  - – OCR, digitised texts
  - – Emails, unprofessional blogs, tweets, …
- May have to use partial analysis techniques, often based on machine learning
- Even apparently ‘clean’ data can be hard to process
  - “Hamburger PDFs” (P. Murray-Rust)

Full-scale analysis

- For formally-written text, full-scale analysis possible
- Deep parsing with MEDIE
  - Allows semantic search
  - Query consists of partial or fully specified fact/event template
  - www.nactem.ac.uk/medie/
  - Developed by Univ. Tokyo (close collaboration)

MEDIE: semantic search of pre-analysed collection

Sentences from MEDLINE where news is the object, with obesity specified in (here unseen) advanced options


Applications Places System
Keyboard En
Mon Nov 7, 17:59
MEDIE - Semantic retrieval engine for MEDLINE - Mozilla Firefox 3.1 Beta 3
File Edit View History Bookmarks Tools Help
http://www.nactem.ac.uk/medie/
Google
OpenSolaris nwiki NaCTeM analytics FactFinder Cafetiere NaCTeM ukpmc test e-journals
MEDIE - Semantic retrieval engi...
subject
verb
object
search
clear
stop
advanced search

object: news

Results **1-8** for **news obesity**

» show query

2.15 seconds (searched 100.00% of Medline)

Sort by ○Rank ○Date Sort

sentence article table
show 10 results
subject verb object gene disease

show next »

1. Break it to me harshly: the effects of intersecting news frames in lung cancer and obesity coverage. »XML
   Lesa Hatley Major, pp. 174-88, Volume 14, Issue 2, Journal of health communication, 2009 [PMID:19283540]
   »sXML

2. The other causality question: estimating attributable fractions for obesity as a cause of mortality. »XML
   B J Levine, pp. S4-7, Volume 32 Suppl 3, International journal of obesity (2005), 2008 [PMID:18695651]

   Recent scientific controversy over the accuracy of population attributable fraction (PAF) estimates for obesity as a cause of mortality has made the concept of PAF visible in both scientific and popular news. »sXML

3. Designing healthier communities through the input of children. »XML
   Kevin M Nelson, pp. 266-71, Volume 14, Issue 3, Journal of public health management and practice : JPHMP [PMID:18408551]

   Recently, the rapid growth in rates of diabetes, obesity, and hypertension has become significant headline news. »sXML

4. Effects of news about genetics and obesity on controllability attribution and helping behavior. »XML
   Se-Hoon Jeong, pp. 221-8, Volume 22, Issue 3, Health communication, 2007 [PMID:17967144]

   This study tested the effects of news stories that offer gene-based explanations of obesity compared to behavior-based and complex (combining genetic and behavioral) explanations on Korean undergraduate students. »sXML

5. Choice and voice: obesity debates in television news. »XML
   Catriona M F Bonfiglioli, Ben J Smith, Lesley A King, Simon F Chapman, Simon J Holding, pp. 442-5, Volume 187, Issue 8, The Medical journal of Australia, 2007 [PMID:17937640]

   MAIN OUTCOME MEASURES: Dominant discourses about causes of overweight and obesity; proposed solutions and location of responsibility for the problem; the age-group focus of television items; the relative prominence of stakeholders; and the aspects of obesity which attract news attention. »sXML

Find: Previous Next Highlight all □ Match case
Done
Ter... [roo... Inbo... Ter... [Log... [Log... mib ... [HM... MED... [Do... [Sea... [Ob... [Co... MED...

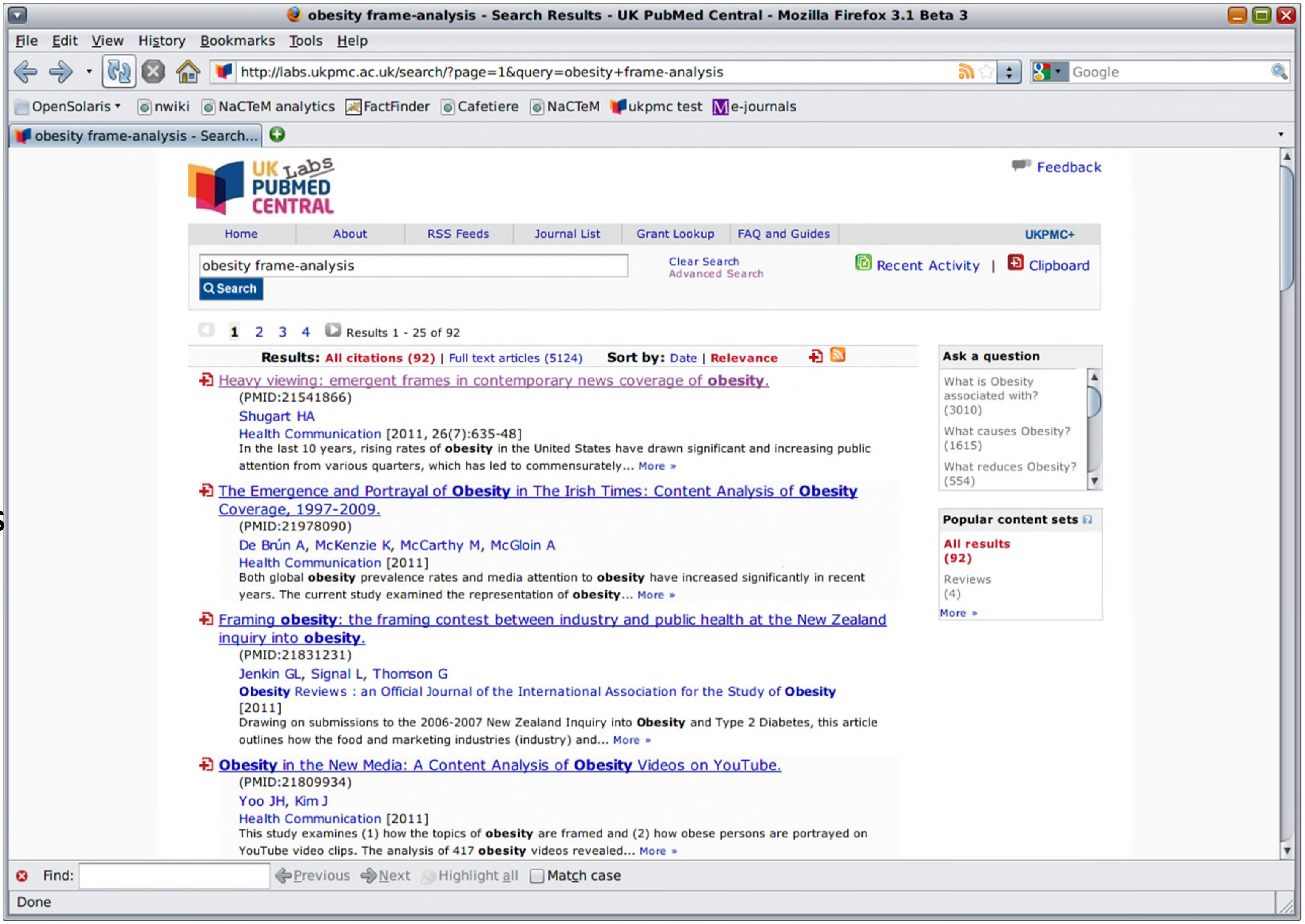
What users want…

- Many users just give a one-word query when searching
  - Very few explore ‘advanced search’
- So NaCTeM uses deep parsing results to *generate questions* relevant to a query that are *known to have answers*
  - UKPMC EvidenceFinder
    - http://labs.ukpmc.ac.uk
  - UKPMC: archive of full texts in life sciences
    - EBI, British Library, Univ of Manchester (Mimas and NaCTeM)

UKPMC EvidenceFinder

Semantic search based on analysis of full text articles

But currently, EF oriented more towards Biology… and can only text mine OA subset


obesity frame-analysis - Search Results - UK PubMed Central - Mozilla Firefox 3.1 Beta 3
File Edit View History Bookmarks Tools Help
http://labs.ukpmc.ac.uk/search/?page=1&query=obesity+frame-analysis
Google
OpenSolaris nwiki NaCTeM analytics FactFinder Cafetiere NaCTeM ukpmc test e-journals
obesity frame-analysis - Search...
UK Labs PUBMED CENTRAL
Feedback
Home About RSS Feeds Journal List Grant Lookup FAQ and Guides UKPMC+
obesity frame-analysis
Search
Clear Search
Advanced Search
Recent Activity | Clipboard
1 2 3 4
Find: Previous Next Highlight all Match case
Done

Results 1 - 25 of 92

Results: All citations (92) | Full text articles (5124) **Sort by:** Date | Relevance

Heavy viewing: emergent frames in contemporary news coverage of **obesity**.
(PMID:21541866)
Shugart HA
Health Communication [2011, 26(7):635-48]
In the last 10 years, rising rates of **obesity** in the United States have drawn significant and increasing public attention from various quarters, which has led to commensurately... More »

The Emergence and Portrayal of **Obesity** in The Irish Times: Content Analysis of **Obesity** Coverage, 1997-2009.
(PMID:21978090)
De Brún A, McKenzie K, McCarthy M, McGloin A
Health Communication [2011]
Both global **obesity** prevalence rates and media attention to **obesity** have increased significantly in recent years. The current study examined the representation of **obesity**... More »

Framing **obesity**: the framing contest between industry and public health at the New Zealand inquiry into **obesity**.
(PMID:21831231)
Jenkin GL, Signal L, Thomson G
Obesity Reviews : an Official Journal of the International Association for the Study of **Obesity** [2011]
Drawing on submissions to the 2006-2007 New Zealand Inquiry into **Obesity** and Type 2 Diabetes, this article outlines how the food and marketing industries (industry) and... More »

Obesity in the New Media: A Content Analysis of **Obesity** Videos on YouTube.
(PMID:21809934)
Yoo JH, Kim J
Health Communication [2011]
This study examines (1) how the topics of **obesity** are framed and (2) how obese persons are portrayed on YouTube video clips. The analysis of 417 **obesity** videos revealed... More »

Ask a question
What is Obesity associated with? (3010)
What causes Obesity? (1615)
What reduces Obesity? (554)

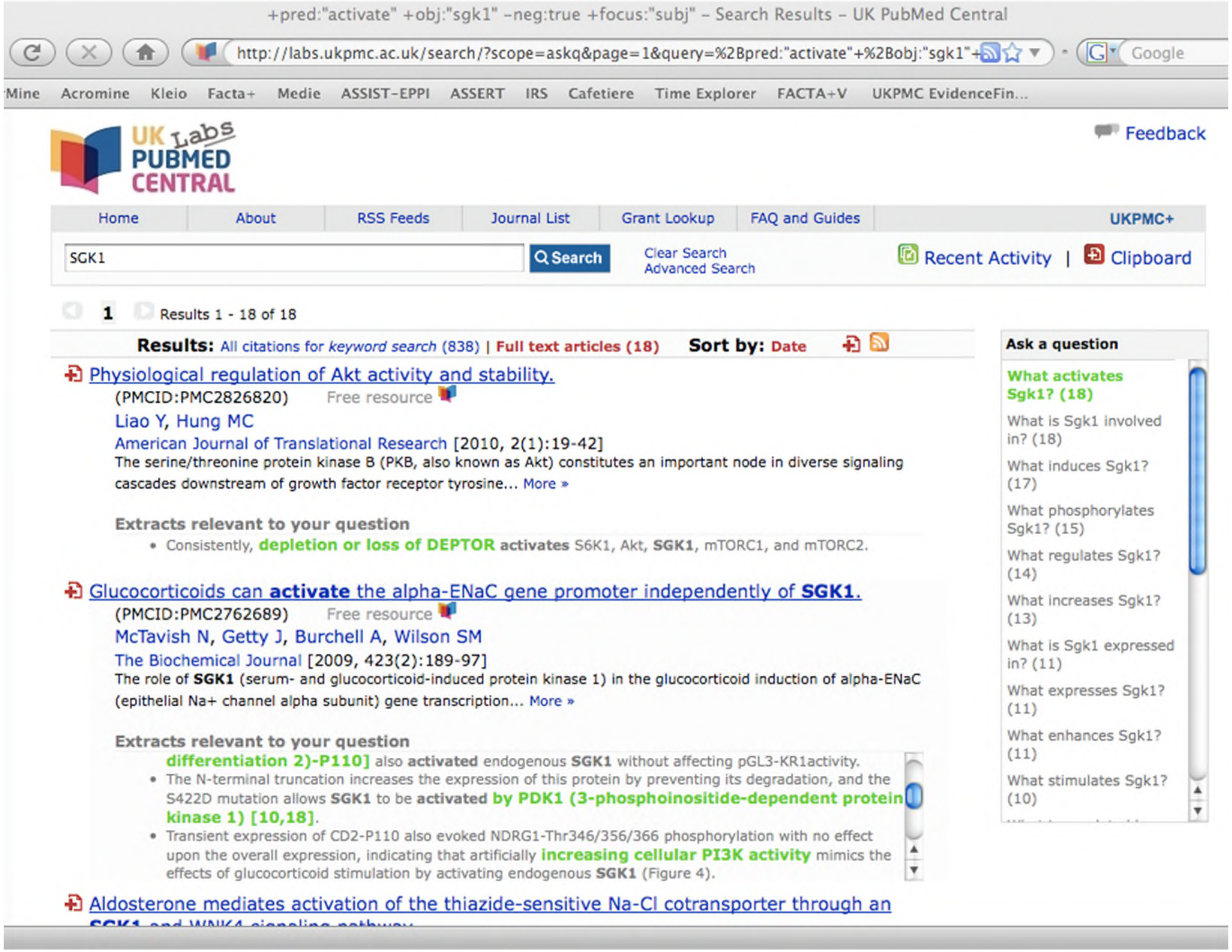
Popular content sets
All results (92)
Reviews (4)
More »

UKPMC EvidenceFinder

But asking about e.g. a protein gives better idea of functionality…

330 thousand OA full text articles analysed.

Generated questions known to have answers.



+pred:"activate" +obj:"sgk1" –neg:true +focus:"subj" – Search Results – UK PubMed Central

http://labs.ukpmc.ac.uk/search/?scope=askq&page=1&query=%2Bpred:"activate"+%2Bobj:"sgk1"+

1 Results 1 - 18 of 18

Results: All citations for *keyword search* (838) | **Full text articles (18)** **Sort by:** Date

Physiological regulation of Akt activity and stability.
(PMCID:PMC2826820) Free resource
Liao Y, Hung MC
American Journal of Translational Research [2010, 2(1):19-42]
The serine/threonine protein kinase B (PKB, also known as Akt) constitutes an important node in diverse signaling cascades downstream of growth factor receptor tyrosine... More »

Extracts relevant to your question

- Consistently, **depletion or loss of DEPTOR activates** S6K1, Akt, **SGK1**, mTORC1, and mTORC2.

Glucocorticoids can **activate** the alpha-ENaC gene promoter independently of **SGK1**.
(PMCID:PMC2762689) Free resource
McTavish N, Getty J, Burchell A, Wilson SM
The Biochemical Journal [2009, 423(2):189-97]
The role of **SGK1** (serum- and glucocorticoid-induced protein kinase 1) in the glucocorticoid induction of alpha-ENaC (epithelial Na+ channel alpha subunit) gene transcription... More »

Extracts relevant to your question

differentiation 2)-P110] also **activated** endogenous **SGK1** without affecting pGL3-KR1activity.

- The N-terminal truncation increases the expression of this protein by preventing its degradation, and the S422D mutation allows **SGK1** to be **activated by PDK1 (3-phosphoinositide-dependent protein kinase 1) [10,18]**.
- Transient expression of CD2-P110 also evoked NDRG1-Thr346/356/366 phosphorylation with no effect upon the overall expression, indicating that artificially **increasing cellular PI3K activity** mimics the effects of glucocorticoid stimulation by activating endogenous **SGK1** (Figure 4).

Aldosterone mediates activation of the thiazide-sensitive Na-Cl cotransporter through an SGK1 and WNK4 signaling pathway.

Ask a question

What activates Sgk1? (18)
What is Sgk1 involved in? (18)
What induces Sgk1? (17)
What phosphorylates Sgk1? (15)
What regulates Sgk1? (14)
What increases Sgk1? (13)
What is Sgk1 expressed in? (11)
What expresses Sgk1? (11)
What enhances Sgk1? (11)
What stimulates Sgk1? (10)

Finding associations

- Known associations (but not to you)
  - Unknown knowns (the one DR missed out)
- Unknown associations (to anyone)
  - Statistics of surprise
  - A related to B, B to C, so infer A related to C
- www.nactem.ac.uk/services.php
- Check out FACTA+ and FACTA+ visualizer

Marrying IR with TM

- Classifying
- Clustering
- Summarisation
- Term extraction
- Related documents based on analysis of returned results
- More detail: talk by James Thomas
- Recommendation systems

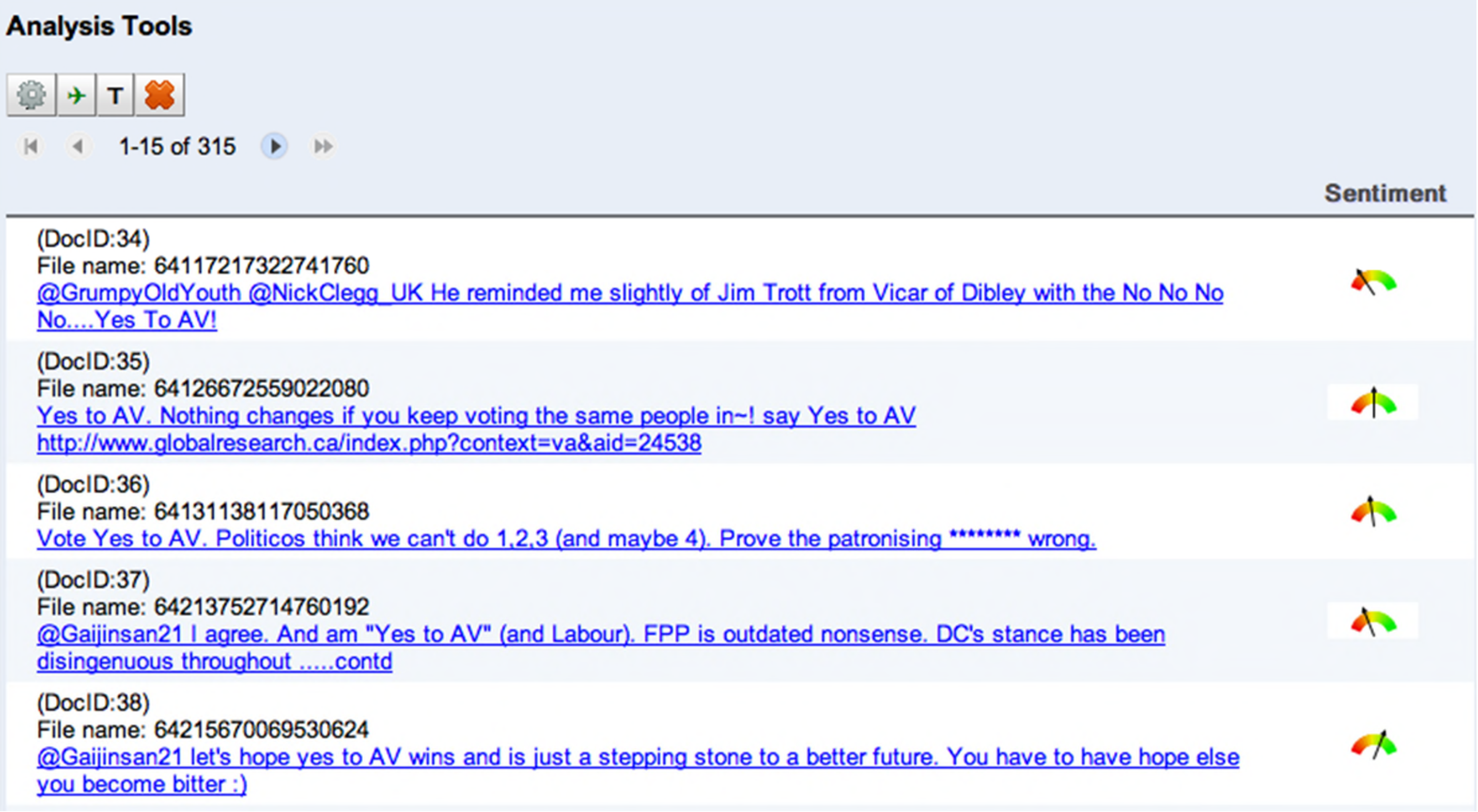
Opinion, sentiment mining

- Opinion polling
  - – Ratios for/against a proposition
- What ± passions are aroused about
  - – A topic?
  - – A social group?
- What are the terms in which the debate is conducted?
- How are media used in the formation and dissemination of opinion?

Case study: Alternative vote debate

- Named entity recognition
  - Finding mentions of political parties, politicians
  - Including alternative designations
- Term and topic discovery
  - Most significant terms used in a (sub) collection of text
- Subjectivity analysis
  - Positive or negative orientation, whether subjective or objective style
- Speech act and rhetorical analysis
  - Whether initiating, reacting, supporting, arguing, citing evidence, etc.
- Semantic search engines based on all the above

NaCTeM analysis of tweets on AV: overall (document-level) sentiment



Analysis Tools

1-15 of 315

Sentiment

(DocID:34)
File name: 64117217322741760
@GrumpyOldYouth @NickClegg_UK He reminded me slightly of Jim Trott from Vicar of Dibley with the No No No No....Yes To AV!

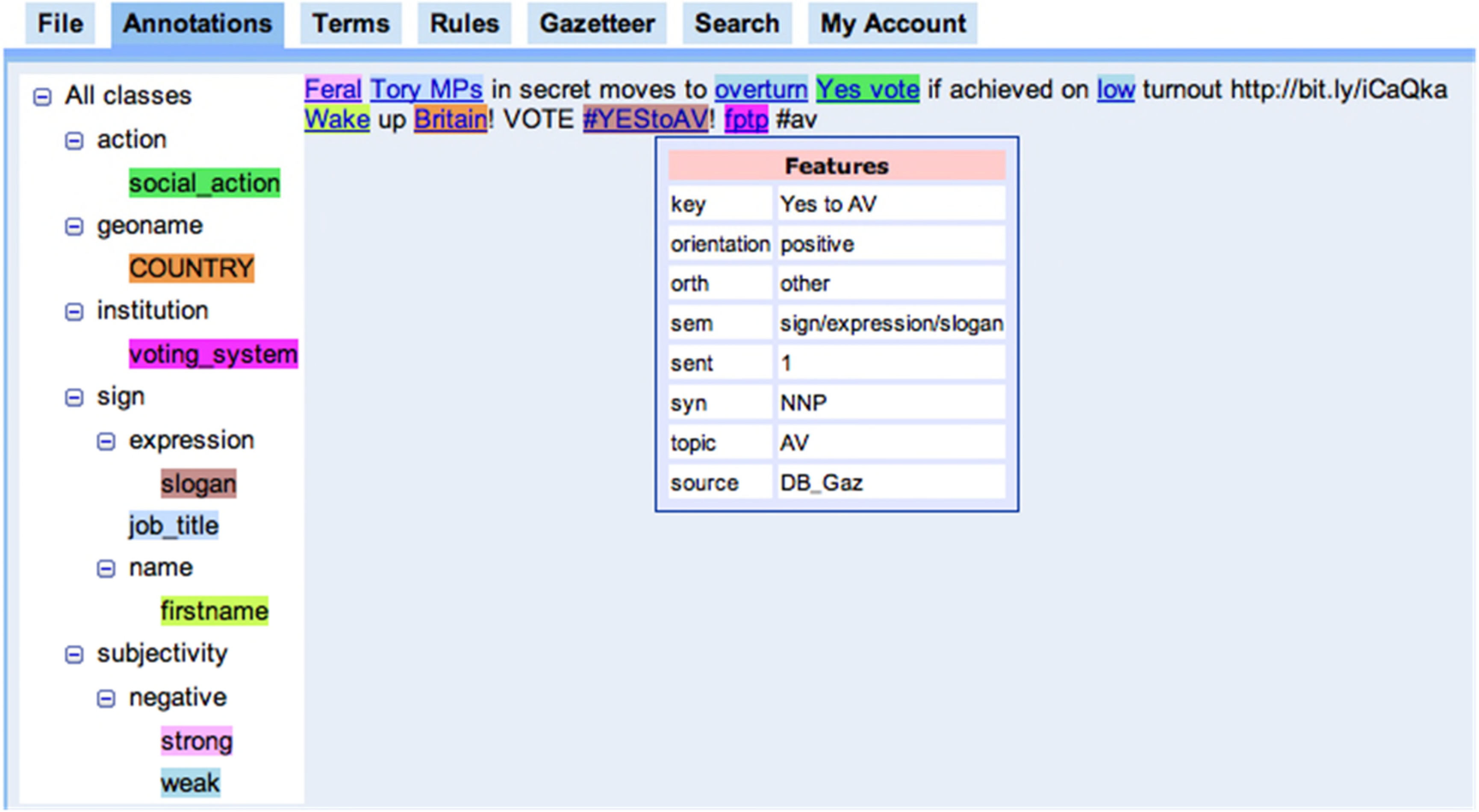
(DocID:35)
File name: 64126672559022080
Yes to AV. Nothing changes if you keep voting the same people in~! say Yes to AV http://www.globalresearch.ca/index.php?context=va&aid=24538

(DocID:36)
File name: 64131138117050368
Vote Yes to AV. Politicos think we can't do 1,2,3 (and maybe 4). Prove the patronising ******** wrong.

(DocID:37)
File name: 64213752714760192
@Gaijinsan21 I agree. And am "Yes to AV" (and Labour). FPP is outdated nonsense. DC's stance has been disingenuous throughout .....contd

(DocID:38)
File name: 64215670069530624
@Gaijinsan21 let's hope yes to AV wins and is just a stepping stone to a better future. You have to have hope else you become bitter :)

Sentiment analysis of 1 document



Feral Tory MPs in secret moves to overturn Yes vote if achieved on low turnout http://bit.ly/iCaQka Wake up Britain! VOTE #YEStoAV! fptp #av

Features	
key	Yes to AV
orientation	positive
orth	other
sem	sign/expression/slogan
sent	1
syn	NNP
topic	AV
source	DB_Gaz

Sentiment: challenges

- Sentiment, subjectivity, emotion
  - Polarity of subjective expression can be confused with being *pro* or *anti* a given proposition (“I support AV” is not subjective)
- Irony
- Rhetoric (last element often overturns predecessors)
- Quality and cultural specificity of resources (e.g. sentiment lexicon)

Opinion/sentiment mining tutorial

http://www.cs.uic.edu/~liub/FBS/Sentiment-Analysis-tutorial-AAAI-2011.pdf

- Recent comprehensive tutorial by Bing Liu at AAAI 2011, August.

Conclusion

- TM has many aspects, components, levels
  - Different combinations for different tasks
  - Can be applied in any domain (customisation)
- No “one size fits all”
- Domain experts and developers must engage to ensure appropriate application for needs
- Copyright & licensing: support adoption of Hargreaves recommendations!

Acknowledgements

- Text mining team at NaCTeM (~19)
- 東京大学 THE UNIVERSITY OF TOKYO
- Funders and sponsors

JISC

bbsrc
biotechnology and biological sciences research council

EPSRC
Engineering and Physical Sciences Research Council

NATIONAL INSTITUTES OF HEALTH

AstraZeneca
UNITED KINGDOM

Pfizer Working together for a healthier world™

npg nature publishing group

Wellcome Trust + 17 other sponsors (UKPMC)

Elsevier

NHS

Information Society Technologies

MANCHESTER 1824
The University of Manchester

NaCTeM
The National Centre for Text Mining