# Skills for the future academic library

Wednesday, 28 November 2018

CILIP The library and information association

#CILIPFutureSkills

# Text mining - new tools, new skills

Julie Glanville

Associate Director, York Health Economics Consortium

Nov 2018

Providing Consultancy &
Research in Health Economics

UNIVERSITY of York  INVESTORS IN PEOPLE

YHEC
York Health Economics Consortium

# Agenda

- What is text mining?
- What is it being used for?
- Advantages and disadvantages of text mining
- Why should we be interested?

# Who am I?

- I am a qualified librarian
- I have been working in the field of systematic reviews for more than 25 years
  - Doing systematic reviews (SRs)
  - Researching information retrieval methods
  - Training in information retrieval
  - Trying to improve the efficiency of conducting SRs
- I am a co-convenor of the Cochrane Information Retrieval Methods Group and a co-author of the Cochrane Handbook chapter on searching for evidence
- My experience is mainly in health care but I have also worked on SRs in many other fields

# Why am I interested in text mining?

- Systematic reviews try to identify as much research as possible to answer specific questions
- This involves developing search strategies
- Sometimes the questions can be simple in term of search strategy design
  - Does one type of hip replacement work better than another in terms of length of time before a second replacement is needed?
- Sometimes questions can be really complex and difficult to search for:
  - Has providing families with discharge instructions/education earlier in admission improved readiness and patient outcomes?

# Are there tools that can help with information retrieval?

- Text mining techniques
- “Text mining is the process of discovering and extracting knowledge from unstructured data. This comprises three main activities:
  - Information retrieval (IR) to gather relevant texts.
  - Information extraction (IE) to identify and extract entities, facts and relationships between them.
  - Data mining to find associations among the pieces of information extracted from many different texts.
- …[TM] can help make the implicit information in your documents more explicit…”
- Source: National Centre for Text Mining http://www.nactem.ac.uk/faq.php?faq=1

# Value of TM, 1

- “Once an information retrieval system has been used to identify the documents that are relevant to a particular problem, these documents then need to be analysed by additional systems which use a mixture of natural language processing and data mining techniques to extract information and identify patterns in that information, **leading to the discovery of new knowledge**. The aim is not to improve the results of searching, but rather to help users **find information which previously may only have been discoverable by reading large numbers of documents**, or which was not in practice discoverable at all. “
- Source: NACTEM

# TM is not a single thing

- TM software comes in many flavours and can do many different things
  - Simple things – word frequency analysis – counting the numbers of times words appear in the text
  - More complex things – word co-occurrence – looking at patterns of words occurring together to identify concepts and relationships between words
  - Semantic analysis – analysing text according to the meaning of words not just their presence or absence
    - "89% of the group achieved smoking cessation"
    - "Five different smoking cessation interventions were explored"
  - Machine learning – relevant and irrelevant records

# What's new?

- Text analysis packages used for qualitative research such as Nudist have text mining elements
- Statistical software such as Simstat has text analysis features
- But the software is highly technical
  - did not lend itself so easily to everyday use in strategy development
- In recent years TM has become more usable by the less expert
- And many packages are available free of charge
- Exciting opportunities for IS

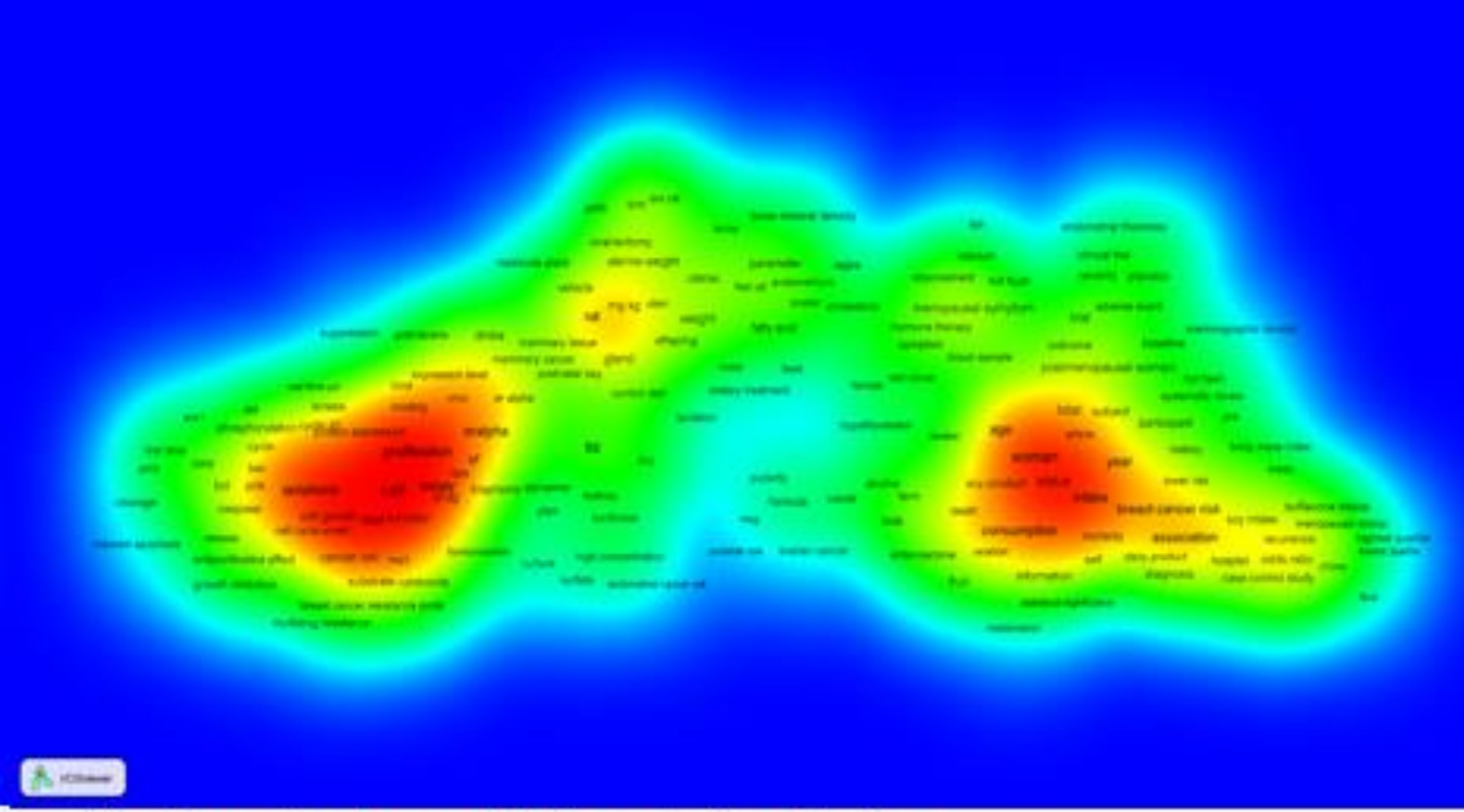
# What do these new “easy to use” tools have to offer?

- Identify concepts in the literature
- Analyse the frequency and relationship of words in records and texts
- Find additional relevant records from seed records
- Help out with other tasks such as record selection

- Free tools are growing in number and availability
- More sophisticated tools are available for purchase

# Types of text analysis tools

- Visual presentations of data within bibliographic records
  - What are the concepts within a literature?
  - Are there concepts I can try to remove from my search
- Word frequency analysis of records
  - What are the words that appear frequently in a set of records?
- Identifying phrases
  - To improve search precision
- Identifying subject headings used within database records
- Identifying trends in word usage
- Identifying research teams



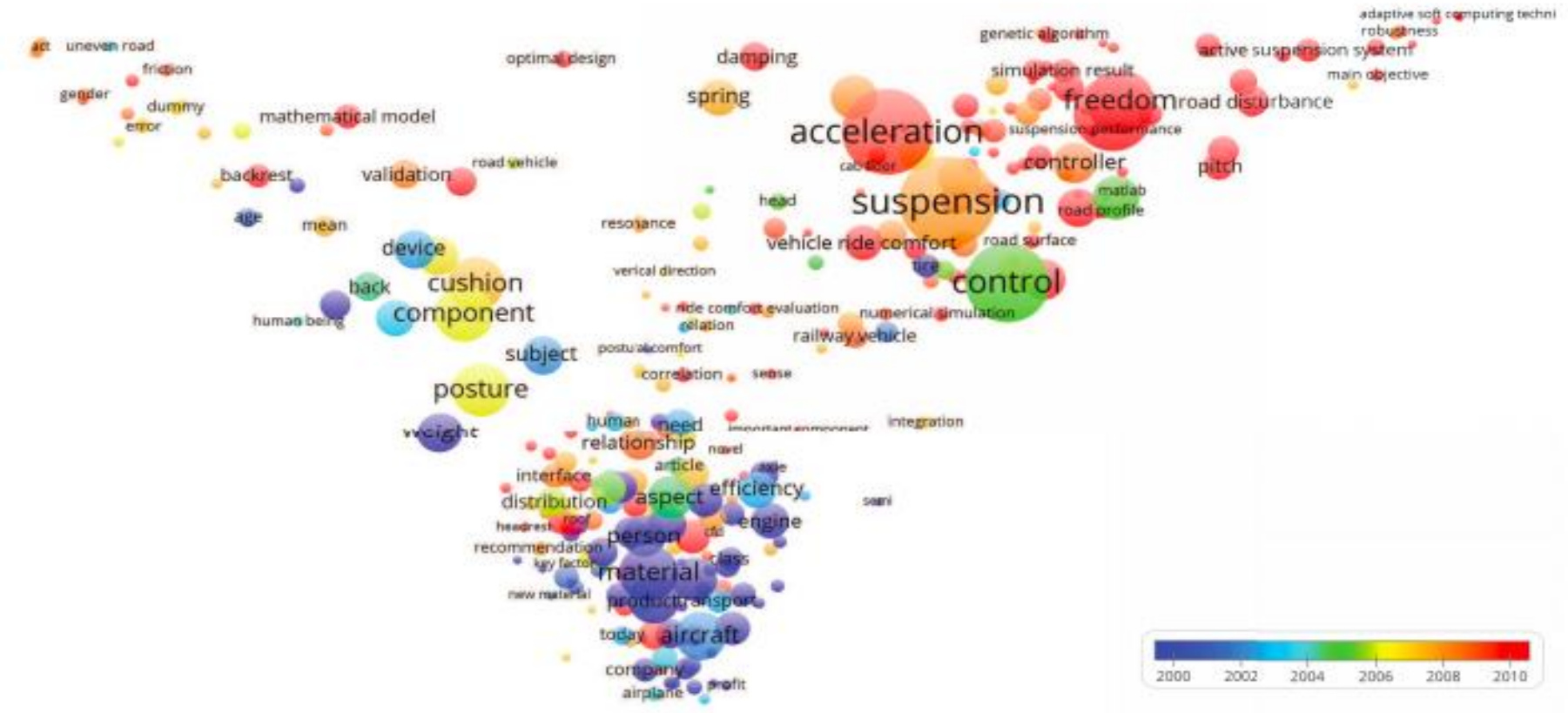
VosViewer: heat map of 500 records from Medline
Isoflavones impact on breast/thyroid/uterus

Vos viewer (http://www.vosviewer.com/)


VOSviewer example

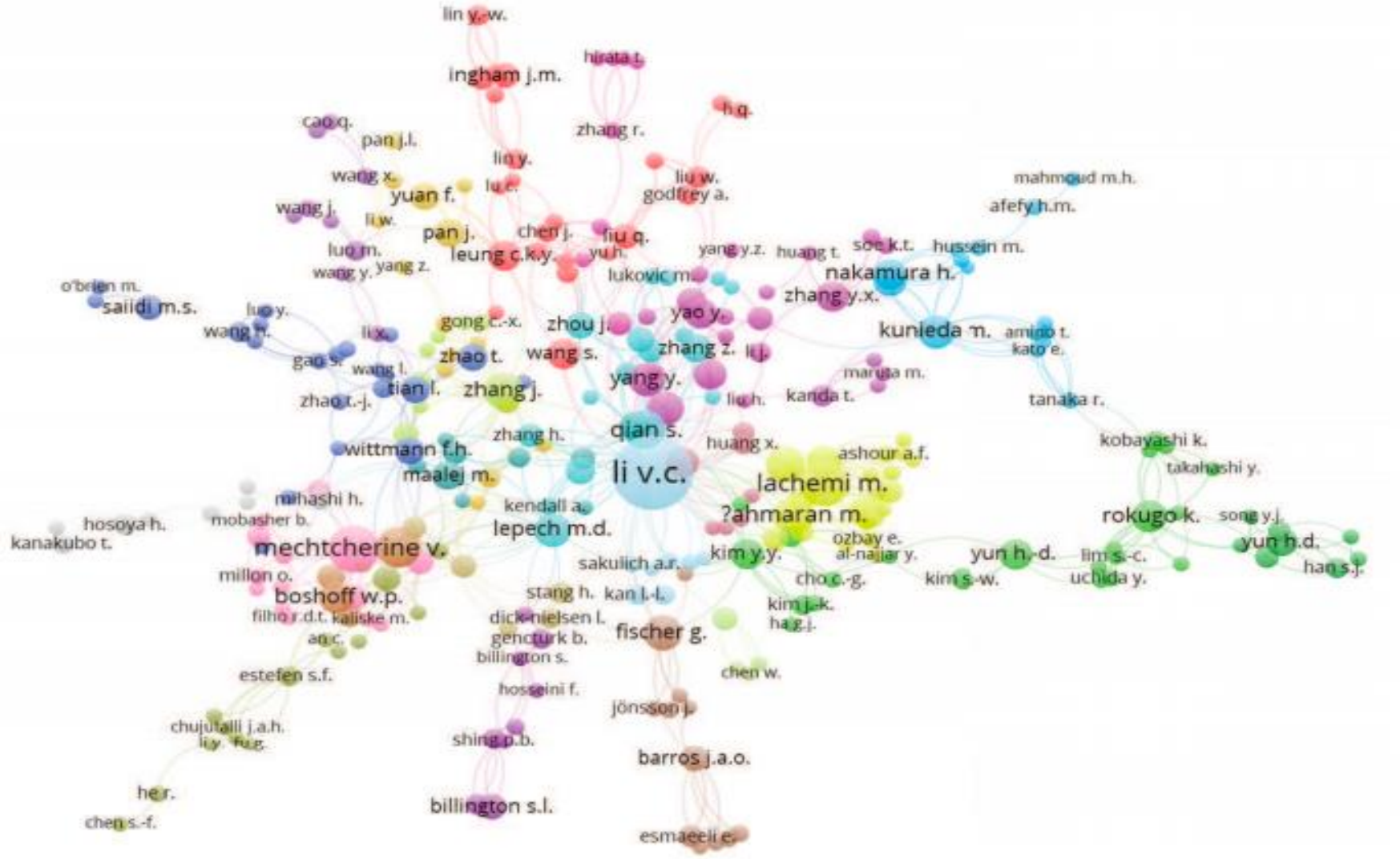
"Recruitment" – this shows how recruitment is a common term in molecular biology as well as recruitment to trials

# Evolution of research



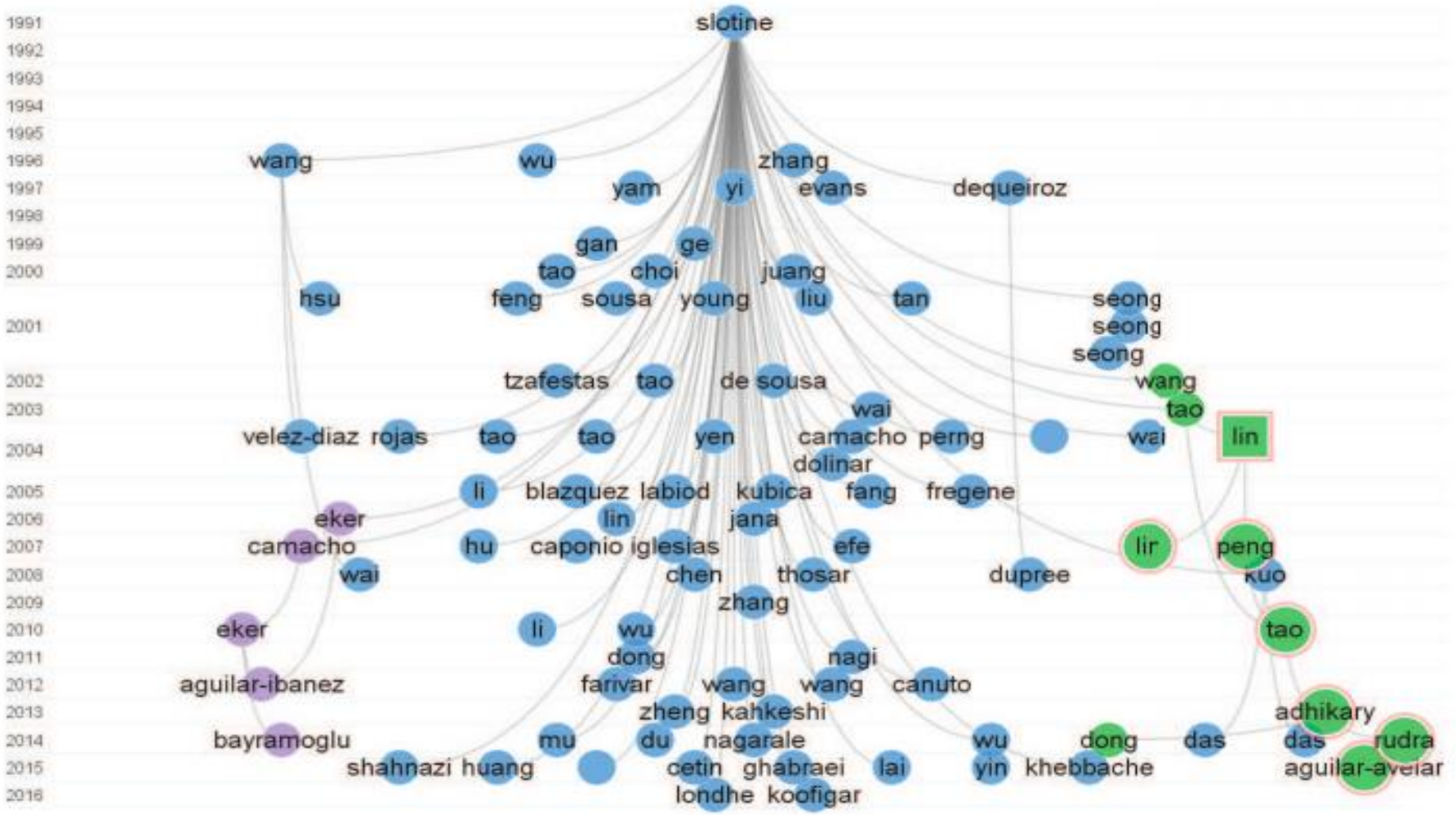

This term map shows the state of the art in research on passenger comfort. The overlay color of each circle corresponds to the average publication year of all the papers that include the corresponding term. In this map, the terms with cold colors (e.g. blue) represent the research activities with older average publication year and the terms with hot colors (e.g. red) show the terms with more recent average publication year.

# Collaboration profile




This co-authorship map shows the names of authors who are publishing papers in a specific domain. The size of the circles corresponds to the number of papers each author in the publication list has published, and the links between the circles show co-authorships of papers.

# Citation networks (1)




This citation network illustrates a set of books and papers published between 1991 and 2016. The top circle represents the book Applied Nonlinear Control, written by J.J. Slotine and W. Li in 1991. The other circles represent the "successor publications" of the book that were published in the following years, have cited the book and have received a high number of citations. Note that the vertical axis represents the year of publication.

# EPPI-Reviewer: Clustering

- EPPI-reviewer is a subscription-based SR package
- Offers text mining options built into a SR packge
  - Categorisation of records into clusters based on the similarity of the words within them
  - Significant terms are extracted and used to code the records by theme. Codes are often arranged in a tree structure to show the relationship between them to be investigated
  - Seen as a tool to increase precision by highlighting irrelevant themes in a set of search results that a later iteration of a strategy can try and exclude (Stansfield et al. 2017)
  - But could also increase sensitivity? Highlight themes we are unaware of that should be searched for explicitly?

- Lingo3G clusters
  - Sun Screen
  - Risk Factors
  - Skin Cancer Prevention
  - UV Exposure
  - Malignant Melanoma
  - Sun Safety
  - Vitamin D Deficiency
  - Clinical Trial
  - Health Communication
  - Quality of Life
  - Side Effects
  - Head and Neck
  - Preliminary
  - Illness
  - Accuracy
  - Regulation
  - Density
  - Failure
  - Uptake
  - Tumor Thickness
    - Western Australia
    - Gene Expression
    - Rural Areas
    - Surface Area
    - Barrier Function
    - Kidney Transplant Recipients
    - Indoor Tanning Facilities
    - Retrospective Review
    - Other Topics

Within this cluster we find more subthemes

YHEC

York Health Economics Consortium

- Health Communication
  - Sun Protection
  - Melanoma Risk
  - Behaviour
  - Risk Perception
  - Program
  - Risk Status
  - Randomized Controlled Trial
  - Skin Self-examination
  - Mass Media
  - Systematic Review
  - UV Radiation
  - (Other topics)

Within this cluster we find the records

YHEC
York Health Economics Consortium

Testing a theory-based health communication program: a replication of Go Sun Smart in outdoor winter recreation.

Effects of web-based intervention on risk reduction behaviors in melanoma survivors.

Dermatology on YouTube.

Randomized trial testing a worksite sun protection program in an outdoor recreation industry.

User-centered development of a smart phone mobile application delivering personalized real-time advice on sun protectior

Validation of sun exposure and protection index (SEPI) for estimation of sun habits.

Prevalence of health promotion policies in sports clubs in Victoria, Australia.

Perceived Relevance of Educative Information on Public (Skin) Health: Results of a Representative, Population-Based Tele

Implementing an Internet-Delivered Skin Cancer Genetic Testing Intervention to Improve Sun Protection Behavior in a Div

Visual feedback of individuals' medical imaging results for changing health behaviour.

The impact of communicating genetic risks of disease on risk-reducing health behaviour: systematic review with meta-ana

Policy interventions implemented through sporting organisations for promoting healthy behaviour change.

Examining direct and indirect pathways to health behaviour: the influence of cognitive and affective probability beliefs.

Social marketing and the creative process: staying true to your social marketing objectives.

Are the arts an effective setting for promoting health messages?.

Melanoma survivors: health behaviors, surveillance, psychosocial factors, and family concerns.

Policy interventions implemented through sporting organisations for promoting healthy behaviour change.

Effects of program exposure and engagement with tailored prevention communication on sun protection by young adolesc

Impact of a sun protection campaign in Medellin (Colombia).

Does personalized melanoma genomic risk information trigger conversations about skin cancer prevention and skin exami

Employee factors associated with interest in improving sun protection in an Australian mining workforce.

Appearance matters: the frame and focus of health messages influences beliefs about skin cancer.

Cancer knowledge and disparities in the information age.

Increasing sun protection in winter outdoor recreation a theory-based health communication program.

Dissemination of go sun smart in outdoor recreation: effect of program exposure on sun protection of guests at high-altitu

Development of an Educational Program Integrating Concepts of Genetic Risk and Preventive Strategies for Children with

# Voyant Tools

- http://voyant-tools.org/
- Paste in a file of database records or a document or a project summary
- Click on Reveal

Cirrus | Terms | Links | Reader | TermsBerry | Trends | Document Terms


Terms:

**Voyant 830 records**

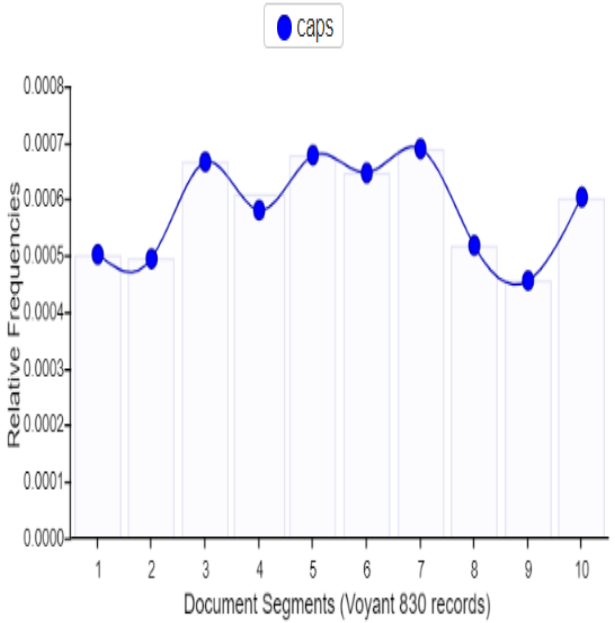
1. Front Physiol. 2017 Sep 21;8:727. doi: 10.3389/fphys.2017.00727. eCollection 2017.

Editorial: Regulation of Endurance Performance: New Frontiers.

Hettinga FJ(1), Renfree A(2), Pageaux B(3), Jones HS(4), Corbett J(5), Micklewright D(1), Mauger AR(6).

Author information:

(1)School of Sport, Rehabilitation and Exercise Science, University of EssexColchester, United Kingdom




Reset | Display

Summary | Documents | Phrases

| Term | Count | Length | Trend |
|---|---|---|---|
| cap with 70 alcohol | 2 | 4 | |
| cap were not | 2 | 3 | |
| cap was recorded | 2 | 3 | |
| cap was not | 2 | 3 | |

cap X | 142 | Length | Overlap

Contexts | Bubblelines | Correlations

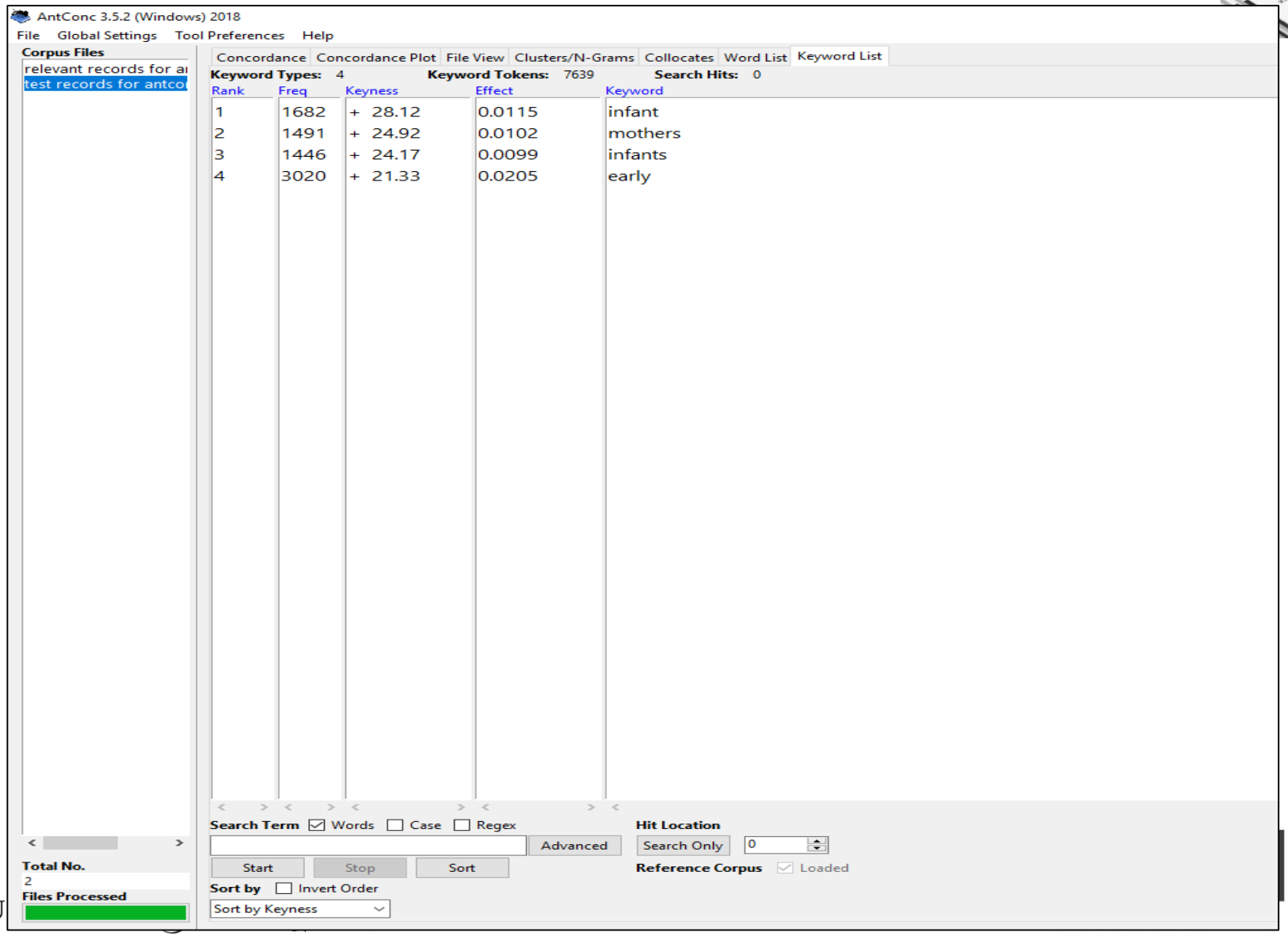
| Document | Left | Term | Right |
|---|---|---|---|
| 1) Voya... | of a simplified arched slender | cap | under intraluminal pressure and supported |
| 1) Voya... | of special interest: the skull | cap | is not open, yet the |
| 1) Voya... | of special interest: the skull | cap | was not cut, yet the |
| 1) Voya... | to test whether a similar | cap | recognition mechanism had independen... |

cap X | 791 context | expand

# Antconc

- Free software that can be downloaded
- Text analysis, phrase analysis, collocation
- Also identifying words that discriminate know relevant records from a more general set of records:
  - You can use your known relevant records
  - To match against a set of other records (that might contain relevant records)
- This may reveal the discriminating words
  - Words that are typical of relevant records but not of a wider set of records

# Antconc results



# Finding Additional Relevant Records

- Medline Ranker
- http://cbdm-01.zdv.uni-mainz.de/~jfontain/cms/?page_id=4
- Use known set of relevant records to find similar records
- Find a set of other records (or use built in options):
  - discharge AND family  (8,926 records)
- Add the PubMed identifiers (accession numbers) from both lists to the program
- It puts the records in order of similarity to the words in the test set collection
- Making use of MeSH as well

# Data mining tools

Mine abstracts, genes, chemicals and diseases!

Medline Ranker | Génie | Alkemio | Gene set to diseases | Phenotypes | Trends | Support

## Medline Ranker

**The query topic (the training set) is defined by:**

- ○ the following PubMed query
- ○ all the following MeSH terms (MeSH browser)
- ◉ the following list of PMIDs

```
OR 18705725  OR 15293587
```

*one per line*

**Examples**: Mitochondria, Neoplasms, Stem Cells, Alzheimer Disease, Computational Biology, Randomized Controlled Trials as Topic

**The abstracts to be ranked (the test set) are defined by:**

- ○ the training set
- ○ the background set
- ○ 10 000 randomly chosen recent abstracts
- ○ publications of the last 2 ▾ month(s)
- ○ the 1 ▾ -year(s) old abstracts
- ◉ the following list of PMIDs

```
19975646
```

*one per line*

Rank it | Reset

## Parameters

| | |
|---|---|
| The training set: | 9 / 19 abstracts. |
| The background set: | the profile of the whole Medline database. |
| The test set: | 8926 abstracts. |
| Other: | Scoring scheme: Bayes, Min word weight: 0.5, Target database: medline |
| Initialisation: | *2 seconds.* |

Top - Results - Discriminative words - Download

## Results

Processing 8349 abstracts: 0% - 9% - 19% - 29% - 39% - 49% - 59% - 69% - 79% - 89% - 99% - done.

*Ranked in 3 seconds.*

*colors: yellow for training set pmids, green for background set pmids,* and brown for *discriminative words* Color code: high low weight.

*Click on the pmid to read the abstract with highlighted discriminative words*

| Rank | PMID | Abstract Title | P-value |
|---|---|---|---|
| 1 | 11034682 | Discharge planning from hospital to home . | 4.52e-05 |
| 2 | 10796533 | Hospital -at-home versus in-patient hospital care . | 5.21e-05 |
| 3 | 25128468 | Quality care outcomes following transitional care interventions for older people from hospital to home : a systematic review. | 6.44e-05 |
| 4 | 26108207 | Protocol for a randomised controlled trial of an outreach support program for family carers of older people discharged from hospital . | 7.02e-05 |
| 5 | 27911489 | Palliative care interventions in advanced dementia. | 7.10e-05 |
| 6 | 12137636 | Education for contraceptive use by women after childbirth. | 7.22e-05 |
| 7 | 27819973 | Association between pacifier use and breast-feeding, sudden infant death syndrome, infection and dental malocclusion. | 7.23e-05 |
| 8 | 26222129 | Education for contraceptive use by women after childbirth. | 7.32e-05 |
| 9 | 28499065 | Not sick enough: Experiences of carers of people with mental illness negotiating care for their relatives with mental health services . | 7.39e-05 |
| 10 | 23076908 | Family -centred care for hospitalised children aged 0-12 years. | 7.61e-05 |
| 11 | 26154426 | Early discharge with home support of gavage feeding for stable preterm infants who have not established full oral feeds. | 7.63e-05 |
| 12 | 27820495 | Effectiveness of strategies to promote safe transition of elderly people across care settings. | 7.86e-05 |
| 13 | 17636778 | Effectiveness of shared care across the interface between primary and specialty care in chronic disease management. | 7.87e-05 |
| 14 | 21631747 | Association between pacifier use and breast-feeding, sudden infant death syndrome, infection and dental malocclusion. | 7.91e-05 |
| 15 | 22986378 | Transitional care after hospitalization for acute stroke or myocardial infarction: a systematic review. | 8.12e-05 |
| 16 | 22895923 | Education for contraceptive use by women after childbirth. | 8.13e-05 |
| 17 | 10796830 | Education for contraceptive use by women after childbirth. | 8.14e-05 |
| 18 | 24153026 | A cluster randomised controlled trial and economic evaluation of a structured training programme for caregivers of inpatients after stroke : the | 8.18e-05 |

# Other uses of TM

- Citation analysis
- Research team identification and inter-relationships
- Research impact
- Record prioritisation
- Research development patterns
- Emerging topics

# Value of TM, 1

- Faster and more consistent than I am in analysing large volumes of records
- “Objective” to a certain extent
- Possible to document the steps involved
- Many software packages available for free
- Useful at various different stages of research process
  - Question definition/disambiguation
  - Strategy development
  - Finding additional records
  - Prioritising most relevant records

# Value of TM, 2

- Health care SR community are appreciating the value of TM
  - More papers on TM and machine learning (ML) being presented at conferences (Cochrane Colloquium and HTAi)
  - More use of TM free tools
  - Also increased facilities within database interfaces and SR software
    - Elsevier introducing simple text mining to Embase.com
    - Ovid experimenting with simple text mining
    - EPPI-reviewer and DistillerSR packages include tex tmining
  - Machine learning techniques embedded within study identification processes for Cochrane CENTRAL
  - Major health technology assessment agencies have commissioned surveys of use and uptake of TM (AHRQ and CADTH)

# Challenges of TM, 1

- Defining what aspect of TM you want to use and how
- May need some element of record processing to get just the fields you want
- Lack of standardisation – no agreed single approach
- Different systems use different algorithms
- How many iterations of strategy development should  involve text mining?
- Subjectivity remains
  - We need to make a series of choices to arrive at final results
  - If needed, developing taxonomies/ontologies can take time

# Challenges of TM, 2

- The more sophisticated the software, the more challenging to learn
- Depending on how involved you want to be - investment in acquiring
  - Knowledge of key issues in linguistics
  - Getting the best out the more complex software
  - For more complex needs/software likely to need specialised support (buy-in) or build in-house capacity
- Building relationships with your computing and linguistic colleagues

# Summary

- Many useful, free text tools available
- For more sophisticated projects a range of more sophisticated commercial packages are available
- TM has applications to a range of research support and monitoring activities (not just search strategies)
- Uptake of TM in specialist areas such as healthcare SRs is moving fast
- Opportunities for information professionals:
  - multi-disciplinary working with colleagues in computer services and linguistics
  - providing knowledge about TM options and resources to researchers
  - providing TM support services

# Selected publications


- O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015;4: 5
- Paynter RA, Bañez LL, Berliner E, Erinoff E, Lege-Matsuura J, Potter S, et al. EPC Methods: An Exploration of the Use of Text-Mining Software in Systematic Reviews. Research White Paper. Rockville, MD: Agency for Healthcare Research and Quality; 2016. Available: https://effectivehealthcare.ahrq.gov/ehc/products/625/2214/text-mining-report-160419.pdf
- van Eck NJ, Waltman L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. Scientometrics. 2017;111: 1053–1070
- Glanville J, Wood H. Text Mining Opportunities: White Paper. Ottawa: CADTH; 2018 May. https://www.cadth.ca/sites/default/files/pdf/methods/2018-05/MG0013_CADTH_Text-Mining_Opportunites_Final.pdf

# Thank You

julie.glanville@york.ac.uk

Telephone: +44 1904 324832

Website: www.yhec.co.uk

http://tinyurl.com/yhec-facebook

http://twitter.com/YHEC1

http://www.linkedin.com/company/york-health-economics-consortium

http://www.minerva-network.com/

Providing Consultancy &
Research in Health Economics

UNIVERSITY of York  INVESTORS IN PEOPLE

YHEC
York Health Economics Consortium

# Skills for the future academic library

Wednesday, 28 November 2018

CILIP The library and information association

#CILIPFutureSkills