Mining the BMJ Group Corpus

Tom Diethe (t.diethe@cs.ucl.ac.uk | tdiethe@bmjgroup.com)

British Medical Association
and
Centre for Computational Statistics & Machine Learning, University College London

19/10/2011

Introduction

Mining the British Medical Journal Group Corpus

- The Content Set
- Related work - two (three?) approaches
- Approach taken so far
- Preliminary Results
- Further work (lots of it!)

http://group.bmj.com

The Content Set

- The British Medical Journal Group content set consists of a variety of content types:
  - Journals
  - Best Health, Best Practice, Clinical Evidence
  - Job adverts
  - BMJ Careers
  - Doc2Doc forum
  - Portfolio
- (Almost) all of the data is in some kind of XML format, with varying levels of structure.
- Possibly include outside content
  - Medline abstracts
  - News feeds (*e.g.* Reuters)

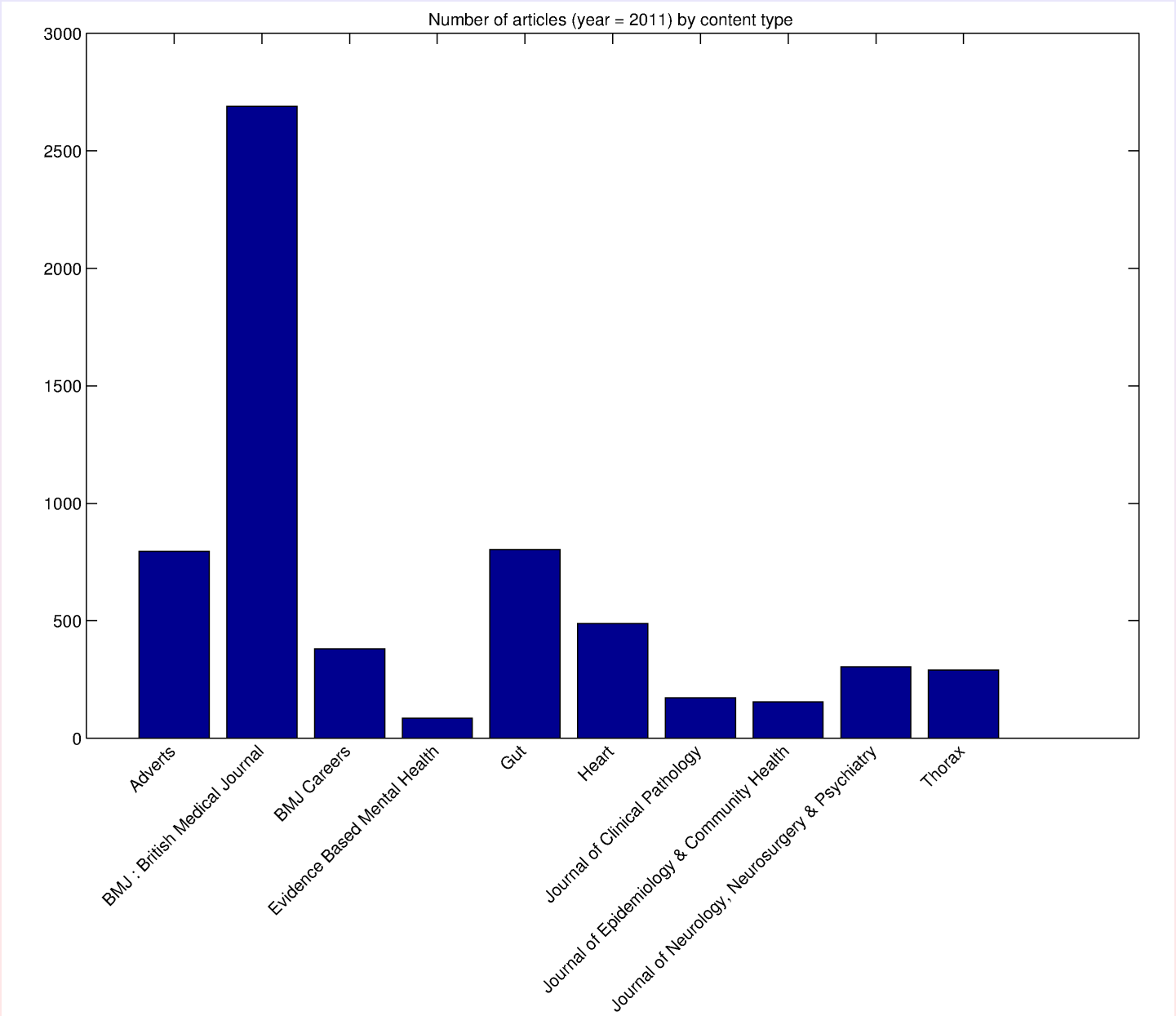
Journals

- Acupuncture in Medicine
- Annals of the Rheumatic Diseases (ARD)
- Archives of Disease in Childhood (ADC)
- British Medical Journal (BMJ)
- BMJ Case Reports
- BMJ Open
- BMJ Quality & Safety
- BMJ Supportive & Palliative Care
- British Journal of Ophthalmology (BJO)
- British Journal of Sports Medicine (BJSM)
- Career Focus
- Drug and Therapeutics Bulletin (DTB)
- Emergency Medicine Journal (EMJ)
- Evidence Based Mental Health (EBMH)
- Evidence Based Medicine (EBM)
- Evidence Based Nursing (EBN)
- Frontline Gastroenterology
- Gut
- Heart
- Heart Asia
- Injury Prevention (IP)
- In Practice
- JAMIA
- Journal of Clinical Pathology (JCP)
- Journal of Epidemiology and Community Health (JECH)
- Journal of Family Planning and Reproductive Health Care (JFPRHC)
- Journal of Medical Ethics (JME)
- Journal of Medical Genetics (JMG)
- Journal of Neurology, Neurosurgery and Psychiatry (JNNP)
- Journal of NeuroInterventional Surgery
- Medical Humanities (MH)
- Occupational and Environmental Medicine (OEM)
- Postgraduate Medical Journal (PGMJ)
- Practical Neurology (PN)
- Sexually Transmitted Infections (STI)
- Student BMJ
- Thorax
- Veterinary Record
- Tobacco Control (TC)
- wjm western journal of medicine (archive)

Goals of the project

- Current tagging being inadequate on some products (notably the BMJ and the journals) and inconsistent
  - current tagging different for each product
  - content stored in different places
  - tagging not maintained
  - content cannot be easily extracted
- Can we come up with a generic solution that is not dependent on HighWire or any other commercial tool to index content across the group in order to provide subject specific portals
- Introduce search capability (replace Semio search engine)
- Link together content across journals and content types
- Easily create “specialty portals”
- Discover knowledge “gaps” in different parts of the content set
- Provide recommendations to services based on searched-for content

2011 content (a selection)



Specialty Portals

- Decide on key terms used to identify relevant articles
- Need to take into account how recent the article is (no set way to do this)
- We can construct a set of terms on which to base the portal

Rheumatology

Ankylosing spondylitis, Connective tissue disease, Fibromyalgia, Osteoarthritis, Osteoporosis, Rheumatoid arthritis, Sjogren's syndrome, Systemic lupus erythematosus,Vascularitis, Rheumatology speciality, rheumatologist, rheumatology department, juvenile arthritis

Diabetes

Diabetes Mellitus, Diabetes, Diabetes Mellitus, Insulin-Dependent, Diabetes Mellitus, Non-Insulin-Dependent (perhaps too simplistic) further analysis suggests Diabetic care, Diabetic, Diabetologist, Antidiabetic concepts

Two (three?) approaches

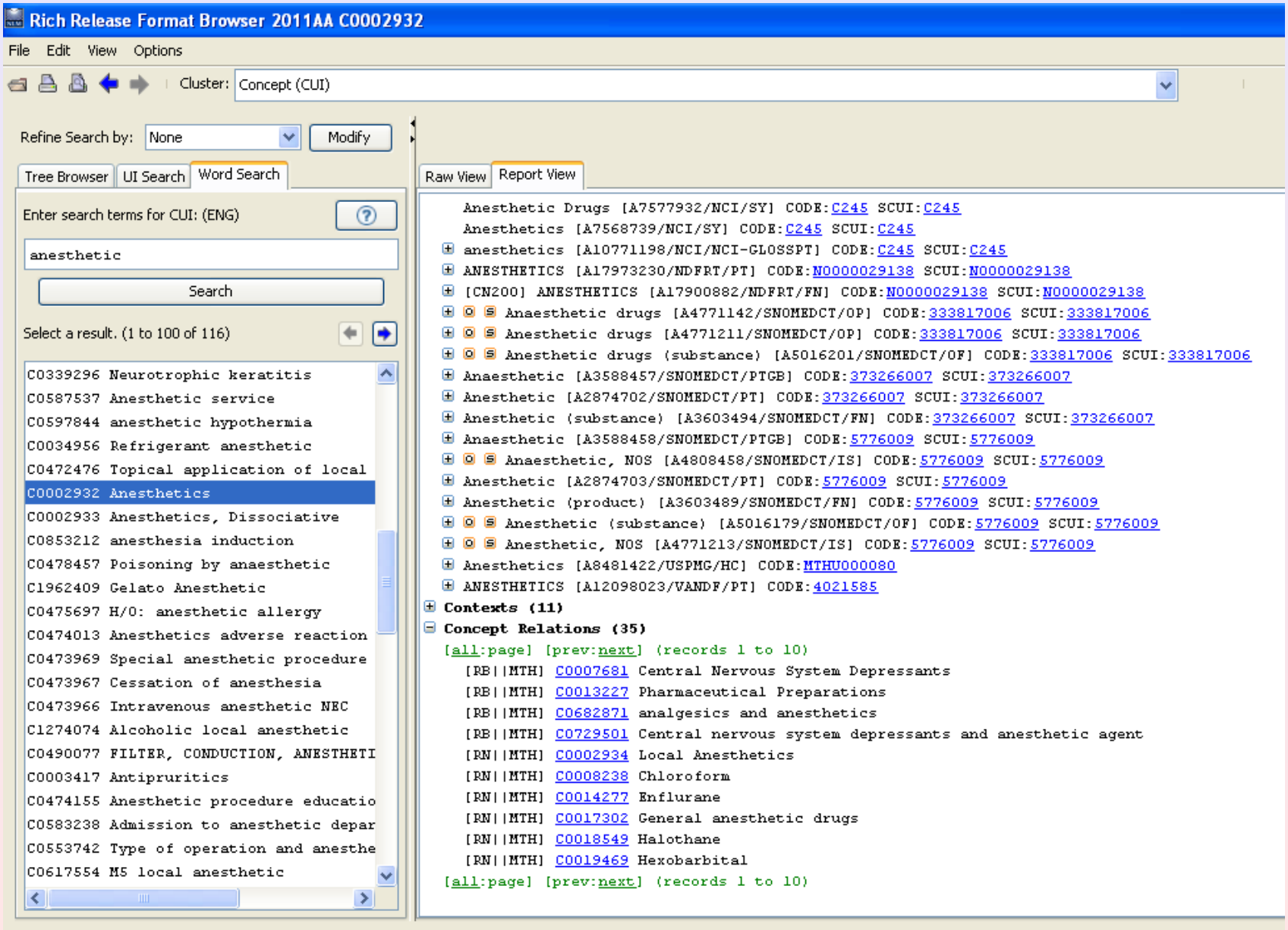
1. Semantic Method:
   Metathesaurus indexing tool “MetaMap” together with UMLS
2. Statistical Method:
   Use word/n-gram frequencies and clustering methods
3. Combination of the two?

UMLS

Unified Medical Language System (UMLS) Metathesaurus by National Library of Medicine (NLM)

- Compendium of many controlled vocabularies in the biomedical sciences (created 1986)
- Provides a mapping structure among these vocabularies
- Thesaurus and ontology of biomedical concepts
- NLP tools

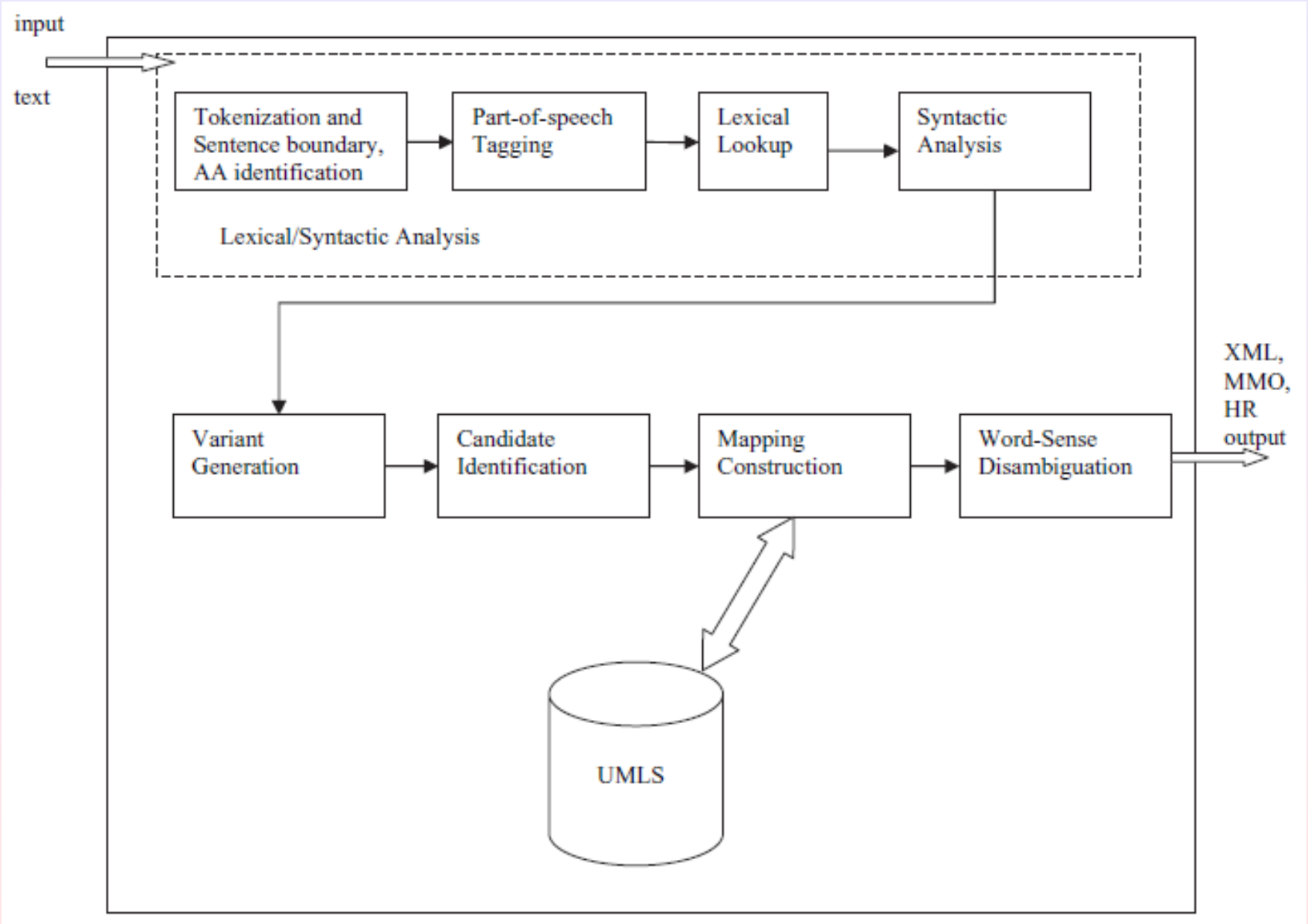
UMLS


Rich Release Format Browser 2011AA C0002932
File Edit View Options
Cluster: Concept (CUI)
Refine Search by: None
Modify
Tree Browser UI Search Word Search
Enter search terms for CUI: (ENG)
anesthetic
Search
Select a result. (1 to 100 of 116)
C0339296 Neurotrophic keratitis
C0587537 Anesthetic service
C0597844 anesthetic hypothermia
C0034956 Refrigerant anesthetic
C0472476 Topical application of local
C0002932 Anesthetics
C0002933 Anesthetics, Dissociative
C0853212 anesthesia induction
C0478457 Poisoning by anaesthetic
C1962409 Gelato Anesthetic
C0475697 H/O: anesthetic allergy
C0474013 Anesthetics adverse reaction
C0473969 Special anesthetic procedure
C0473967 Cessation of anesthesia
C0473966 Intravenous anesthetic NEC
C1274074 Alcoholic local anesthetic
C0490077 FILTER, CONDUCTION, ANESTHETI
C0003417 Antipruritics
C0474155 Anesthetic procedure educatio
C0583238 Admission to anesthetic depar
C0553742 Type of operation and anesthe
C0617554 M5 local anesthetic
Raw View Report View
Anesthetic Drugs [A7577932/NCI/SY] CODE:C245 SCUI:C245
Anesthetics [A7568739/NCI/SY] CODE:C245 SCUI:C245
anesthetics [A10771198/NCI/NCI-GLOSSPT] CODE:C245 SCUI:C245
ANESTHETICS [A17973230/NDFRT/PT] CODE:N0000029138 SCUI:N0000029138
[CN200] ANESTHETICS [A17900882/NDFRT/FN] CODE:N0000029138 SCUI:N0000029138
O S Anaesthetic drugs [A4771142/SNOMEDCT/OP] CODE:333817006 SCUI:333817006
O S Anesthetic drugs [A4771211/SNOMEDCT/OP] CODE:333817006 SCUI:333817006
O S Anesthetic drugs (substance) [A5016201/SNOMEDCT/OF] CODE:333817006 SCUI:333817006
Anaesthetic [A3588457/SNOMEDCT/PTGB] CODE:373266007 SCUI:373266007
Anesthetic [A2874702/SNOMEDCT/PT] CODE:373266007 SCUI:373266007
Anesthetic (substance) [A3603494/SNOMEDCT/FN] CODE:373266007 SCUI:373266007
Anaesthetic [A3588458/SNOMEDCT/PTGB] CODE:5776009 SCUI:5776009
O S Anaesthetic, NOS [A4808458/SNOMEDCT/IS] CODE:5776009 SCUI:5776009
Anesthetic [A2874703/SNOMEDCT/PT] CODE:5776009 SCUI:5776009
Anesthetic (product) [A3603489/SNOMEDCT/FN] CODE:5776009 SCUI:5776009
O S Anesthetic (substance) [A5016179/SNOMEDCT/OF] CODE:5776009 SCUI:5776009
O S Anesthetic, NOS [A4771213/SNOMEDCT/IS] CODE:5776009 SCUI:5776009
Anesthetics [A8481422/USPMG/HC] CODE:MTHU000080
ANESTHETICS [A12098023/VANDF/PT] CODE:4021585
Contexts (11)
Concept Relations (35)
[all:page] [prev:next] (records 1 to 10)
[RB||MTH] C0007681 Central Nervous System Depressants
[RB||MTH] C0013227 Pharmaceutical Preparations
[RB||MTH] C0682871 analgesics and anesthetics
[RB||MTH] C0729501 Central nervous system depressants and anesthetic agent
[RN||MTH] C0002934 Local Anesthetics
[RN||MTH] C0008238 Chloroform
[RN||MTH] C0014277 Enflurane
[RN||MTH] C0017302 General anesthetic drugs
[RN||MTH] C0018549 Halothane
[RN||MTH] C0019469 Hexobarbital
[all:page] [prev:next] (records 1 to 10)

MetaMap

- Tool by NLM for indexing documents against UMLS concepts [Aronson, 2001]
- Open Source but based on Sictus Prolog which is a commercial product, current cost 2100 Euros for a single license.
- Procedure:
  1. transforms the text in a document by limited syntactic analysis that recognizes simple noun phrases
  2. variant generation
  3. resulting phrases are matched to UMLS concepts, where possible, and then replaced with the preferred form of the matched concept
  4. “surrogate text” is then indexed with a retrieval system to allow query

MetaMap


input
text
Tokenization and Sentence boundary, AA identification
Part-of-speech Tagging
Lexical Lookup
Syntactic Analysis
Lexical/Syntactic Analysis
Variant Generation
Candidate Identification
Mapping Construction
Word-Sense Disambiguation
XML, MMO, HR output
UMLS

Semantic Method

- indexed about 12 months of BMJ content 2010 to Jan 2011 using MetaMap with the UMLS metathesaurus 2010AB
- Metamap often comes up with multiple matches. Use a bit of “black art” to select the best candidate
- Output is a table with a list of each UMLS entry found in an article, which are then scored using tf-idf
- Took about 36 hours on a desktop PC.
- 37k concepts used 1.3 million times
- Metamap may come up with more than match. More could be done to pick the most appropriate
- Metamap struggles with UK specific acronyms such as NHS, GPs. as they don’t exist with UMLS

Statistical Method

- Approach taken
  - Convert XML to plain text using XSLT
  - Tokenising
  - Acronym and Abbreviations: Schwartz & Hearst method [Schwartz and Hearst, 2003]
  - Stripping urls and emails
  - Stemming?
  - Dictionary creation using hashtables
  - TF-IDF
  - Cosine similarity
  - N-Grams
- Preliminary Results
  - Searching for terms
  - Specialty Portals

Vector Space Model (VSM)

A corpus of ℓ documents can be represented as a document-term matrix whose rows are indexed by the documents and whose columns are indexed by the terms. Each entry in position (i, j) is the term frequency of the term t_j in document i.

$$\mathbf{D} = \begin{pmatrix} tf(t_1, d_1) & \dots & tf(t_N, d_1) \\ | & X & | \\ tf(t_1, d_\ell) & \dots & tf(t_N, d_\ell) \end{pmatrix}.$$

From matrix $\mathbf{D}$, we can construct:

- the term-document matrix: $\mathbf{D}'$
- the term-term matrix: $\mathbf{D}'\mathbf{D}$
- the document-document matrix: $\mathbf{D}\mathbf{D}'$ ($\equiv$ gram or kernel matrix)

Disadvantages of VSM

- BOW is not able to map documents that contain semantically equivalent words into the same feature vectors.
- *e.g.* synonymous words which contain the same information, but are assigned distinct components.
- Complete loss of context information around a word
  - apply different weight w_i to each coordinate
  - remove uninformative words, called (stop words)
- Long documents contain more words than the short ones, and hence they are represented by feature vectors with greater norm

$$idf = \ln\left(\frac{\ell}{df(t)}\right).$$

tf-idf is able to highlight discriminative terms and downweight irrelevant terms, but can't take into account semantic information about two or more terms or about two or more documents. Semantic information can be introduced into the semantic kernel using the proximity matrix $\mathbf{P}$. This matrix needs to have non-zero off-diagonal entries, $\mathbf{P}_{ij} > 0$ for$i \neq j$, when the term i is semantically correlated with term j.

Acronyms & Abbreviations

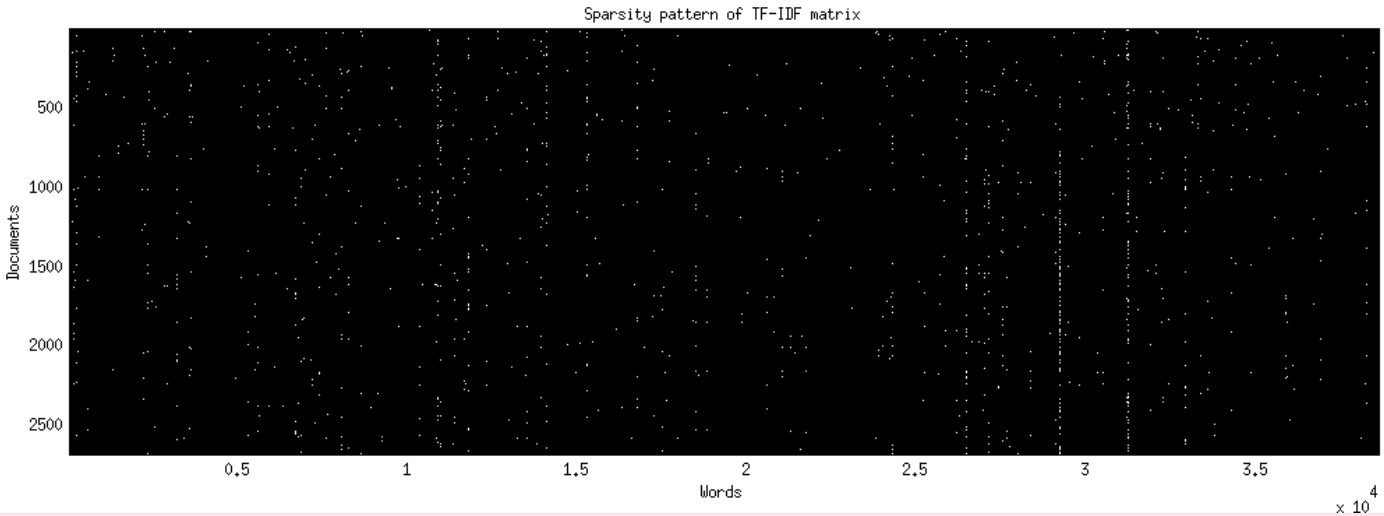
AA,Alcoholics Anonymous
ABPI,Association for the British Pharmaceutical industry
ABPI,Association of the British Pharmaceutical Industry
ABPI,ankle brachial pressure index
ACE,angiotensin converting enzyme
ACMD,Advisory Council on the Misuse of Drugs
ACPGBI,Association of Coloproctology of Great Britain and Ireland
ACT,artemisinin based combination therapy
ACT,artemisinin based combination treatment
ACTs,artemisinin based combination therapies
ADASS,Adult Social Services
ADHD,attention deficit hyperactivity disorder
AGREE,Appraisal of Guidelines Research and Evaluation
AHSC,academic health science centre
AMA,American Medical Association
AMD,age related macular degeneration
ANH,and hydration
ANU,Australian National University
APACHE,acute physiological and chronic health evaluation
APMS,alternative provider medical services
ART,assisted reproductive technology
ARVs,antiretrovirals
ASCO,American Society of Clinical Oncology
ASCOT,Anglo-Scandinavian cardiac outcomes trial
ASH,Action on Smoking and Health
ASR,articular surface replacement
Avandia,advisory committee on the antidiabetes drug rosiglitazone
Avite,Association of Victims of Thalidomide
BAT,British American Tobacco

...

- How to discover which AA's are referring to the same thing? Can use Leventshein distance but will miss some
- Some false positives
- Some will slip through the net too

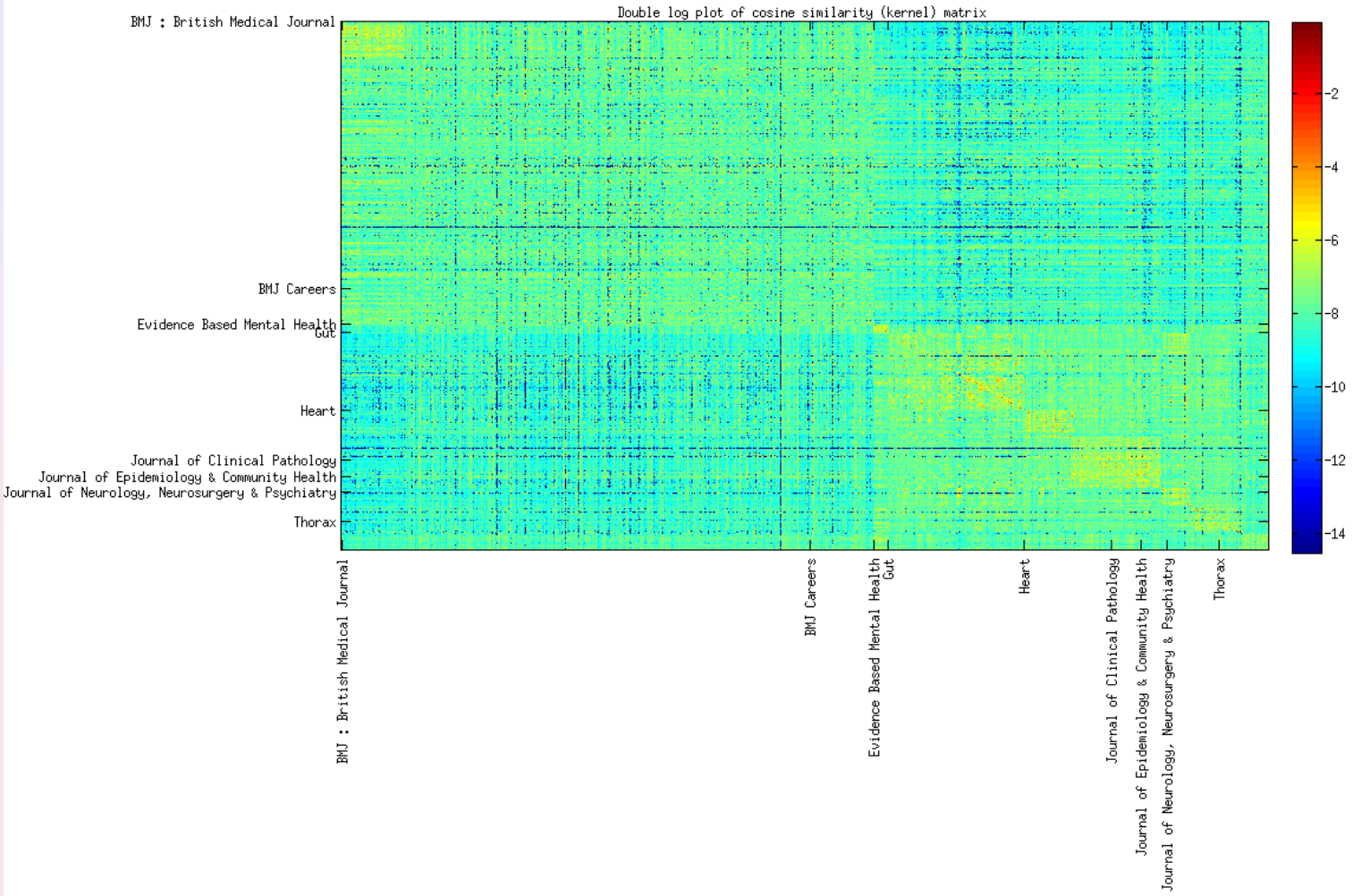
Sparsity

tf-idf matrix is sparse - enables efficient storage and processing



Typically $<$ 1% nonzero entries

Cosine-similarity matrix



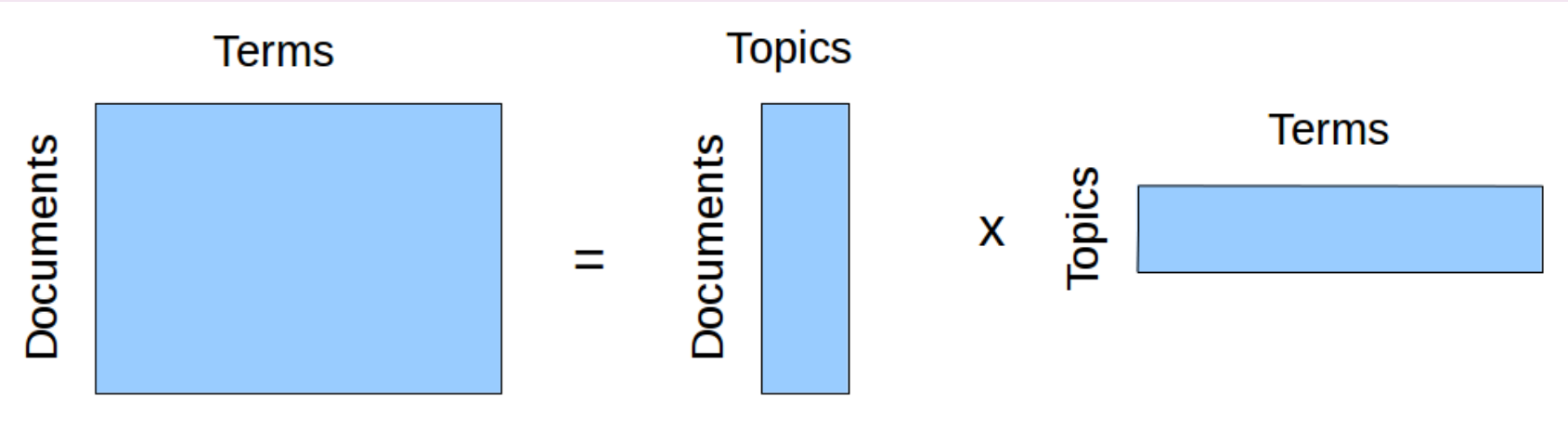
Gives an idea of homo-/hetero-geneity of content types

Clustering

- Hard Clustering: k-means
- Soft Clustering I: Non-negative Matrix Factorisation (NNMF) [Lee and Seung, 1999, Xu et al., 2003]

$$F(\mathbf{W}, \mathbf{H}) = \|\mathbf{X} - \mathbf{WH}\|_F^2$$

- Soft Clustering II: Latent Dirichlet Allocation (LDA) [Blei et al., 2003]



Clustering - Examples

k = 20; [IDX,C,sumd,D] = kmeans(tfidf, k, 'distance', 'cosine');
[CS CSI] = sort(full(C), 'descend'); dictcut(CSI(:,1:10))
ans =

Infectious dis.	Nat. Disasters	Alcohol/tabacco	organ donation	child abuse	cancer	general?	nhs trusts
'hiv'	'aid'	'alcohol'	'wife'	'mental'	'cancer'	'doctors'	'nhs'
'aids'	'humanitarian'	'tobacco'	'transplant'	'children'	'screening'	'doctor'	'trusts'
'malaria'	'haiti'	'smoking'	'organ'	'child'	'breast'	'medicine'	'commissioning'
'tuberculosis'	'gaza'	'ban'	'leaves'	'detention'	'women'	'children'	'care'
'tb'	'food'	'smoke'	'grandchildren'	'physical'	'cancers'	'back'	'gps'
'countries'	'sudan'	'drinking'	'donation'	'abuse'	'prostate'	'men'	'services'
'sex'	'cholera'	'minimum'	'organs'	'depression'	'survival'	'don'	'commission'
'treatment'	'international'	'cigarettes'	'retirement'	'psychiatric'	'cervical'	'patients'	'bn'
'infections'	'water'	'price'	'hospital'	'health'	'diagnosis'	'time'	'trust'
'million'	'earthquake'	'drink'	'royal'	'illness'	'ovarian'	'pain'	'quality'

Of course there difficulties with this approach:

- How to select the number of clusters
- How to decide what the "topic" of a cluster is
- How to incorporate new data (are clusters stable??)
- Clustering algorithm is quite expensive

Search Results

Search for ”diabetes”
1: [news/bmj.c3228.xml] (1782) More than half of diabetic patients do not get recommended annual tests
2: [news/bmj.c980.xml] (2676) Nearly 23000 people in England aged under 18 have diabetes, survey shows
3: [views/bmj.c1216.xml] (342) Bad medicine: type 2 diabetes

Search for ”nhs”
1: [news/bmj.c1823.xml] (1506) NHS hopes to repeat success of BBC Worldwide in health market
2: [news/bmj.b4029.xml] (857) Auditors find NHS financial performance good overall
3: [news/bmj.c2909.xml] (1724) NHS Confederation chief quits over plans for staffing services for trusts

Search for ”reform”
1: [news/bmj.c4330.xml] (2028) State of Virginia launches legal challenge to US health reforms
2: [news/bmj.c463.xml] (2071) Election of Republican senator derails US healthcare reform
3: [news/bmj.c2429.xml] (1629) Implementing health reform in the US is going to require huge change, briefing told

Search for ”nhs” + ”reform”
1: [news/bmj.c1823.xml] (1506) NHS hopes to repeat success of BBC Worldwide in health market
2: [news/bmj.c4330.xml] (2028) State of Virginia launches legal challenge to US health reforms
3: [news/bmj.c463.xml] (2071) Election of Republican senator derails US healthcare reform

Search for ”nhs” + ”reform” + ”-republican” + ”-congress”
1: [news/bmj.c1823.xml] (1506) NHS hopes to repeat success of BBC Worldwide in health market
2: [news/bmj.b4029.xml] (857) Auditors find NHS financial performance good overall
3: [news/bmj.c2909.xml] (1724) NHS Confederation chief quits over plans for staffing services for trusts

Issues with this method:

- Name (+places?) resolution, Acronym and abbreviation (AA) resolution both not included
- ’US’ became ’us’ (lowercase) and was removed as a stopword
- Synonym resolution - maybe ontology can help?

Neurology Portal I

Top terms in cluster:
'patients' 'mri' 'cases' 'clinical' 'seizures' 'ms' 'lesions' 'onset' 'diagnosis' 'years' 'epilepsy' 'study' 'neurological' 'eeg' 'brain' 'pnes' 'lesion' 'seizure' 'patient' 'abnormalities'

25 possible clusters found, best cluster index: 11, documents in cluster: 208

Method 1: Use top documents from cluster

Method 2: Perform search using augmented set of terms:
Search for "brain" + "cerebellum" + "cerebral" + "cns" + "coma" + "cord" + "cranial" + "creutzfeld-jakob" + "disease" + "disorders" + "epilepsy" + "headache" + "hydrocephalus" + "injury" + "intracranial" + "memory" + "migraine" + "motor" + "movement" + "multiple" + "muscle" + "nerve" + "nerves" + "neuroimaging" + "neurological" + "neurology" + "neuromuscular" + "neurone" + "neurooncology" + "palsy" + "parkinson" + "peripheral" + "pns" + "sclerosis" + "seizures" + "sleep" + "spinal" + "stem" + "stroke" + "trauma" + "variant" + "vcjd"

Search term "creutzfeld-jakob" not found in dictionary!
Search term "neurooncology" not found in dictionary!
Ignoring any missing search terms and continuing ...

Neurology Portal II

Method 1

Index (score):	Journal Name	Title
1 (0.116)	JNNP	Complementary roles of grey matter MTR and T2 lesions in predicting progression in ea
2 (0.112)	Thorax	Progression of idiopathic pulmonary fibrosis: lessons from asymmetrical disease
3 (0.111)	JNNP	Clinical classification of psychogenic non-epileptic seizures based on video-EEG analys
4 (0.108)	JNNP	Jakob disease: a retrospective analysis of the first 150 cases in the UK
5 (0.108)	JNNP	Age at onset predicts good seizure outcome in sporadic non-lesional and mesial tempor
6 (0.0941)	Thorax	Long-term follow-up high-resolution CT findings in non-specific interstitial pneumonia
7 (0.0935)	JNNP	Amygdalar enlargement in patients with temporal lobe epilepsy
8 (0.0863)	JNNP	T2 lesion location really matters: a 10 year follow-up study in primary progressive multip
9 (0.0855)	JNNP	CADASIL with cord involvement associated with a novel and atypical NOTCH3 mutation
10 (0.0825)	JNNP	Psychogenic seizures and frontal disconnection: EEG synchronisation study

Method 2

Index (score):	Journal	Title
1 (0.00783)	JNNP	Isolated abducens and facial nerve palsies due to a facial collicular plaque in multiple sc
2 (0.00683)	Gut	The role of the spinal cord in bowel dysfunction secondary to multiple sclerosis: a comp
3 (0.00672)	BMJ	Magnitude, impact, and stability of primary headache subtypes: 30 year prospective Sw
4 (0.00555)	BMJ	John Douglas Mitchell
5 (0.00553)	BMJ	Total hip replacement
6 (0.00542)	BMJ	Headache, migraine, and structural brain lesions and function: population based Epiden
7 (0.00538)	JNNP	Are subjects with spondylotic cervical cord encroachment at increased risk of cervical sp
8 (0.00529)	JNNP	How is stroke thrombolysis affecting neurology training?
9 (0.00525)	JNNP	Jakob disease?
10 (0.00523)	JNNP	All-night sleep organisation and distinction of schizophrenia, paranoid and affective psyc

(JNNP = Journal of Neurology, Neurosurgery & Psychiatry)

NHS Reform Portal I

Top terms in cluster:

'health' 'nhs' 'commissioning' 'care' 'government' 'gps' 'bill' 'public' 'services' 'local' 'gp' 'competition' 'consortiums' 'reforms' 'patients' 'providers' 'general' 'bma' 'committee' 'service'

59 possible clusters found, best cluster index: 48, documents in cluster: 497

Method 1: Simple Search:
Search for "nhs" + "reform" + "-republican" + "-american"

Method 2: Use top documents from cluster

NHS Reform Portal II

Method 1

Index (score):	Journal Name	Title
1 (0.0235)	BMJ	Is the NHS only a means of delivering healthcare?
2 (0.0235)	BMJ	Is the NHS only a means of delivering healthcare? (quick responses)
3 (0.0204)	BMJ	Reform
4 (0.0175)	BMJ Careers	NHS reforms to dominate LMCs conference
5 (0.0171)	BMJ Careers	Doctors slow to get flu vaccine
6 (0.0157)	BMJ	NHS managers add voice to calls for changes to health bill
7 (0.0157)	BMJ	What's happening to NHS spending across the UK?
8 (0.0156)	BMJ	NHS complexity, not bureaucracy, is issue in health bill
9 (0.0151)	BMJ	What's happening to NHS spending across the UK? (quick responses)
10 (0.0145)	BMJ Careers	Doctors' satisfaction with NHS and support for reform are lower than health profession a

Method 2

Index (score):	Journal	Title
1 (0.153)	BMJ	How the secretary of state for health proposes to abolish the NHS in England
2 (0.121)	BMJ	Can the government's proposals for NHS reform be made to work?
3 (0.116)	BMJ	NHS rethink: charade or cause for new hope?
4 (0.109)	BMJ	Challenges of EU competition law for general practice commissioning
5 (0.0997)	BMJ	We need to shake the bill to make sure it works for patients
6 (0.0964)	BMJ Careers	PMS SOS
7 (0.0897)	BMJ	Reaction: what they say about the health bill
8 (0.0736)	BMJ	Changes to NHS reforms will increase bureaucracy and treble the number of statutory c
9 (0.072)	BMJ	Dr Lansley's Monster
10 (0.066)	BMJ	BMA calls special representative meeting to discuss concerns over reforms

N-Grams

- N- contiguous terms (within sentence boundaries)
- Gives some contextual information
- Might help with cross-referencing to UMLS
- But ... combinatorial explosion

For the present corpus we have the following statistics:

Unique dictionary terms: 92966
Unique 1-grams: 1761381
Unique 2-grams: 2850714
Unique 3-grams: 3089736

Not actually done anything with these yet ....

Other Issues and Further Work I

- Still lots more content to analyse (more journals, other types of content) ...
- Amalgamation of the two approaches? *e.g.* use statistical methods but then cross-reference with UMLS, and then use the UMLS tree
- Negative Expressions: NegEx locates trigger terms indicating a clinical condition is negated or possible and determines which text falls within the scope of the trigger terms.
- Part-of-speech tagging
- Lexical lookup: Lexical lookup requires a morphological analyzer to associate each token with one or more readings. MetaMap uses the SPECIALIST lexicon. Lexical Variant Generation (lvg) written in Java looks like it'd do the job ok
  Syntactic analysis: This is needed to deal with partial matching
- Acronym & Abbreviation Detection: Although we're doing this, we're not yet using the information
- Variant generation (and suppression): Variants of words need to be generated. Increased precision can result if variants of 1-2 character words are suppressed.

Other Issues and Further Work II

- Word sense disambiguation (WSD): How to deal with Metathesaurus (MT) ambiguity, where two or more MT concepts share a common synonym (e.g. 'cold' appears in 6 different concepts: 'common cold', 'cold temperature' and 'cold sensation', 'cold therapy', 'cold brand of chlorpheniramin-phenylpropanolamine', 'chronic obstructive airway disease'). There are were 38266 manually edited suppressions of synonyms in the 2008AA release!! There is a WSD algorithm in the latest MetaMap, which chooses the most likely semantic type for a given context.
- Medical specific stopwords? And what to do about drug names etc?
- Output: XML and/or Java API
- Clustering
  - Currently we have to pre-specify the number of clusters
  - Non-parametric Bayesian approach (*e.g.* Hierarchical Dirichlet Processes (HDP)) [Teh et al., 2007]
- Online/Incremental Learning
  - Currently we are learning on a static corpus, but new content is generated at regular intervals (weekly for BMJ, continuously for Doc2Doc etc.)
  - Online NNMF, e.g. Cao *et. al.* at IJCAI [Cao et al., 2007]

Other Issues and Further Work III

- Online version of LDA (Matthew Hoffman, David Blei, and Francis Bach [Hoffman et al., 2010]) has been implemented in Vowpal Wabbit (John Langford's toolbox)
- Online verison of HDP using Stochastic E-M by David Blei and Yee Whye Teh (personal communication)

- Semi-Supervised Learning?
- User feedback (*e.g.* relevance)?
- Data storage
  - Use RDF triples?
  - Database format:
    - Relational database may not be suitable ...
    - Graphing database, such as Neo4j
    - Cassandra: Map/reduce possible with Apache Hadoop. Machine Learning tools available using Apache Mahout (k-means, LDA, dirichlet processes)

Selected References

Aronson, A. (2001).

Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).

Latent dirichlet allocation.
Journal of Machine Learning Research, 3(4-5):993–1022.

Cao, B., Shen, D., Sun, J.-T., Wang, X., Yang, Q., and Chen, Z. (2007).

Detect and track latent factors with online nonnegative matrix factorization.
In *Proceedings of IJCAI*, volume 2689–2694.

Hoffman, M. D., Blei, D. M., and Bach, F. (2010).

Online learning for latent dirichlet allocation.

Lee, D. D. and Seung, H. S. (1999).

Learning the parts of objects by non-negative matrix factorization.
Nature.

Schwartz, A. S. and Hearst, M. A. (2003).

A simple algorithms for identifying abbreviation definitions in biomedical text.
In *Proceedings of the Pacific Symposium on Biocomputing*, volume 8, pages 451–462.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2007).

Hierarchical dirichlet processes.
Journal of the American Statistical Association, 101(476):1566–1581.

Xu, W., Liu, X., and Gong, Y. (2003).

Document clustering based on non-negative matrix factorization.
In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–273.

Drug Names

A-Methapred (Methylprednisolone Sodium Succinate)
Abacavir Sulfate (Ziagen)
Abacavir Sulfate and Lamivudine Tablets (Epzicom)
Abacavir Sulfate, Lamivudine, and Zidovudine (Trizivir)
Abarelix (Plenaxis)
Abatacept (Orencia)
Abciximab (ReoPro)
Abelcet (Amphotericin B)
Abilify (Aripiprazole)
abiraterone acetate (Zytiga)
Ablavar (Gadofosveset Trisodium Injection)
Abobotulinumtoxin A Injection (Dysport)
Abraxane (Albumin-bound Paclitaxel)
Abstral (Fentanyl Sublingual Tablets)
Acamprosate Calcium (Campral)
Acanya Gel (Clindamycin Phosphate 1.2% and Benzoyl Peroxide 2.5%)
Acarbose (Precose)
Accolate (Zafirlukast)
Accretropin (Somatropin Injection)
AccuNeb (Albuterol Sulfate Inhalation Solution)
Accupril (Quinapril Hydrochloride)
Accuretic (Quinapril HCl/Hydrochlorothiazide)
Accutane (Isotretinoin)
Accuzyme (Papain and Urea)

...