

An overview of text mining tools and services

at the National Centre for Text Mining

**John McNaught
Deputy Director**

www.nactem.ac.uk

Outline

- Overview of NaCTeM
 - Why, What, ...
 - Role in e-Infrastructure
- Quick tour of NaCTeM services/tools
- Challenges

UK National Centre for Text Mining (NaCTeM)

- 1st national text mining centre in the world www.nactem.ac.uk
- **Location:** Manchester Interdisciplinary Biocentre (MIB)
- **Remit:** Provision of text mining services to support UK research
- **Funded by:** the JISC, BBSRC, EPSRC
- **Initial Focus:** Biology
- **Now:** Social Sciences, Medicine, ...

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**
- UK **policy** to 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

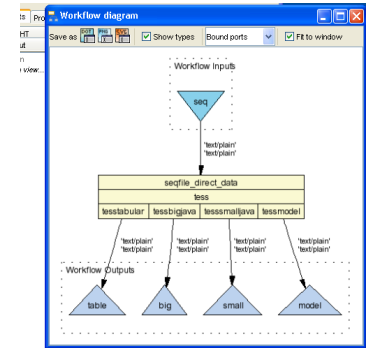
e-Research Infrastructure

- Access to information, data resources, distributed computing essential for bio-scientists
- e-Infrastructure provides services and facilities *enabling* advanced research
- Deploying e-Infrastructure increases the pace and efficiency of new research methods and techniques

E-Infrastructure and Text Mining: for whom?

- End-users
- Software engineers
- Generic tool developers
- Service and resource providers

- Workflow developers
 - Text miners
- Swiss-Prot, Nature



What users want to do with their data (minimally)

- Easier access to data
- Share data with their peers
- Annotate data with metadata
- Manage data across locations
- Integrate data within workflows, Web Services
- Aids for semantic metadata creation; enriching data with related metadata e.g. experimental results

**TEXT MINING CAN
SUPPORT USERS**

Science is data-driven

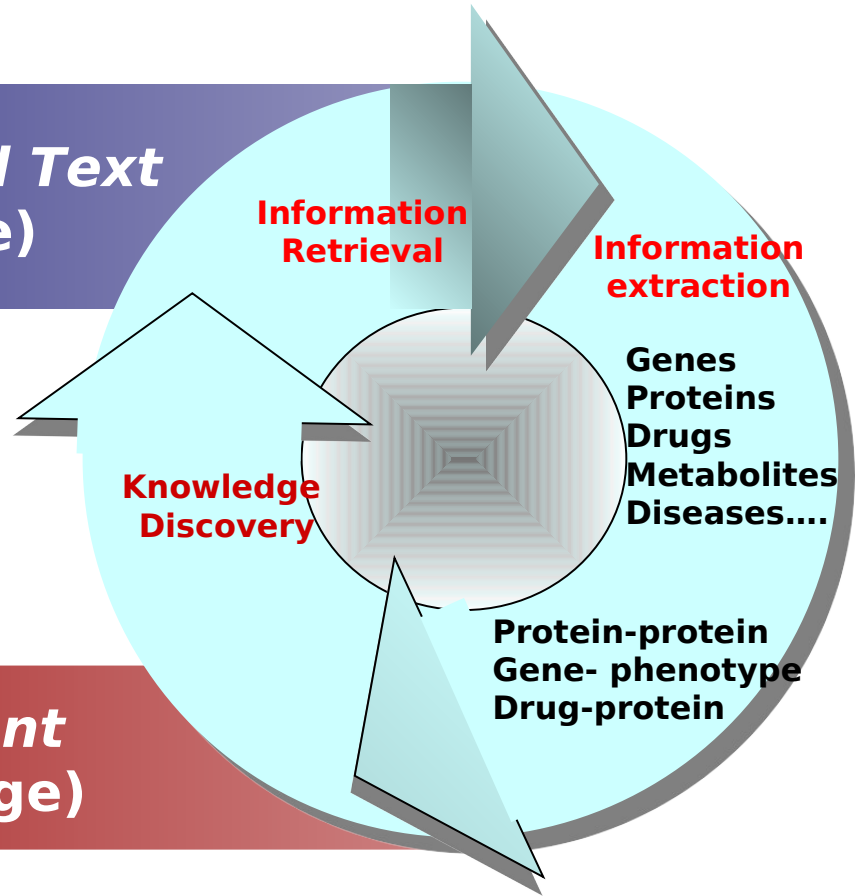
“the current scientific literature, were it to be presented in semantically accessible form, contains huge amounts of undiscovered science”

Peter Murray-Rust, *Data-driven science*

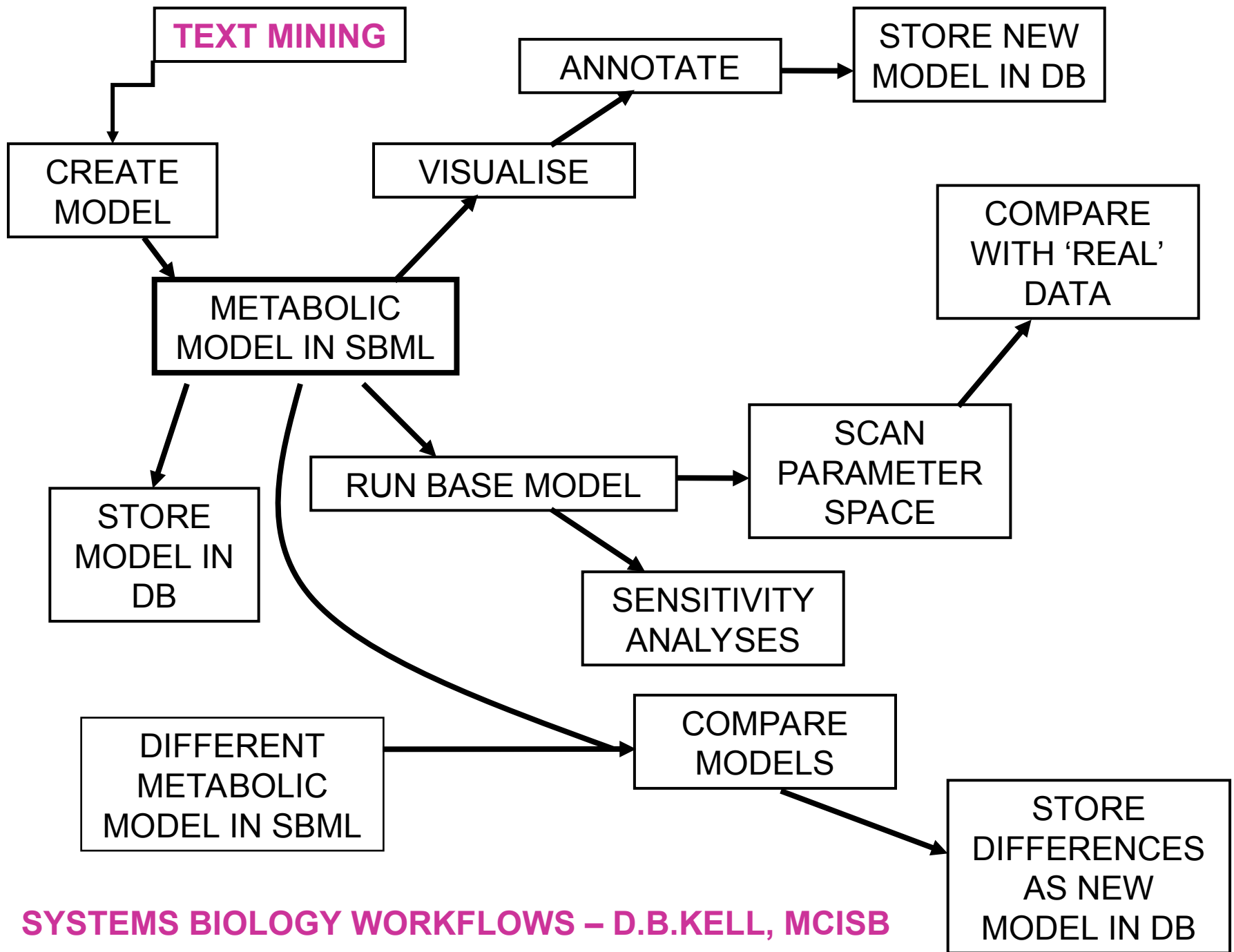
From text to knowledge



**Unstructured Text
(implicit knowledge)**



**Structured content
(explicit knowledge)**



What do we provide and build?

- **Resources:** lexica, terminologies, thesauri, grammars, annotated corpora
- **Tools:** tokenisers, taggers, chunkers, parsers, NE recognisers, semantic analysers
- **Services**
 - Proof of concept: evolving to large scale Grid-enabled services
 - Service provision through  *Mimas*
- **Customised solutions**

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
- Need to be active in all areas to effectively support researchers

Development strategy

- **Re-use** tools where possible
 - Forge strategic relationships (UTokyo, IBM)
- Use **integration** framework (UIMA)
- Develop **generic** TM tools
- **Customise** for specific domains/scenarios (pharmas, repository search, systematic reviews)
 - While actively engaging with user communities (requirements, evaluation)
- Encapsulate in **services**

How do we provide services?

Modes of use

- **Demonstrators:** for small-scale online use
- **Batch mode:** upload data, get email with link to download site when job done
- **Web Services**
- Embedding text mining Web Services into **Workflows**

- Some services are compositions of tools
- Individual tools to process user data

Services based on pre-processed collections

- Pre-process (analyse and index) **popular collections**, e.g.
 - MEDLINE
 - Intute repository
 - Evolving to handle full text (UKPMC)
- Provide **advanced search interfaces** to these
 - Based on user scenarios
- Rapid results for end-user
- Regular up-dating of analyses carried out

NaCTeM services and tools

- TerMine: extract candidate terms
- AcroMine: acronyms ↔ fullforms
- TM for IRS: repository search
- ASSIST: search, browse and cluster
- KLEIO: semantic search, concepts
- FACTA: semantic search, associations
- MEDIE: semantic search over facts
- InfoPubMed: protein-protein interactions

TerMine (C-value) analysis

Found **1163** terms in 2.6 seconds - all terms ([in table](#)) ([in text](#)) - threshold:

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 rapid control of receptor sensitivity in repeatedly stimulated cells such as neurons.

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 high degree of genetic heterogeneity in the human population and the distributions of single-nucleotide polymorphisms (SNPs) at amino acid positions 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, 700, 701, 702, 703, 704, 705, 706, 707, 708, 709, 710, 711, 712, 713, 714, 715, 716, 717, 718, 719, 720, 721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 731, 732, 733, 734, 735, 736, 737, 738, 739, 740, 741, 742, 743, 744, 745, 746, 747, 748, 749, 750, 751, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 798, 799, 800, 801, 802, 803, 804, 805, 806, 807, 808, 809, 810, 811, 812, 813, 814, 815, 816, 817, 818, 819, 820, 821, 822, 823, 824, 825, 826, 827, 828, 829, 830, 831, 832, 833, 834, 835, 836, 837, 838, 839, 840, 841, 842, 843, 844, 845, 846, 847, 848, 849, 850, 851, 852, 853, 854, 855, 856, 857, 858, 859, 860, 861, 862, 863, 864, 865, 866, 867, 868, 869, 870, 871, 872, 873, 874, 875, 876, 877, 878, 879, 880, 881, 882, 883, 884, 885, 886, 887, 888, 889, 890, 891, 892, 893, 894, 895, 896, 897, 898, 899, 900, 901, 902, 903, 904, 905, 906, 907, 908, 909, 910, 911, 912, 913, 914, 915, 916, 917, 918, 919, 920, 921, 922, 923, 924, 925, 926, 927, 928, 929, 930, 931, 932, 933, 934, 935, 936, 937, 938, 939, 940, 941, 942, 943, 944, 945, 946, 947, 948, 949, 950, 951, 952, 953, 954, 955, 956, 957, 958, 959, 960, 961, 962, 963, 964, 965, 966, 967, 968, 969, 970, 971, 972, 973, 974, 975, 976, 977, 978, 979, 980, 981, 982, 983, 984, 985, 986, 987, 988, 989, 990, 991, 992, 993, 994, 995, 996, 997, 998, 999, 1000.

Done

http://www.nactem.ac.uk - Term List - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Rank	Term	Score
1	beta2-adrenergic receptor	68.25
2	blood pressure	16.75
3	beta2-adrenergic receptor gene	11.6797
4	odd ratio	10
5	protein kinase	9.599999
6	single nucleotide polymorphism	9.509775
7	adrenergic receptor	9.142857
8	gly16 allele	8
8	a549 cell	8
10	body mass index	7.924812
10	cystic fibrosis patient	7.924812
12	cystic fibrosis	7.428571
13	metabolic syndrome	7
13	confidence interval	7
15	bioluminescence resonance energy transfer	6.8
16	blood flow response	6.424812
17	gene polymorphism	6.4
18	diastolic blood pressure	6.22095

Done

mine_cvalue.cgi#

Google

Apply

on to be mediated by phosphorylation of the agonist-stimulated binding of beta-arrestins to the phosphorylated receptor. Binding of sensitization, internalization via clathrin-coated pits, and the initiation of interactions between receptors and beta-arrestin2 in living cells using assays of beta-arrestin2 binding to the receptor is limited by the kinetics of association to the accumulation of GRK2-phosphorylated receptor, which can compete with the receptor depends on the activation of the receptor by agonist-stimulated beta-arrestin2 complex. This fast agonist-controlled association and dissociation permit rapid control of receptor sensitivity in repeatedly stimulated cells.

beta-arrestin-coupled receptors and is present on skeletal and cardiac muscle. beta-arrestin2 has a moderate degree of heterogeneity in the human population and polymorphisms at positions 16, 27, and 164 are changed in asthma, obesity, and hypertension. The presence of the beta2-AR has also been suggested in human rheumatoid arthritis. The prevalence of the alleles Arg16 and Gln27 and a lower prevalence of the genotype combination Glu16/Gln27 had higher levels of

start | Microsoft Office P... | Mozilla Firefox | http://www.nactem... | Links EN Address | 18:42

TerMine

- C-value is a **hybrid** technique; extracts multi-word terms; language independent
- Combines **linguistic filters** and **statistics**
 - total frequency of occurrence of string in corpus
 - frequency of string as part of longer candidate terms (nested terms)
 - number of these longer candidate terms
 - length of string (in number of words)

Frantzi, K., Ananiadou, S. & Mima, H. (2000)
International Journal of Digital Libraries 3(2)

TerMine analysis of Obama's inauguration speech: close to your perception?

2.000000 common dangers

2.000000 health care

2.000000 new age

2.000000 new era

1.584962 few worldly possessions

1.584962 gross domestic product

1.584962 long rugged path

1.584962 many big plans

1.584962 stale political arguments

1.000000 american people

1.000000 bad habits

1.000000 better history

1.000000 better life

1.000000 bitter swill

1.000000 brave americans

1.000000 childish things

1.000000 civil war

1.000000 clean waters

1.000000 collective failure

1.000000 common defense

1.000000 common good

<SNIP>

Ordered by descending C-value, then by ascending alphabetic order

The importance of acronym recognition

- Acronyms are among the most productive type of term variation
- Acronyms are used more frequently than full terms
 - 5,477 documents could be retrieved by using the acronym JNK while only 3,773 documents could be retrieved by using its full term, c-jun N-terminal kinase [Wren et al. 05]
- No rules or exact patterns for the creation of acronyms from their full form

Top 20 acronyms in MEDLINE

Rank	Parenthetic phrase	# contextual sentence	# unique long-forms
1	CT	30,982	171
2	PCR	25,387	39
3	HIV	19,566	13
4	LPS	18,071	51
5	MRI	16,966	18
6	ELISA	16,527	25
7	SD	15,760	165
8	BP	14,860	145
9	DA	14,518	129
10	CSF	14,035	34
11	CNS	13,573	47
12	IL	13,423	60
13	PKC	13,414	11
14	TNF-ALPHA	12,228	14
15	HPLC	12,211	16
16	ER	12,155	140
17	RT-PCR	12,153	21
18	TNF	12,145	13
19	LDL	11,960	24
20	5-HT	11,836	20
..
—	(overall 50 acronyms)	600,375	4,212

Acromine Demonstration

After using this service, please complete a [questionnaire](#).

Enter an acronym in "Acronym" field to search its expanded forms. Alternatively, enter an expanded form in "Fullform" field to search its acronyms.

Acronym:

Fullform:

Found 41 definitions

Acronym	Full-form	Freq	Score
CPR	cardiopulmonary resuscitation	1505	1467.1
CPR	computer-based patient record	57	55.8
CPR	c-peptide immunoreactivity	52	46.8
CPR	cefprome	38	36.6
CPR	nadph-cytochrome p450 reductase	34	32.6
CPR	receptor	28	15.8
CPR	contraceptive prevalence rate	27	25.4
CPR	computerized patient record	20	19.0
CPR	cardio-pulmonary resuscitation	19	17.9
CPR	c-peptide reactivity	10	8.8

Intute repository search: single-interface search & browse

- NaCTeM provides core TM components for IRS:
 - Query builder for added usability
 - Real time clustering of search results
 - Term extraction for improved browsing options
 - Metadata creation for improved search capability
 - Personalisation tools to make the most of the information available
- Final deliverable includes web demonstrator and machine-to-machine interfaces for further integration into JISC e-Infrastructure
- www.nactem.ac.uk/intute/

Click **Generate Query** button to generate query.

Basic Query

Query field:

Use semicolon to separate words/phrases, e.g. "image; human brain". **DO NOT** use uppercase AND, OR and NOT in your query terms.

Terms must occur:

Terms may occur:

Unwanted terms:

Optional filters

Author: e.g. [Smith? P*] for "Paul Smith"

You can choose multiple items below.

Repositories:

- ALL
- Aberdeen University Research Archive
- Access to Research Resources for Teachers
- Advanced Knowledge Technologies EPrints Archive
- Birkbeck ePrints

Generate Query and Search

Generated query:

```
(ftTermNorm:(+health nursing -children)) AND dc_publisher:(20 1 30)
dc_creator:("Smith? P*")
```

Sort by

- An addition to the now standard query interface box
- Removes the need to learn complex query languages
- Build up your search in steps including wildcards
- Use additional filters to remove unwanted words or collections
- Option to edit query for more experienced users
- Continually updated based upon user requests

Document clustering

Visual Cluster Map

- Clustered Documents: 158
- [Randomised Controlled Trial](#) (15)
 - [Learning Disability Nurse](#) (9)
 - [Depression](#) (6)
 - [Care Services For Older People](#) (7)
 - [The Development Research Method](#) (8)
 - [Higher Education](#) (9)
 - [Social Work](#) (9)
 - [Midwives In Systems](#) (6)
 - [General Practice](#) (7)
 - [Chest Pain Unit](#) (7)
 - [Children](#) (5)
 - [Mental Health](#) (4)
 - [Community Centre](#) (5)
 - [NHS Library](#) (5)
 - [Clinical Skills And Practice](#) (6)
 - [Family Support For Stroke](#) (6)
 - [Problem Specific Knowledge](#) (4)
 - [Measure The Effect](#) (3)
 - [\(Other\)](#) (50)

Results: 1 - 10 of 158 found for "nursing". (0.028 seconds)

TITLE: [An Estimation of Distribution Algorithm for Nurse Scheduling](#)
AUTHORS: Aickelin, Uwe, Li, Jingpeng
FULL TEXT: http://eprints.nottingham.ac.uk/568/1/07annals_edu.pdf
ABSTRACT SNIPPETS: Schedules can be built in a similar way to a human involve domain knowledge. This paper presents an Estimation of Distribution scheduling problem, which involves choosing a suitable scheduling rule from **nurse**. Unlike previous work that used Genetic Algorithms (GAs) to impleme

TITLE: [Building Better Nurse Scheduling Algorithms](#)
AUTHORS: Aickelin, Uwe, White, Paul
FULL TEXT: http://eprints.nottingham.ac.uk/612/1/04annals_nurse.pdf
ABSTRACT SNIPPETS: The aim of this research is twofold: Firstly, to model scheduling problem with an integer programming formulation and evolution novel statistical method of comparing and hence building better scheduling algorithm modifications. The comparison method captures the results of alg

TITLE: [Building Better Nurse Scheduling Algorithms](#)
AUTHORS: Aickelin, Uwe, White, Paul
FULL TEXT: http://eprints.nottingham.ac.uk/662/1/04annals_nurse.pdf
ABSTRACT SNIPPETS: The aim of this research is twofold: Firstly, to model scheduling problem with an integer programming formulation and evolution novel statistical method of comparing and hence building better scheduling algorithm modifications. The comparison method captures the results of alg

TITLE: [Future career pathways in nursing and midwifery. A Delphi survey of England](#)
AUTHORS: Beattie, A., Hek, G., Ross, Kath, Galvin, Kathleen
FULL TEXT: http://eprints.bournemouth.ac.uk/1166/1/Galvin_Output_4.pdf
ABSTRACT SNIPPETS: There is growing interest, both nationally and intern research, regarding future career pathways for **nurses** and midwives and w represent ... **nurses** and midwives representing a wide range of **nurses** ar ... in **nursing** and midwifery could take ...

- Filter your results based upon regular underlying themes, in real time
- Lingo algorithm merges instances of commonly occurring phrases, keeping the best candidate to describe the documents
- Human readable labels make reaching your goal easier, faster and more efficient
- Visualisation option allows users to gain an overview and examine relationships between the clusters and documents.

Full Document Information

Similar Documents:

[Building Better Nurse Scheduling Algorithms](#)

[On the job rotation problem](#)

[Identifying and Assessing the Critical Risk Factors in an Underground Rail Project in Thailand: A Factor Analysis Approach](#)

[Exploiting problem structure in a genetic algorithm approach to a nurse rostering problem](#)

[An Estimation of Distribution Algorithm for Nurse Scheduling](#)

Term Extraction

- Identifies the most significant multi-word phrases within a document and adds them as metadata
- Uses TerMine
 - Can be used to browse towards related topics
- Useful for those new to or unfamiliar with a topic by suggesting other areas that may be of interest

Similar Documents

- Identifies conceptually similar documents using the most commonly occurring terms and words in the source document
- Useful for indentifying documents or repositories that you may not normally investigate
- Helps to solve **information overlook**

non-numeric or missing due to infeasibility. The final algorithm outperforms all previous evolutionary algorithms, which relied

Terms from Document
[nurse scheduling problem/](#)
[scheduling problem/](#)
[nurse scheduling/](#)
[comparison method/](#)
[integer programming formulation/](#)
[traditional statistical technique/](#)
[evolutionary algorithm](#)
[Dr Uwe Aickelin Dr P](#)
[White School/](#)
[scheduling algorithm/](#)
[Building Better Nurse Scheduling Algorithms](#)
[Annals/](#)
[integer programming](#)
[Nottingham NG8 1BE](#)
[UK Bristol/](#)
[complex nurse scheduling problem/](#)
[objective procedure/](#)
[single figure/](#)

TM for Social Sciences: ASSIST

- Innovative search engine that qualitatively analyses social sciences documents
- Domain knowledge facilitates query expansion
 - Term extraction for improved browsing capabilities
- Real time clustering of search results
- Semantic information enrichment for targeting the main topics
- Web demonstrator for further integration into JISC e-Infrastructure <http://www.nactem.ac.uk/assist/>

Conventional engines vs ASSIST

- Conventional
 - Return long list of documents, hard to filter
- ASSIST improves (case studies):
 - Research process with domain knowledge for the Educational Evidence Portal (EPPI-Centre)
 - Content access through semantic information for sociological analysis of mass-media documents (NCeSS)

Query interface

NaCTeM - ASSIST Browser

ASSIST Browser

authors:Wintour AND date:2005 AND educat* Search

Search results for: authors:Wintour AND date:2005 AND educat*
'ID', 'ID*' or 'ID OR Scientist'

Browse...

- Curriculum subject and skills
 - [School Sixth form](#) [Higher Adult learning...](#)
- Learning/teaching needs and practice
 - [Early years](#) [Primary](#) [Secondary](#) [Homework...](#)
- Performance, assessment, standards
 - [Inspection](#) [Early years education](#) [Secondary...](#)
- Careers and work experience
 - [Career guidance](#) [Employment](#) [Apprenticeships...](#)
- Administration, governance and finance
 - [Finance Buildings](#) [ICT](#) [Admissions](#) [Governance...](#)
- Teachers, CPD, Recruitment, Teacher education...
 - [Teachers](#) [CPD](#) [Recruitment](#) [Teacher education...](#)
- Home, community and society
 - [Parents](#) [Equality and diversity](#) [Crime](#) [NEETs...](#)
- Care, welfare and psychology
 - [Adoption](#) [Childcare](#) [Truancy...](#)

The NaCTeM ASSIST Browser is an beta version demonstrator for the [ASSIST](#) project funded by [JISC](#) to investigate the benefits of text mining in 2 case studies within the social science disciplines. This includes a review of the requirements gathering stage in order to advise future projects in this area and the development of high profile exemplars demonstrating how text mining software can solve, in part at least, major challenges facing e-Researchers across all domains.

- ♦ Automatic grouping of documents into *useful categories*
- ♦ Automatic creation of *descriptive labels* for categories based upon content
- ♦ Automatic identification of similar or *related documents*

Expanding the standard query interface

- ✓ Semantic operators to build complex queries
- ✓ Browsing documents through a domain taxonomy

Improving the rank of query results

- Resolution of Pronominal Anaphora relations to compute the real frequency of search words
(e.g. **The dog** eats **the cat**. **It** sleeps now)

Search result interface

NaCTeM - ASSIST

Cluster results for: education Visualize

	Query Results	Semantic Content
Documents (274)		
Blair Seeks (101)		
ID Cards (70)		
Labour at MANCHESTER (44)		
Brown (50)		
Willfully (42)		
Election AFTERMATH (34)		
David Cameron (42)		
Blunkett Affair (31)		
Government (33)		
Just Week (25)		
New (28)		
Us' (22)		
Delay School Reform (20)		
Religious Hatred Bill (21)		
Britain (21)		
(Other) (275)		

Query Results | Semantic Content

TITLE: The State of Intellectual Property Education Worldwide
AUTHORS: Lakhan, Shaheen F., Khurana, Meenakshi K.
FULL TEXT: http://cogprints.org/5640/1/IP_Education.pdf
ABSTRACT SNIPPETS: , **education** in intellectual property is required and must be advocated. We must make individuals ... property rights and provides a detailed overview of the state of intellectual property education worldwide. The discussion ...

TITLE: Embracing ignorance in Higher Education
AUTHORS: Soetendorp, Ruth
FULL TEXT: <http://eprints.bournemouth.ac.uk/3411/1/722.pdf>
ABSTRACT SNIPPETS: Ignorance receives a bad press which it doesn't deserve. Negative and unwarranted associations with stupidity and foolishness can make ignorance a quality from which to shy or for which to apologise particularly when **education** is on the agenda ...

TITLE: Exclusive Brethren: an Educational Dilemma.
AUTHORS: Bigger, Stephen
FULL TEXT: <http://eprints.worc.ac.uk/241/1/EXCLUSIVES.pdf>
ABSTRACT SNIPPETS: An article from 1990 on the Exclusive Brethren and **Education**, particularly focusing on the ICT National Curriculum regulations that came out at that time, since Exclusive Brethren parents wished to withdraw their children from ICT on conscientious grounds. The paper follows their arguments. An update from 2007 has been added ...

TITLE: Citizenship education in the UK: devolution, diversity and divergence
AUTHORS: Andrews, Rhys, Mycock, Andrew
FULL TEXT: <http://eprints.hud.ac.uk/563/1/MycockCitizenship.pdf>
ABSTRACT SNIPPETS: of **education** in the UK. But to what extent does citizenship **education** receive equal attention within the four UK Home Nations? And, what are the implications of

- ✓ Clustering the query results in real time
- Lingo algorithm merges instances of commonly occurring phrases, keeping the best candidate to describe each cluster
- ✓ A familiar presentation of query results including snippets

Search result interface

NaCTeM - ASSIST

Cluster results for: 'education' [Visualize](#)

	Query Results	Semantic Content
All Documents (684)		
Blair Seeks (101)		
ID Cards (70)		
Labour at MANCHESTER (64)		
Brown (50)		
Willfully (42)		
Election AFTERMATH (34)		
David Cameron (42)		
Blunkett Affair (31)		
Government (33)		
Just Week (25)		
New (28)		
Us' (22)		
Delay School Reform (20)		
Religious Hatred Bill (21)		
Britain (21)		
(Other) (275)		

DEAR SU...
AUTHORS: Unknown, **DATE:** May 27, 2005
TERMS: [MARK TAYLFORTH Kensington, ID card, tangible benefit, pension shortfall, health service, We...

Reply: Letters and emails: Towards a surveillance society -
AUTHORS: Robin Hull, **DATE:** April 26, 2006 Wednesday
TERMS: [legal compensation shakeup, false court action, Robin Hull London, compensation issue, dangerous issue, lingering sentence, ID card, human error, good faith, good evidence]

PUPILS FACE CLOCKING IN -
AUTHORS: Unknown, **DATE:** April 19, 2004
TERMS: [ID card scanning, human right issue, swipe card, St Thomas, High School, pilot project, Education chief, Marion Pagani, Glasgow Children]

Learner numbers are a step towards ID cards -
AUTHORS: Unknown, **DATE:** February 15, 2008, Friday
TERMS: [national insurance number, income tax purpose, robert steel Salisbury, Government intent, unique number, birth certificate, HM Revenue, education purpose]

Learner numbers are a step towards ID cards -
AUTHORS: Unknown, **DATE:** February 15, 2008, Friday
TERMS: [national insurance number, income tax purpose, robert steel Salisbury, Government intent, unique number, birth certificate, HM Revenue, education purpose]

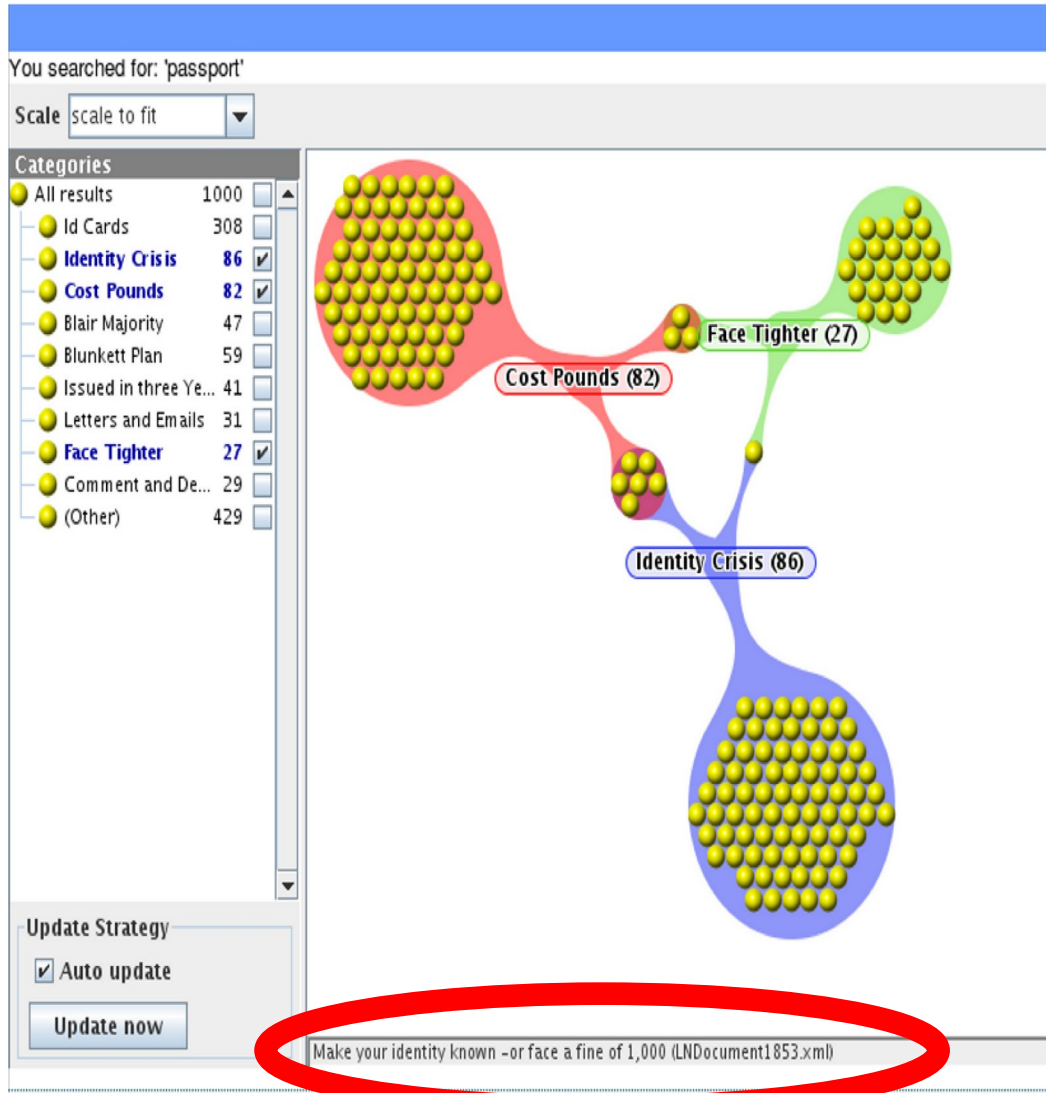
Whitehall poised for shuffle of top posts -
AUTHORS: David Hencke, Westminster correspondent, **DATE:** July 18, 2005
TERMS: [Sir John, permanent secretary, principal private secretary, Home Office, cabinet secretary, Gordon Brown, public service spending programme, Sir David, Mr Clarke, senior civil servant]

Labour reshuffle: The winners -
AUTHORS: Unknown, **DATE:** May 6, 2006 Saturday
TERMS: [John Reid Who Former communist bruiser, Alan Johnson Who Alan Johnson, Hazel Blears Who

Document content is described using semantic information

✓ makes document analysis easier, faster and more efficient

Query result visualisation



- Examination of cluster memberships via a friendly visualisation interface
- Graphical representation of the intersection between the clusters provides immediate visualization of cluster relations
- ✓ Information regarding membership of particular cluster

Document analysis

NaCTeM - ASSIST

Document: LNDocument1135.xml

Interviews to cut passport fraud

Fingerprinting and eye scans may also be brought in to tighten security.

Cath Urquhart reports No, it's not Dubai, it's Portsmouth.

This is the Spinnaker Tower, due to open this month, which at 170m (547ft) is higher than the London Eye or Blackpool Tower.

The tower, which has three viewing decks looking towards the Isle of Wight, is part of the Harbour Renaissance of Portsmouth Project. **new passport system**, as claimed, "will be simple for people who really are who they claim to be". **THE INTRODUCTION** of "e-passports", starting **early next year**, will see passport prices rise. Only a few applicants being called in for one-to-one interviews to obtain the travel document.

E-passports, or **biometric passports**, are to be phased in from February.

They will bear a chip containing biometric data -initially, a facial scan taken from a photograph although a fingerprint scan is likely to be included from 2008.

Britons who need to apply for their **first adult passport**, or whose passport is lost or stolen, will have to attend a face-to-face interview before they will be granted a **biometric passport**.

A network of 70 **new passport offices** will be created across the country, to supplement the existing seven offices, where the interviews, likely to start from October 2006, will take place.

This autumn the **UK Passport Service** (UKPS) is likely to announce a **huge price rise** to cover the cost of **biometric passports**.

Figures are not yet available, but the projected **unit cost** of the passport in 2006-07, according

A safer, more convenient passport. Now would you like chips with that?
A face-to-face interview to get your first passport

Face-to-face queuing for 600,000 first-time passport applicants
National: Passport price to rise for third time in less than two years: Increase to fund consular service, says Foreign Office
Bill is underwriting cost of ID cards, say opponents
Fingerprints plan for new passports

Related Topics LOW MED HIGH

biometric passport
ID card
human right group
adult passport
unit cost
early next year
first adult passport
full adult passport
huge price rise
ID card centre
ID card legislation
interesting moral dilemma
new passport office
new passport system
passport price increase
UK Passport Service

- ✓ Identification of conceptually similar documents using the most commonly occurring terms and words in the source document
- ✓ Highlighting selected semantic information within the document
- ✓ Selecting terms according to their importance and using them to browse documents

ASSIST architecture

Multi-format documents



Conversion tools

.PDF with *pdfbox*
.DOC with *POI*
.HTML with *Jtidy*
.XML



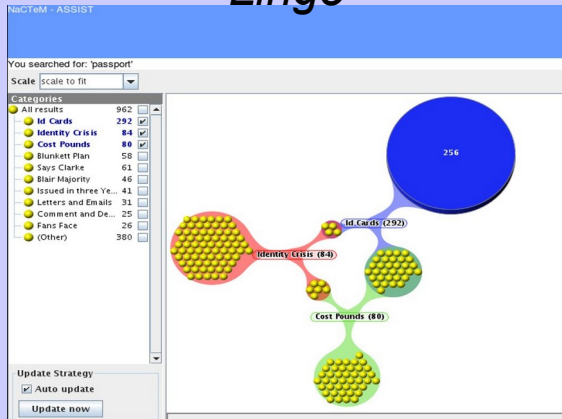
TM components

- Named Entity Recognizer
BaLIE
- Term Extractor
Termine
- Anaphora resolver
Bayaphora
- Lexical Chain extractor



Search Engine
Lucene

Search result clustering *Lingo*



Web Query Interface

NaCTeM - ASSIST Browser

ASSIST Browser

authors:Wintour AND date:2005 AND educat* Search

Sample queries include:
10, 10* or 10 OR Scientist

Browse...

- Curriculum subject and skills
- Administration, governance and finance
- School Sixth form Higher Adult learning...
- Finance Buildings ICT Admissions Governors...
- Learning/teaching needs and practice
- Teachers, lectures and support staff
- Early years Primary Secondary Homework...
- Teachers CPD Recruitment Teacher education...
- Performance, assessment, standards
- Home, community and society
- Inspection Early years education Secondary...
- Parents Equality and diversity Crime NEETS...
- Careers and work experience
- Care, welfare and psychology
- Career guidance Employment Apprenticeships...
- Adoption Childcare Truancy...

The NaCTeM ASSIST Browser is a beta version demonstrator for the ASSIST project funded by JISC to investigate the benefits of text mining in 2 case studies within the social science disciplines. This includes a review of the requirements gathering stage in order to advise future projects in this area and the development of high profile exemplars demonstrating how text mining solutions can solve, in part at least, major challenges facing e-Researchers across all domains.

- Automatic grouping of documents into useful categories
- Automatic creation of descriptive labels for categories based upon content
- Automatic identification of similar or related documents

User Query



Indexed Documents



Querying without semantic annotation

KLEIO

The screenshot shows a Mozilla Firefox browser window with the address bar displaying 'http://nactem4.mc.man.ac.uk:8080/Kleio/SimpleSearch.c'. The page header includes the NaCTeM logo and the text 'The National Centre for Text Mining'. On the left side, there is a navigation menu with 'New Query' and 'NaCTeM Services' (Termines, Acromine, Cheshire/Termines, Medie, Info-Pubmed). The main content area shows a search query 'cat' circled in red, with 'Results 1 - 10 of 60711' displayed below it. Navigation links 'First', 'Previous', 'Next', and 'Last' are present. A list of search results follows, each with a title, journal information, and a PubMedID link. The first result is 'Axillary lymphadenopathy secondary to cat-scratch disease.' with PubMedID 16255119. The second result is 'Early relief of osteoarthritis symptoms with a natural mineral supplement and a herbomineral combination: a randomized controlled trial [ISRCTN38432711].' with PubMedID 16242032. The third result is 'The cover. Black Cat on a Chair.' with PubMedID 16234486. The fourth result is 'Secondary calcium-binding parameter of Bacillus amyloliquefaciens alpha-amylase obtained from inhibition kinetics.' with PubMedID 16233519.

- False Positives: similarity with non-protein entities
- False Negatives: search ignores synonym forms
- Poor accuracy (e.g. more than 60,000 results for “cat”)

Querying with semantic annotation PROTEIN: cat

The screenshot shows the NaCTeM website interface. At the top, there is a navigation menu with 'File', 'Edit', 'View', 'Go', 'Bookmarks', 'Tools', and 'Help'. Below this is a search bar containing the URL 'http://nactem4.mc.man.ac.uk:8080/Kleio/S'. The NaCTeM logo is prominently displayed, with the text 'The National Centre for Text Mining' underneath. On the left side, there is a sidebar with 'New Query' and 'NaCTeM Services' including 'Termine', 'Acromine', 'Cheshire/Termine', 'Medie', and 'Info-Pubmed'. The main content area shows the search results for the query 'PROTEIN:cat'. The query is highlighted with a red circle and a speech bubble. Below the query, it says 'Results 1 - 10 of 237' and provides navigation links: 'First', 'Previous', 'Next', and 'Last'. A list of results follows, each with a title, journal information, and a PubMedID link. The results are:

- 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**

- Provides a more focused query
- Returns only documents with annotated protein
- Allows better integration with external protein databases and resources
- Returns fewer documents (237 for "PROTEIN:cat")

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

File Edit View History Bookmarks Tools Help

NaCTeM
The National Centre for Text Mining

New Query

NaCTeM Services
Termino
Acromine
Cheshire/Termino
Medie
Info-Pubmed

Kleio

Query info:
Query String: pain AND SYMPTOM:"neuropathic pain" AND ORGAN:"spinal cord"

Search Results

Facets:

MeshHeading [show/hide the rest terms]
[Spinal Cord -- metabolism \(24\)](#) [Time Factors \(20\)](#) [Research Support, Non-U.S. Gov't \(18\)](#) [Spinal Cord -- physiology \(15\)](#) [Spinal Cord -- physiopathology \(14\)](#)

PROTEIN [show/hide the rest terms]
[substance P \(16\)](#) [protein kinase C \(11\)](#) [tumor necrosis factor-alpha \(8\)](#) [L6 \(7\)](#) [substance P receptor \(6\)](#)

GENE [show/hide the rest terms]
[c-fos \(7\)](#) [enkephalin \(5\)](#) [calcitonin gene-related peptide \(3\)](#) [cyclooxygenase \(3\)](#) [tumor necrosis factor \(2\)](#)

METABOLITE [show/hide the rest terms]
[morphine \(2\)](#) [peptide \(16\)](#) [saline \(15\)](#) [condition \(10\)](#) [galanin \(10\)](#)

DISEASE [show/hide the rest terms]
[spinal cord \(5\)](#) [chronic pain \(31\)](#) [syndrome \(19\)](#) [heat \(16\)](#) [cold \(16\)](#)

SYMPTOM [neuropathic pain \(34\)](#) [thermal hyperalgesia \(60\)](#) [peripheral neuropathic pain \(6\)](#) [secondary hyperalgesia \(3\)](#)

ORGAN [show/hide the rest terms]
[spinal cord \(28\)](#) [spinal nerve \(49\)](#) [spinal nerves \(20\)](#) [synaptic transmission \(5\)](#) [substantia gelatinosa \(5\)](#)

GENERAL_PHENOM [receptor internalization \(1\)](#)

HUMAN_PHENOM

NATURAL_PHENOM [show/hide the rest terms]
[reduction \(52\)](#) [plasticity \(26\)](#) [heat \(23\)](#) [compression \(5\)](#) [light \(5\)](#)

DIAG_PROC [MRI \(4\)](#) [Neurostimulation \(3\)](#) [electrophysiological studies \(2\)](#) [US \(1\)](#) [neuroimaging \(1\)](#)

THERAPEUTIC_PROC [show/hide the rest terms]
[intrathecal injection \(27\)](#) [axotomy \(14\)](#) [Intrathecal injection \(8\)](#) [Spinal cord stimulation \(6\)](#) [pharmacotherapy \(5\)](#)

INDICATOR [antisense oligonucleotide \(5\)](#) [AMPA \(3\)](#) [radioligand \(2\)](#) [PTIO \(2\)](#) [acetic acid \(2\)](#)

- PMID:** [15479977](#) (Score: [-11.318087](#) Date: 20041101)
Title: [DREAMING about arthritic pain](#)
Abstract: The experience of acute [pain](#) serves a crucial biological purpose: it alerts a living organism. In contrast, chronic [pain](#) arising from disease states and/or pathological functioning of the nervous system understanding of [pain](#) mechanisms, the satisfactory management of pathological [pain](#) eludes current treatment
- PMID:** [12022219](#) (Score: [11.317133](#) Date: 20011201)
Title: [Neuropathic pain--recent trends in management](#)
Abstract: Injury to central or peripheral nervous system causes neuropathic [pain](#). Initially it affects are affected simultaneously and the symptoms develop simultaneously. Spontaneous [pain](#), abnormal evoked [pains](#) and paroxysmal evoked [pain](#) are the pesenting symptoms. Types of neuropathies are: Peripheral nerve lesion, spinal
- PMID:** [11838651](#) (Score: [-11.293789](#) Date: 20020201)

Select listed entities to add them to query and narrow down the abstract list

List of retrieved documents is updated with the new queries

MEDIE - Semantic retrieval engine for MEDLINE - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www-tsujii.is.s.u-tokyo.ac.jp/medie3/search.cgi

Most Visited Customize Links Free Hotmail Windows Marketplace Windows Media Windows Statistics for text0.mi... take! home page

MEDIE — Semantic query based on facts

Specify the subject

subject	verb (base form)	object
<input type="text" value="p53"/>	<input type="text" value="activate"/>	<input type="text"/>

search clear

advanced search

What is MEDIE?

Specify the verb

Click to search!

MEDIE is an intelligent search engine to retrieve biomedical correlations from MEDLINE, based on indexing by Natural Language Processing and Text Mining techniques. You can find abstracts/sentences in MEDLINE by specifying semantics of correlations; for example, "What activates p53" and "What causes colon cancer".

Currently, 18,018,361 MEDLINE articles are in... Sat Nov

What does p53 activate?

GCL Sea

A GCL... used to a GCL server. See the following table for the details of GCL. If you want to see examples of GCL queries, just click "Show...". In the search summary of semantic/keyword search, and you will see a GCL query submitted to a server.

[The customization of the number of results](#) is available as in Semantic Search.

Done

MEDIE - Semantic retrieval engine for MEDLINE - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www-tsujii.is.s.u-tokyo.ac.jp/medie3/search.cgi?search_type=semantic_search&subject=p53

MEDIE is a demo system presented by Tsujii Laboratory

subject	verb (base form)	object
<input type="text" value="p53"/>	<input type="text" value="activate"/>	<input type="text"/>

search clear advanced search

Results 1-10 for **p53 activate** 0.21 seconds (searched 0.24% of Medline)

» show query

sentence **article** table show 10 results subject verb object [gene](#) [disease](#)

show next »

- [Oscillations by the p53-Mdm2 feedback loop. »XML](#)
Galit Lahav, pp. 28-38, Volume 641, Advances in experimental medicine and biology, 2008 [PMID:18783169]
p53 also **activates** the transcription of **Mdm2**, which in turns target **p53** for degradation, therefore creating a negative feedback loop on **p53**.
- [ERK and JNK mediate p53 activation in apoptotic and autophagic L1210 cell death. »XML](#)
- [Triphala inhibits both in vitro and in vivo xenograft growth of pancreatic tumor cells by inducing apoptosis. »XML](#)
Yan Shi, Ravi P Sahu, Sanjay K Srivastava, pp. 294, Volume 8, BMC cancer, 2008 [PMID:18647491]
Our data also suggests that the growth inhibitory effects of Triphala is mediated by the activation of ERK and p53 and shows potential for the treatment and/or prevention of human pancreatic cancer.

Done

Click to change the view

the growth inhibitory effects of Triphala is mediated by the activation of ERK and p53 ...

MEDIE — **Perform advanced search** a demo system presented by [Tsuji Laboratory](#)

subject	verb (base form)	object
<input type="text" value="p53"/>	<input type="text" value="activate"/>	<input type="text"/>
		<input type="button" value="search"/> <input type="button" value="clear"/>

[advanced search](#)

Results 1-10 for **p53 activate** 0.21 seconds (searched 0.24% of Medline)
 » show query

show

[gene](#)
[disease](#)

show next »

title	subject		verb	object	
	subject	entities	entities	object	entities
Oscillations by the p53-Mdm2 feedback loop. »XML	p53		activates	the transcription of Mdm2	
ERK and JNK mediate TNFalpha-induced p53 activation in apoptotic and autophagic L929 cell death. »XML		TNFalpha-induced MAPKs mediate p53 activation in apoptotic and autophagic cell death, as well as autophagy	amplify	apoptosis	
Regulation and pathological role of p53 in cisplatin nephrotoxicity. »XML	p53		leading	the development of effective renoprotective strategies during cancer therapy	
Triphala inhibits both in vitro and in vivo xenograft growth of pancreatic tumor cells by inducing apoptosis. »XML		by the activation of ERK and p53	mediated	the growth inhibitory effects of Triphala	
Pharmacogenetics and pharmagenomics trends in normal and pathological aging studies: focus on p53. »XML	p53		activate	an apoptotic program	
p53 family in development. »XML		Imbalance of p53 protein family	contribute	a significant proportion of congenital developmental abnormalities in humans	
Involvement of p53 and Raf/ MEK / ERK pathways in hematopoietic drug resistance. »XML		Dominant-negative (DN) p53 genes	increased	the resistance to chemotherapeutic drugs: MDM2 and MEK inhibitors	



Search only the conclusion sentences

	subject	verb (base form)	object	
	<input type="text" value="p53"/>	<input type="text" value="activate"/>	<input type="text"/>	<input type="button" value="search"/> <input type="button" value="clear"/>
enable ontology search	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	hide options
treat as base forms	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	help
search category	<input checked="" type="checkbox"/> gene <input checked="" type="checkbox"/> product <input checked="" type="checkbox"/> disease		<input checked="" type="checkbox"/> gene <input checked="" type="checkbox"/> product <input checked="" type="checkbox"/> disease	help
verb modifiers				help
additional keywords	<input type="text"/>			help
author	<input type="text"/>			help
journal title	<input type="text"/>			help
MeSH keywords	<input type="text"/>			help
sentence types	<input type="checkbox"/> title <input checked="" type="checkbox"/> conclusion <input type="checkbox"/> method <input type="checkbox"/> objective <input type="checkbox"/> result			help

Results 1-10 for **p53 activate** 6.68 seconds (searched 54.44% of Medline)
» show query

sentence | **article** | table show 10 results ▼ subject verb object gene disease

[show next »](#)

1. [Oncogenic mutation of the p53 gene derived from head and neck cancer prevents cells from undergoing apoptosis after DNA damage. »XML](#)
Hitoshi Kawamata, Fumie Omotehara, Koh-Ichi Nakashiro, Daisuke Uchida, Yasuhiro Shinagawa, Masatsugu Tachibana, Yutaka Imai, Takahiro Fujimori, pp. 1089-97, Volume 30, Issue 5, International journal of oncology, 2007 [PMID:17390010]

A mutant-p53 (Glu17Lys, His193Leu) or a truncated p53 (Delta121) did not activate the reporters containing p53 responsive elements from p21waf1, BAX, MDM2, p53AIP1, and PUMA genes at all.

MEDIE - Semantic retrieval engine for MEDLINE - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www-tsujii.is.s.u-tokyo.ac.jp/medie3/search.cgi?search_type=semantic_search&subject=p!

Most Visited Customize Links Free Hotmail Windows Marketplace Windows Media Windows Statistics for text0.mi... take' home page

MEDIE — **only conclusion facts** MEDIE is a demo system presented by [Tsuji Laboratory](#)

subject	verb (base form)	object
<input type="text" value="p53"/>	<input type="text" value="activate"/>	<input type="text"/>

search clear advanced search

Results 1-10 for p53 activate

» show

1. [ERK and JNK mediate TNFalpha-induced apoptosis in apoptotic and autophagic L929 cell death.](#) »XML
 Yan Cheng, Feng Qiu, Shin-ichi Tashiro, Satoshi Onodera, Takashi... 183-8, Volume 376, Issue 3, Biochemical and biophysical research communications, 2008 [PMID:18796294]

In conclusion, these results demonstrate that TNFalpha-induced MAPKs mediate **p53** activation in apoptotic and autophagic cell death, as well as autophagy may **amplify** apoptosis when associated with a death signaling pathway.

2. [Triphala inhibits both in vitro and in vivo growth of pancreatic tumor cells by inducing apoptosis.](#) »XML
 Yan Shi, Ravi P Sahu, Sanjay K Srivastava, pp. 294, Volume 8, BMC... 2008 [PMID:18847491]

Our data also suggests that the growth inhibitory effects of Triphala is **mediated** by the activation of ERK and **p53** and shows potential for the treatment and/or prevention of human **pancreatic cancer**.

3. [p53 family in development.](#) »XML
 Nadia Danilova, Kathleen M Sakamoto, Shuo Lin, pp. 919-31, Volume 125, Issue 11-12, Mechanisms of development, YYYY [PMID:18835440]

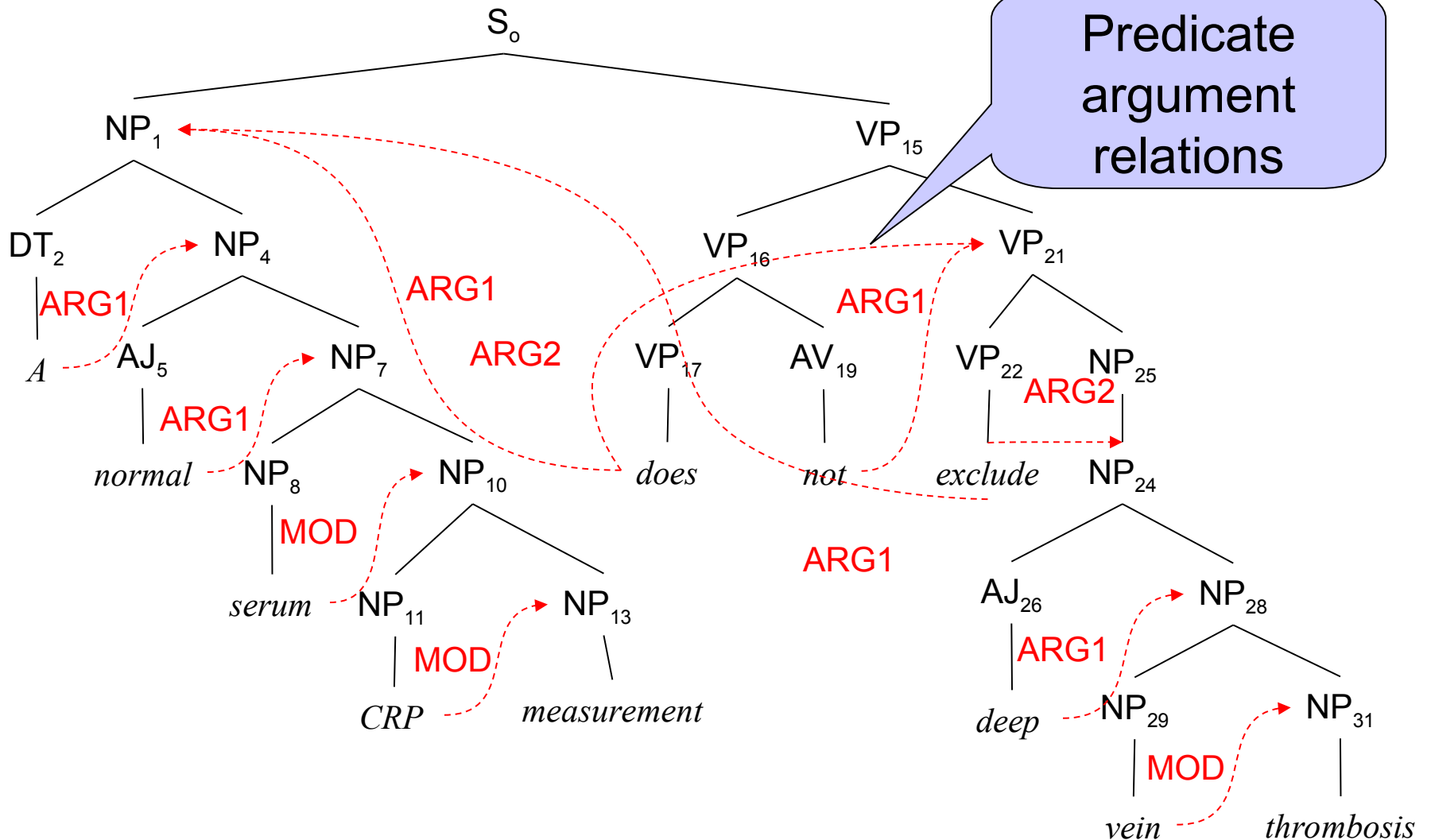
Imbalance of **p53** protein family may contribute to a significant proportion of congenital developmental abnormalities in humans.

Done

In conclusion, ...

Our data also suggests that ...

Semantic structure



FACTA: finding associated concepts

nicotine - FACTA Search

http://text0.mib.man.ac.uk/software/facta/a.cgi?query=nicotine|111111|20|2&cat=human&cat=disease&cat=sympt

FACTA

nicotine

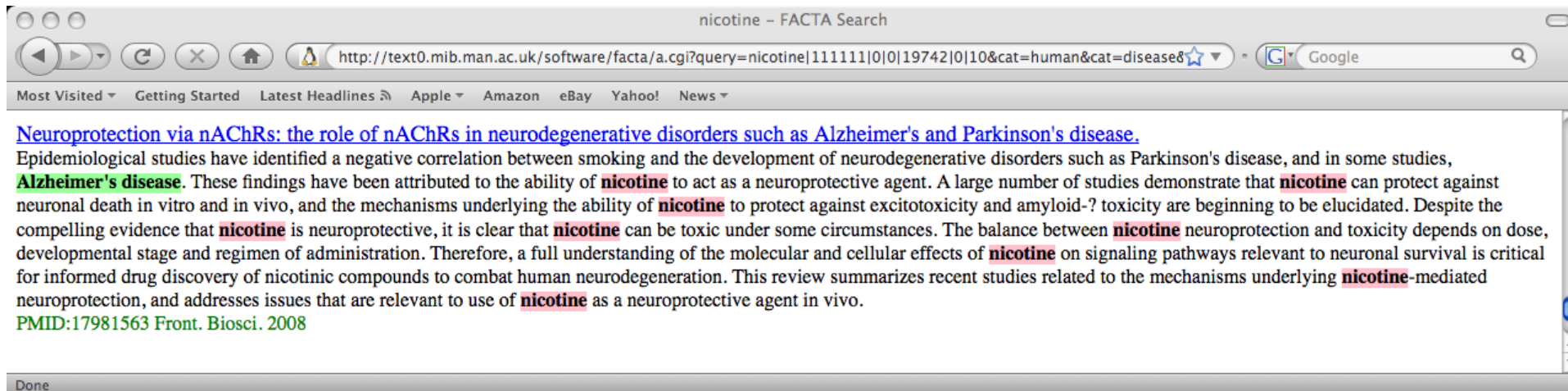
Gene/Protein Disease Symptom Drug Enzyme Compound

Query: **nicotine**
 19,779 document(s) hit in 17,702,258 MEDLINE articles (0.04 seconds). [Excerpts](#) (click to show).

Concepts found in the documents ranked by [[Frequency](#) | [Pointwise Mutual Information](#) | [Freq. * PMI](#)] .

Human Gene/Protein	Disease	Symptom	Drug	Enzyme	Compound
nicotinic acetylcholine receptor 9087.7	nicotine addiction 8854.6	nausea 215.6	caffeine 1659.3	acetylcholinesterase 808.1	Nicotine 45193.9
CYP2A6 1122.4	addiction 4010.1	hunger 190.5	Nicorette 597.1	tyrosine hydroxylase 484.0	ACh 2735.0
muscarinic receptor 971.9	tobacco dependence 2553.6	agitation 189.9	Atropine 553.3	cholinesterase 404.9	CO2 2334.9
neuronal nicotinic acetylcholine receptor 965.0	depression 1074.5	tremor 160.3	Nicorette 548.1	choline acetyltransferase 376.7	Cotinine 1839.5
acetylcholine receptor 808.3	Alzheimer's disease 919.8	seizures 156.7	Zyban 394.0	6-hydroxy-D-nicotine oxidase 356.0	Mecamylamine 1797.6
acetylcholinesterase 533.7	drug addiction 915.4	insomnia 147.3	Ethanol 216.4	nicotine oxidase 254.9	caffeine 1651.2
tyrosine hydroxylase 485.8	alcoholism 893.8	hypothermia 135.7	Rimonabant 148.8	protein kinase C 181.4	ethanol 1540.1
choline acetyltransferase 397.1	schizophrenia 854.7	dizziness 110.0	Clonidine 140.3	putrescine N-methyltransferase 168.0	noradrenaline 1406.8
substance P 309.8	Nicotine withdrawal 764.8	analgesia 97.0	Menthol 140.2	monoamine oxidase 156.3	calcium 1356.8
CA1 305.5	lung cancer 727.0	hypercapnia 85.6	Scopolamine 122.8	NADPH-diaphorase 116.0	Bupropion 1217.4
vasopressin 304.8	substance abuse 670.7	vomiting 72.5	Tubocurarine 122.3	6-hydroxy-L-nicotine oxidase 115.0	trans-3'-hydroxycotinine 836.3
AChR 270.9	substance use 600.4	prostration 70.3	FAD 117.6	NO synthase 100.9	glutamate 730.0
NMDA receptor 263.9		hyperoxia 48.5	Norepinephrine 109.5	dopamine 85.6	DSM 723.9
dopamine transporter 258.9		skin irritation 46.6	nitroglycerin 83.0		pyridine 689.5
		cough 38.3	Vardenafil 69.5		Epibatidine 595.1
		lightheadedness 33.8	Pilocarpine 64.2		

Nicotine and AD



The screenshot shows a web browser window with the title "nicotine - FACTA Search". The address bar contains the URL: <http://text0.mib.man.ac.uk/software/facta/a.cgi?query=nicotine|111111|0|0|19742|0|10&cat=human&cat=disease&>. The search engine is Google. The browser's menu bar includes "Most Visited", "Getting Started", "Latest Headlines", "Apple", "Amazon", "eBay", "Yahoo!", and "News".

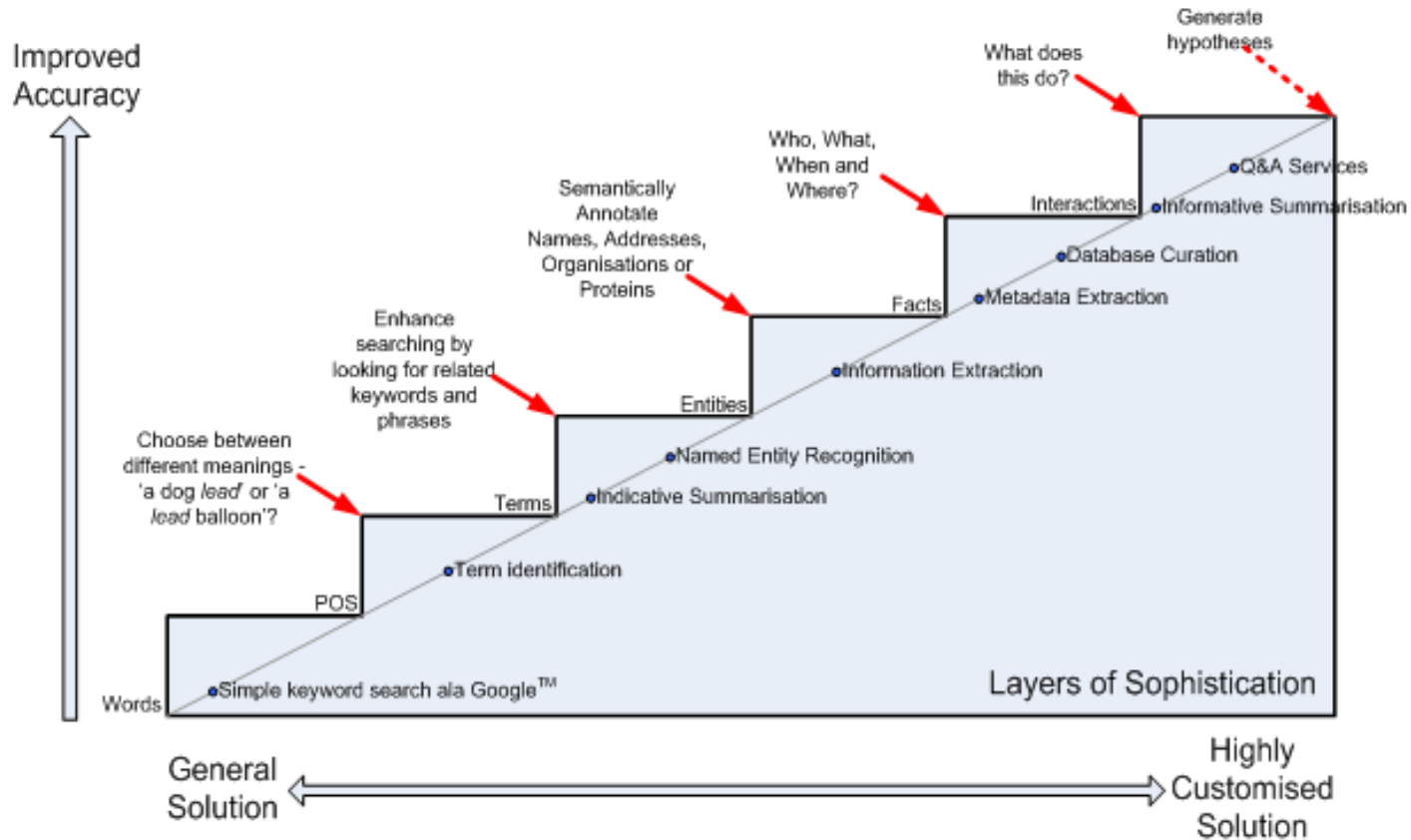
[Neuroprotection via nAChRs: the role of nAChRs in neurodegenerative disorders such as Alzheimer's and Parkinson's disease.](#)

Epidemiological studies have identified a negative correlation between smoking and the development of neurodegenerative disorders such as Parkinson's disease, and in some studies, **Alzheimer's disease**. These findings have been attributed to the ability of **nicotine** to act as a neuroprotective agent. A large number of studies demonstrate that **nicotine** can protect against neuronal death in vitro and in vivo, and the mechanisms underlying the ability of **nicotine** to protect against excitotoxicity and amyloid- β toxicity are beginning to be elucidated. Despite the compelling evidence that **nicotine** is neuroprotective, it is clear that **nicotine** can be toxic under some circumstances. The balance between **nicotine** neuroprotection and toxicity depends on dose, developmental stage and regimen of administration. Therefore, a full understanding of the molecular and cellular effects of **nicotine** on signaling pathways relevant to neuronal survival is critical for informed drug discovery of nicotinic compounds to combat human neurodegeneration. This review summarizes recent studies related to the mechanisms underlying **nicotine**-mediated neuroprotection, and addresses issues that are relevant to use of **nicotine** as a neuroprotective agent in vivo.

PMID:17981563 Front. Biosci. 2008

Done

Challenge: Complex analysis currently requires highly customised solutions



Challenge: Dealing with full text

- Need to be able to handle very large amounts of text
- Other issues besides linguistic/NLP ones (already hard)
 - Efficiency, scalability, distributed processing
- Porting TM tools to UK and European Grid environment

Need for processing full texts

- Allow researchers to discover hidden relationships from text that were not known before
 - an abstract's length is on average 3% of the entire article
 - an abstract includes only 20% of the useful information that can be learned from text

Parallelising TM

- TM applications are data independent
 - Scale linearly in an ideal world
- HPC implementation
- Scaled linearly to 100 processors
- Porting to DEISA to scale over 1000s processors to process TBs of data in reasonable time

TM of full texts for UK PubMed Central (UKPMC)

- Free archive of life sciences journals
- British Library, European Bioinformatics Institute & UManchester (NaCTeM, Mimas)
- Phase 3 tasks: integration of UKPMC in biomedical DB infrastructure with TM solutions for improved search and knowledge discovery



NaCTeM in UKPMC

- TM “behind the scenes” on full texts
- Named entity recognition
 - Link entities in texts to bioDB entries
- Fact extraction
 - E.g., protein-protein interactions
- Add extracted info as semantic metadata
 - Index for efficient access
- Semantic search capability
 - Based on user needs, evaluation workshops

Uses of our tools and services

- Searching
- Metadata creation
- Controlled vocabularies
- Ontology building
- Data integration
- Linking repositories
- Database curation
- Reviewing
- Gene – disease mining
- Enriching pathway models
- Indexing
- Document classification
-

NaCTeM phase II (2008-2011)

TM supporting service provision

- **Web Services**
- **Embedding TM within workflows**
- **Adaptive learning**
- **Integration of data / text mining**

Issues

- **Full paper processing**
- **open access collections**
- **IPR in data derived via text mining**
- **Interoperability**
- **Education and training**

NaCTeM phase II (2008-2011)

Service exemplars

- **Intelligent semantic searching for construction of biological networks**
- **Support for qualitative data analysis for social sciences**
- **Intelligent semantic search of gene-disease associations for health**

e-Research and e-Science

- **Knowledge discovery**
- **Collaborative research**
- **E-publishing**
- **Personalised searching**

Acknowledgments

- Text Mining Team: 16 members
- NaCTeM funding agencies:



- Wellcome Trust
- Close collaboration with University of Tokyo



Acknowledgements

- User group



Systems Biology Centres

- Middleware provider



<http://taverna.sourceforge.net>

- Usability and evaluation



- Service provision MIMAS



Further reading

- Visit our site at www.nactem.ac.uk for TM briefing paper and other publications on our work
- If you're a biologist/bioinformatician:
Ananiadou, S. & McNaught, J. (eds) (2006) Text Mining for Biology and Biomedicine. Norwood, MA: Artech House.