Bibliothèque
universitaire de médecine

Mine and combine

Text Mining Tools Used for Search Term Identification
Meet and Greet

Cécile Jaques
Jolanda Elmers

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Overview

- Brief introduction and presentation

  Focus on PubMed PubReMiner
TerMine
Yale MeSH Analyzer

- Hands-on session
- Discussion together
- More tips and tricks from our experience

Questions

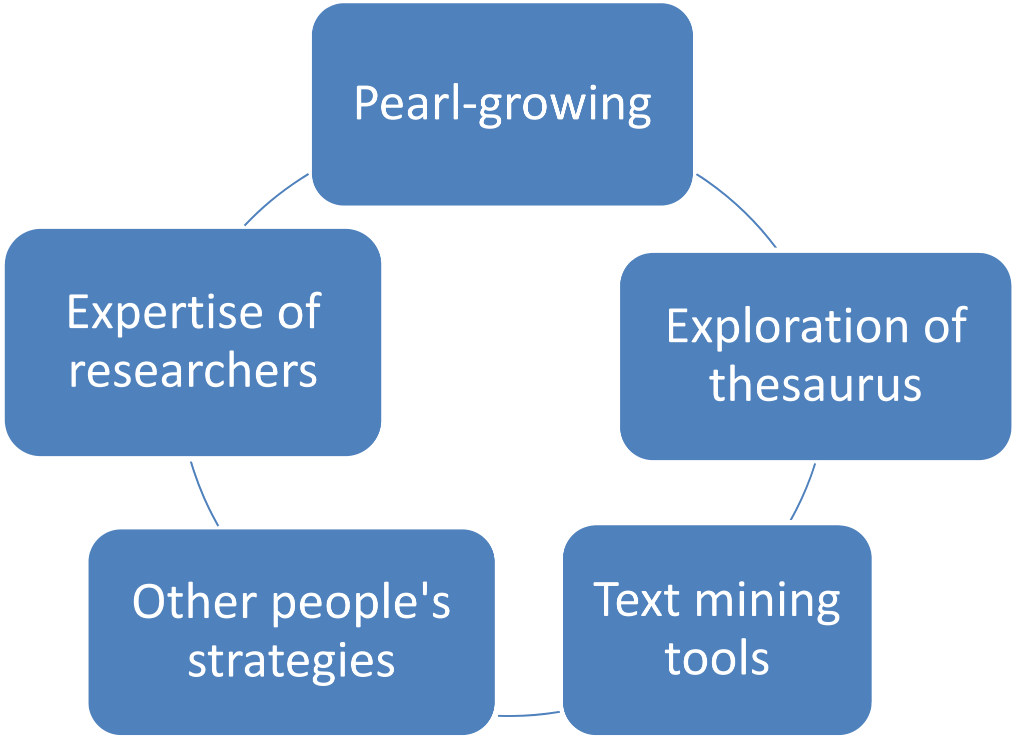
Who has been involved in developing a search strategy for a systematic review?

Who has already tested the following tools (or one of them)?

PubMed PubReMiner
TerMine
Yale MeSH Analyzer

Who uses these tools (or one of them) on a regular basis?

Search term identification methods



Pearl-growing

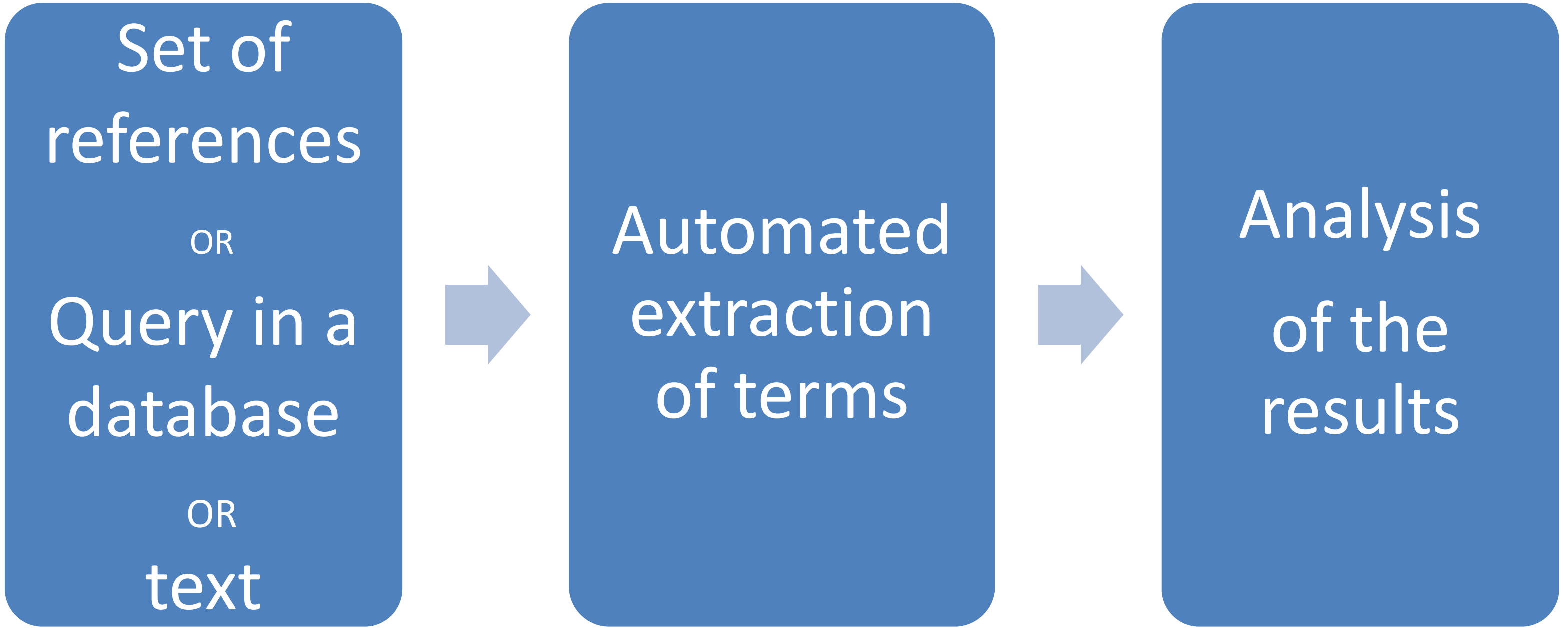
Exploration of thesaurus

Text mining tools

Other people's strategies

Expertise of researchers

Text mining tools



Set of references

OR

Query in a database

OR

text

Automated extraction of terms

Analysis of the results

Different programmes offer different levels of analysis

Overview of the tools

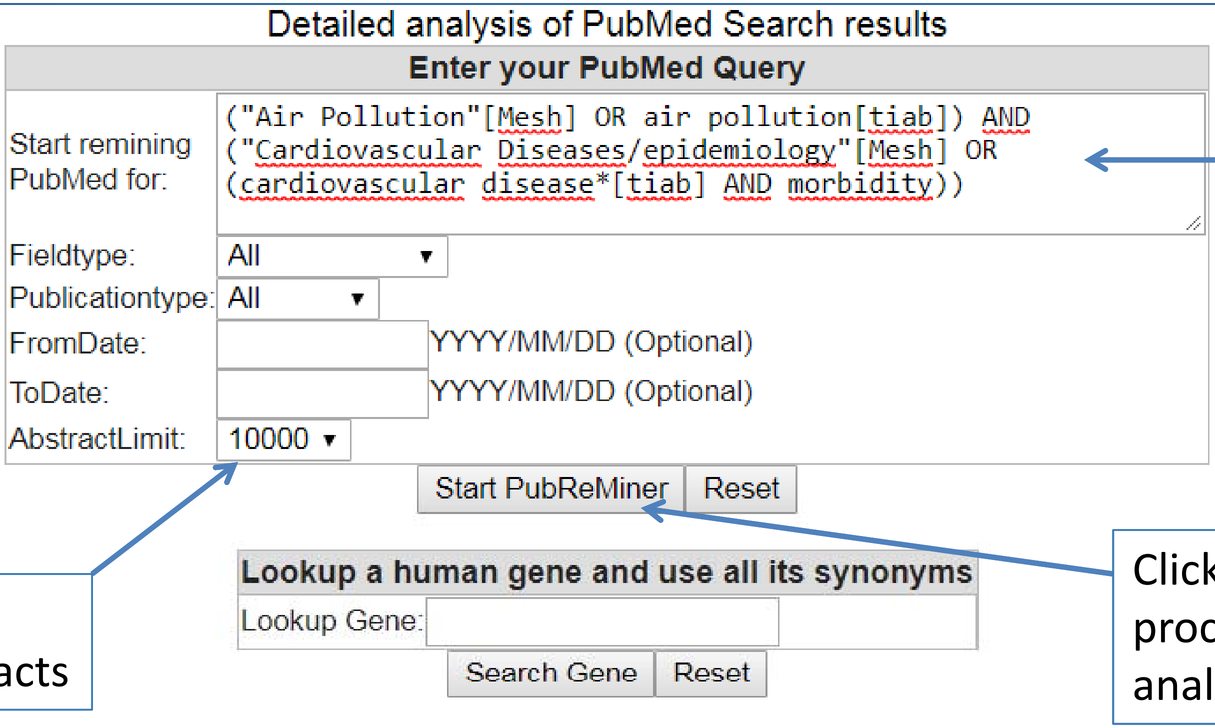
PubMed PubReminer	Frequency count of words in a set of articles, based on a query in PubMed
Yale MeSH Analyzer	MeSH analysis grid to identify MeSH Terms from a set of PubMed articles
Termine	Word combinations are extracted from raw text from any database, based on statistical and linguistic analysis

PubMed PubReminer

- Uses frequency tables for terms to display results of search queries in PubMed

- Offers a quick view of the words which are occurring most frequently in your results set

PubMed PubReminer



Detailed analysis of PubMed Search results

Enter your PubMed Query

Start remining PubMed for:	("Air Pollution"[Mesh] OR air pollution[tiab]) AND ("Cardiovascular Diseases/epidemiology"[Mesh] OR (cardiovascular disease*[tiab] AND morbidity))
Fieldtype:	All ▾
Publicationtype:	All ▾
FromDate:	YYYY/MM/DD (Optional)
ToDate:	YYYY/MM/DD (Optional)
AbstractLimit:	10000 ▾

Lookup a human gene and use all its synonyms

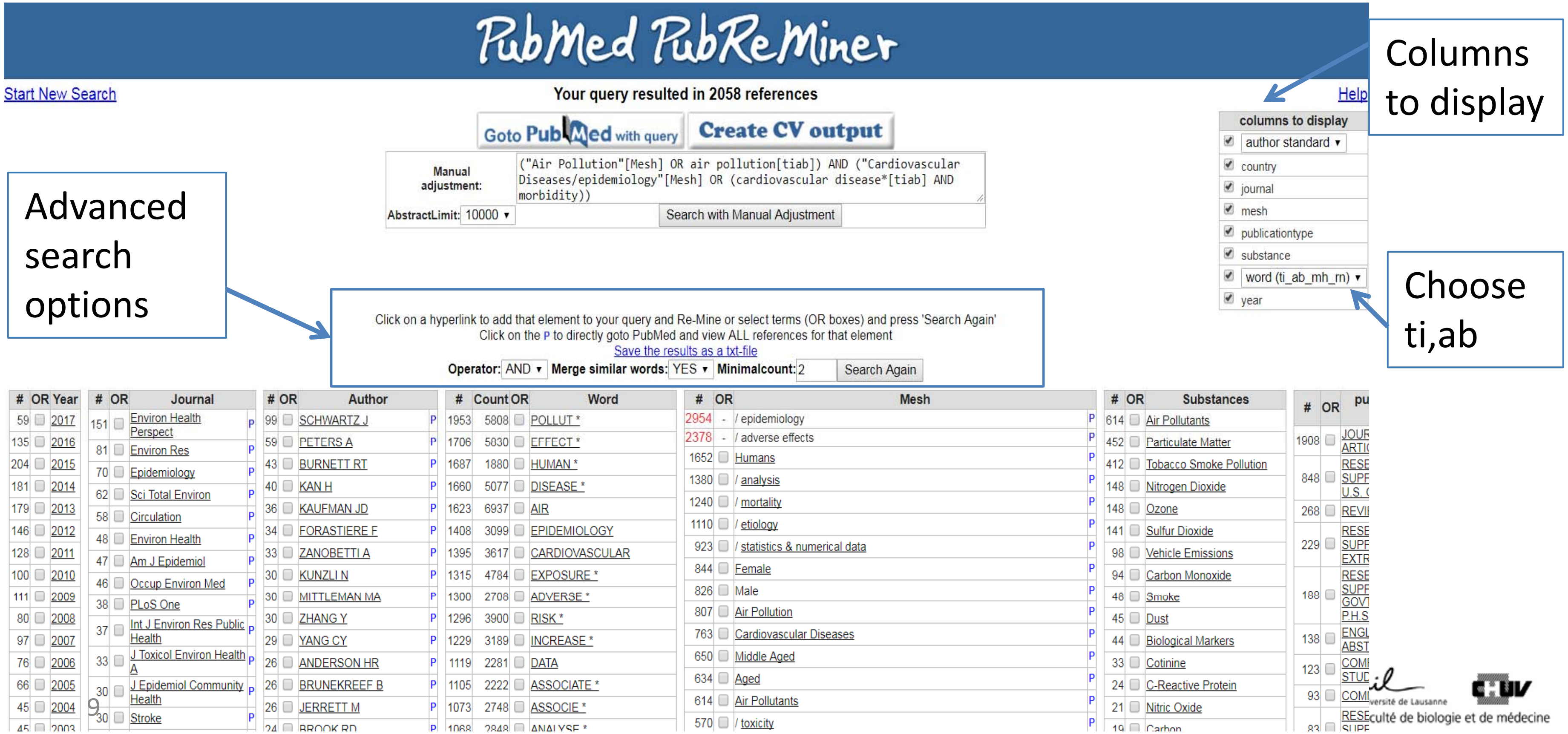
Lookup Gene:

Any query that can be processed by PubMed

Up to 10 000 abstracts

Click on start to proceed to the analysis

PubMed PubReminer - results



Your query resulted in 2058 references

Manual adjustment: ("Air Pollution"[Mesh] OR air pollution[tiab]) AND ("Cardiovascular Diseases/epidemiology"[Mesh] OR (cardiovascular disease*[tiab] AND morbidity))

AbstractLimit: 10000 ▾ Search with Manual Adjustment

columns to display
- ☑ author standard ▾
- ☑ country
- ☑ journal
- ☑ mesh
- ☑ publicationtype
- ☑ substance
- ☑ word (ti_ab_mh_rn) ▾
- ☑ year

Columns to display

Advanced search options

Choose ti,ab

Click on a hyperlink to add that element to your query and Re-Mine or select terms (OR boxes) and press 'Search Again'
Click on the P to directly goto PubMed and view ALL references for that element
Save the results as a txt-file

Operator: AND ▾ Merge similar words: YES ▾ Minimalcount: 2 Search Again

#	OR	Year
59	□	2017
135	□	2016
204	□	2015
181	□	2014
179	□	2013
146	□	2012
128	□	2011
100	□	2010
111	□	2009
80	□	2008
97	□	2007
76	□	2006
66	□	2005
45	□	2004
45	□	2003

#	OR	Journal	
151	□	Environ Health Perspect	P
81	□	Environ Res	P
70	□	Epidemiology	P
62	□	Sci Total Environ	P
58	□	Circulation	P
48	□	Environ Health	P
47	□	Am J Epidemiol	P
46	□	Occup Environ Med	P
38	□	PLoS One	P
37	□	Int J Environ Res Public Health	P
33	□	J Toxicol Environ Health A	P
30	□	J Epidemiol Community Health	P
30	□	Stroke	P

#	OR	Author	
99	□	SCHWARTZ J	P
59	□	PETERS A	P
43	□	BURNETT RT	P
40	□	KAN H	P
36	□	KAUFMAN JD	P
34	□	FORASTIERE F	P
33	□	ZANOBETTI A	P
30	□	KUNZLI N	P
30	□	MITTLEMAN MA	P
30	□	ZHANG Y	P
29	□	YANG CY	P
26	□	ANDERSON HR	P
26	□	BRUNEKREEF B	P
26	□	JERRETT M	P
24	□	BROOK RD	P

#	Count	OR	Word
1953	5808	□	POLLUT *
1706	5830	□	EFFECT *
1687	1880	□	HUMAN *
1660	5077	□	DISEASE *
1623	6937	□	AIR
1408	3099	□	EPIDEMIOLOGY
1395	3617	□	CARDIOVASCULAR
1315	4784	□	EXPOSURE *
1300	2708	□	ADVERSE *
1296	3900	□	RISK *
1229	3189	□	INCREASE *
1119	2281	□	DATA
1105	2222	□	ASSOCIATE *
1073	2748	□	ASSOCIE *
1068	2848	□	ANALYSE *

#	OR	Mesh	
2954	-	/ epidemiology	P
2378	-	/ adverse effects	P
1652	□	Humans	P
1380	□	/ analysis	P
1240	□	/ mortality	P
1110	□	/ etiology	P
923	□	/ statistics & numerical data	P
844	□	Female	P
826	□	Male	P
807	□	Air Pollution	P
763	□	Cardiovascular Diseases	P
650	□	Middle Aged	P
634	□	Aged	P
614	□	Air Pollutants	P
570	□	/ toxicity	P

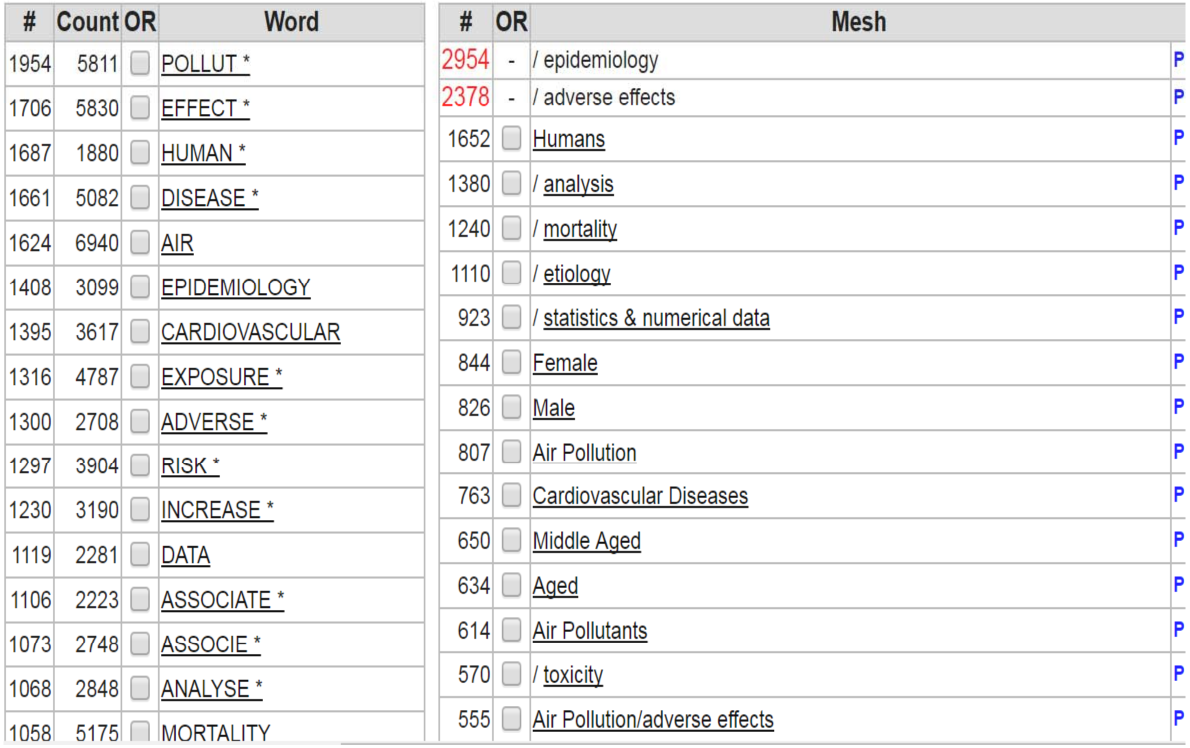
#	OR	Substances
614	□	Air Pollutants
452	□	Particulate Matter
412	□	Tobacco Smoke Pollution
148	□	Nitrogen Dioxide
148	□	Ozone
141	□	Sulfur Dioxide
98	□	Vehicle Emissions
94	□	Carbon Monoxide
48	□	Smoke
45	□	Dust
44	□	Biological Markers
33	□	Cotinine
24	□	C-Reactive Protein
21	□	Nitric Oxide
19	□	Carbon

#	OR	pu
1908	□	JOUR ARTI
848	□	RESE SUPF U.S. (
268	□	REVII
229	□	RESE SUPF EXTR
188	□	RESE SUPF GOVT P.H.S
138	□	ENGL ABST
123	□	COMI STUD
93	□	COMI
83	□	RESE SUPF

PubMed PubReminer - results

# refers to the number of references the term appears in

Count refers to the number of times the term appears in total



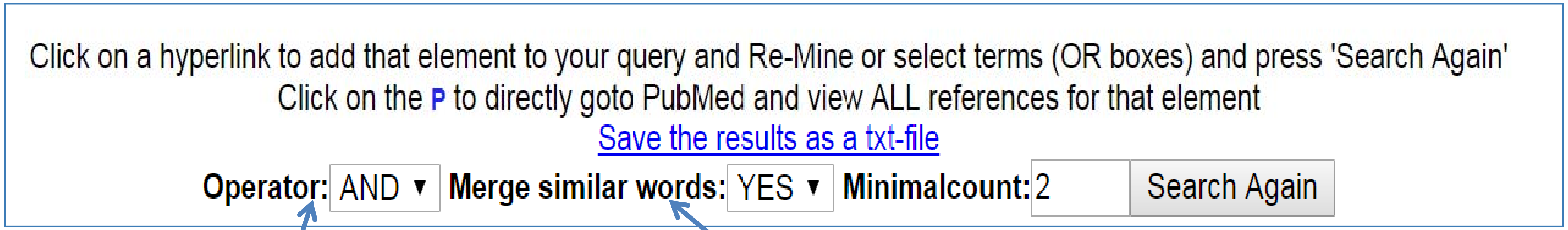
#	Count	OR	Word
1954	5811	☐	POLLUT *
1706	5830	☐	EFFECT *
1687	1880	☐	HUMAN *
1661	5082	☐	DISEASE *
1624	6940	☐	AIR
1408	3099	☐	EPIDEMIOLOGY
1395	3617	☐	CARDIOVASCULAR
1316	4787	☐	EXPOSURE *
1300	2708	☐	ADVERSE *
1297	3904	☐	RISK *
1230	3190	☐	INCREASE *
1119	2281	☐	DATA
1106	2223	☐	ASSOCIATE *
1073	2748	☐	ASSOCIE *
1068	2848	☐	ANALYSE *
1058	5175	☐	MORTALITY

#	OR	Mesh	
2954	-	/ epidemiology	P
2378	-	/ adverse effects	P
1652	☐	Humans	P
1380	☐	/ analysis	P
1240	☐	/ mortality	P
1110	☐	/ etiology	P
923	☐	/ statistics & numerical data	P
844	☐	Female	P
826	☐	Male	P
807	☐	Air Pollution	P
763	☐	Cardiovascular Diseases	P
650	☐	Middle Aged	P
634	☐	Aged	P
614	☐	Air Pollutants	P
570	☐	/ toxicity	P
555	☐	Air Pollution/adverse effects	P

P will process a search in PubMed for the chosen term

A click on the term will add it to the query in PubReminer

PubMed PubReminer advanced search options



Click on a hyperlink to add that element to your query and Re-Mine or select terms (OR boxes) and press 'Search Again'
Click on the P to directly goto PubMed and view ALL references for that element
Save the results as a txt-file

Operator: AND ▾ **Merge similar words:** YES ▾ **Minimalcount:** 2

Add selected items from the results analysis to the query
available operators : AND / NOT

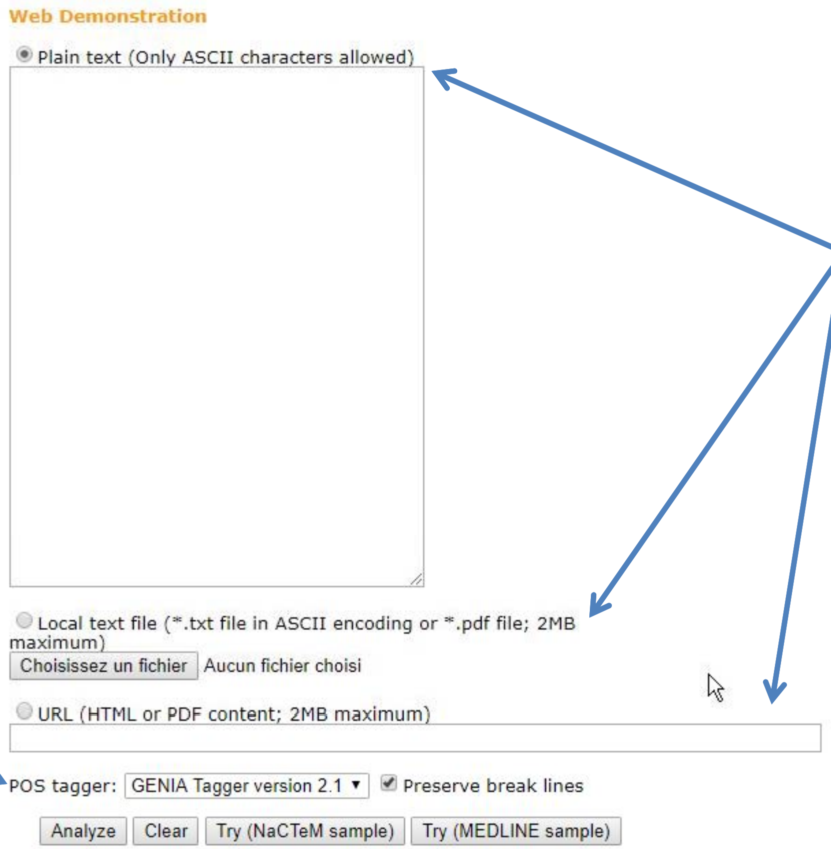
does not work with MeSH terms with multiple subheadings

Only for the word column

TerMine

- Tool used for term recognition
  - recognises automatically candidate multiword terms from documents.
- It annotates raw text from any database
- It recognises acronyms

TerMine Web Demonstration



Web Demonstration

● Plain text (Only ASCII characters allowed)

○ Local text file (*.txt file in ASCII encoding or *.pdf file; 2MB maximum)

Choisissez un fichier Aucun fichier choisi

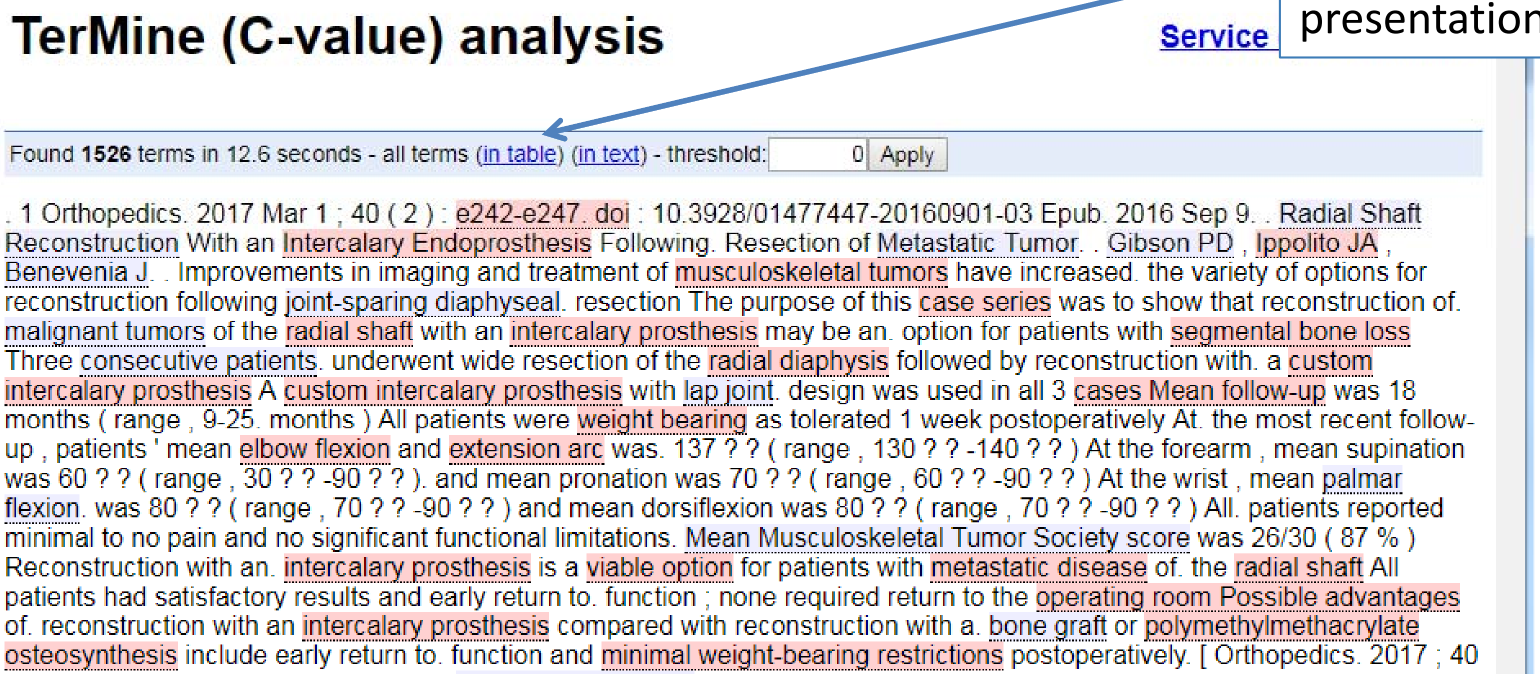
○ URL (HTML or PDF content; 2MB maximum)

POS tagger: GENIA Tagger version 2.1 ▾ ☑ Preserve break lines

Different options of introducing text
(2MB maximum)

Check that the selected tagger is GENIA tagger.
This tagger is specifically tuned for biomedical text such as MEDLINE abstracts

TerMine results



TerMine (C-value) analysis

Found **1526** terms in 12.6 seconds - all terms (in table) (in text) - threshold: 0 Apply

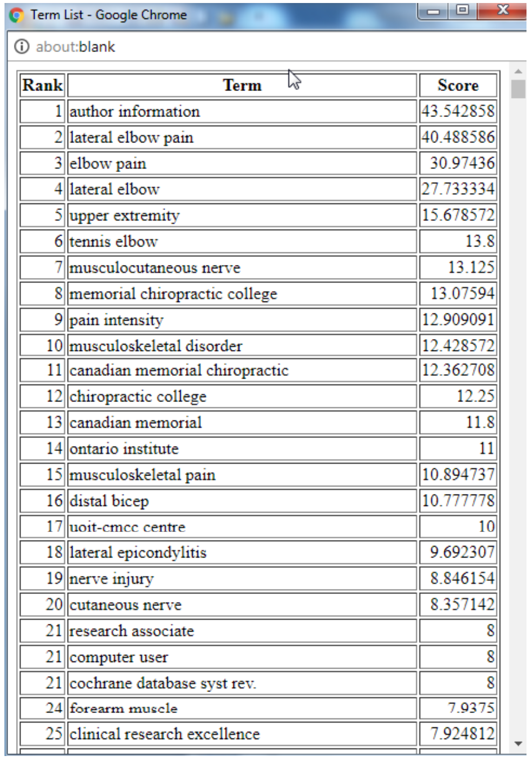
. 1 Orthopedics. 2017 Mar 1 ; 40 (2) : e242-e247. doi : 10.3928/01477447-20160901-03 Epub. 2016 Sep 9. . Radial Shaft Reconstruction With an Intercalary Endoprosthesis Following. Resection of Metastatic Tumor. . Gibson PD , Ippolito JA , Benevenia J. . Improvements in imaging and treatment of musculoskeletal tumors have increased. the variety of options for reconstruction following joint-sparing diaphyseal. resection The purpose of this case series was to show that reconstruction of. malignant tumors of the radial shaft with an intercalary prosthesis may be an. option for patients with segmental bone loss Three consecutive patients. underwent wide resection of the radial diaphysis followed by reconstruction with. a custom intercalary prosthesis A custom intercalary prosthesis with lap joint. design was used in all 3 cases Mean follow-up was 18 months (range , 9-25. months) All patients were weight bearing as tolerated 1 week postoperatively At. the most recent follow-up , patients ' mean elbow flexion and extension arc was. 137 ? ? (range , 130 ? ? -140 ? ?) At the forearm , mean supination was 60 ? ? (range , 30 ? ? -90 ? ?). and mean pronation was 70 ? ? (range , 60 ? ? -90 ? ?) At the wrist , mean palmar flexion. was 80 ? ? (range , 70 ? ? -90 ? ?) and mean dorsiflexion was 80 ? ? (range , 70 ? ? -90 ? ?) All. patients reported minimal to no pain and no significant functional limitations. Mean Musculoskeletal Tumor Society score was 26/30 (87 %) Reconstruction with an. intercalary prosthesis is a viable option for patients with metastatic disease of. the radial shaft All patients had satisfactory results and early return to. function ; none required return to the operating room Possible advantages of. reconstruction with an intercalary prosthesis compared with reconstruction with a. bone graft or polymethylmethacrylate osteosynthesis include early return to. function and minimal weight-bearing restrictions postoperatively. [Orthopedics. 2017 ; 40

Choose the
in table
presentation

TerMine

Results in table format

- Score indicates the c-value. This value is calculated using automatic term recognition, based on the following characteristics:
  - the occurrence frequency of the candidate term
  - the frequency of the candidate term as part of other longer candidate terms
  - the number of these longer candidate terms
  - the length of the candidate term

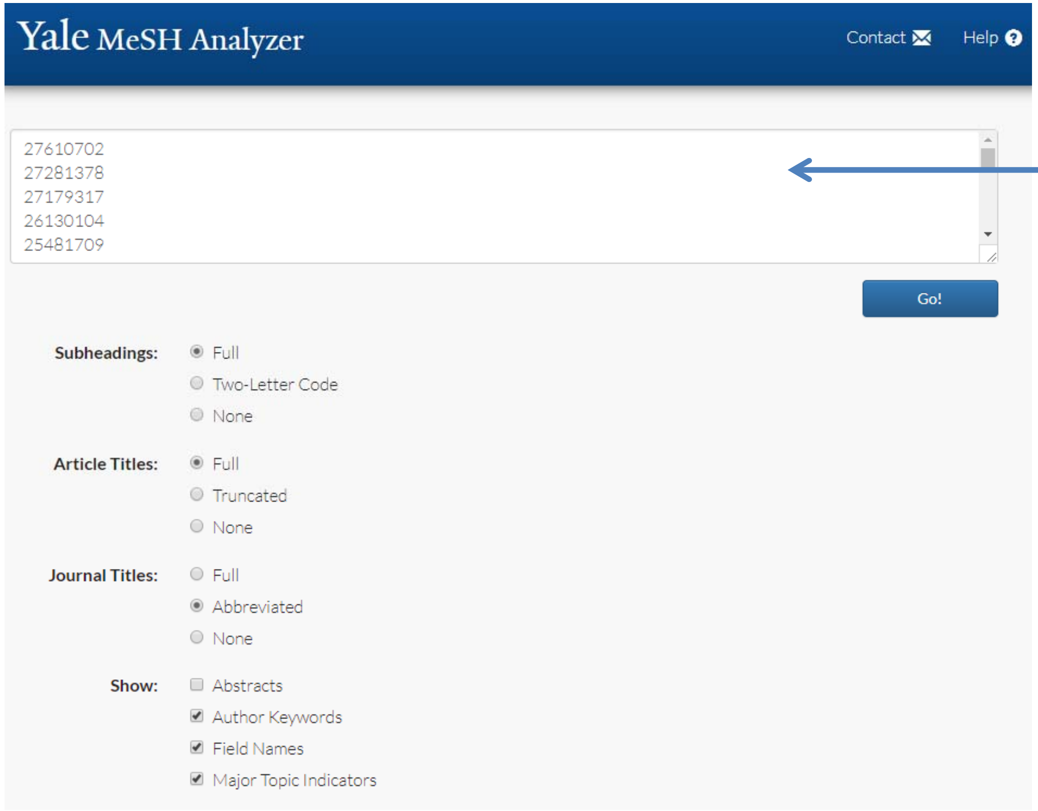


Rank	Term	Score
1	author information	43.542858
2	lateral elbow pain	40.488586
3	elbow pain	30.97436
4	lateral elbow	27.733334
5	upper extremity	15.678572
6	tennis elbow	13.8
7	musculocutaneous nerve	13.125
8	memorial chiropractic college	13.07594
9	pain intensity	12.909091
10	musculoskeletal disorder	12.428572
11	canadian memorial chiropractic	12.362708
12	chiropractic college	12.25
13	canadian memorial	11.8
14	ontario institute	11
15	musculoskeletal pain	10.894737
16	distal bicep	10.777778
17	uoit-cmcc centre	10
18	lateral epicondylitis	9.692307
19	nerve injury	8.846154
20	cutaneous nerve	8.357142
21	research associate	8
21	computer user	8
21	cochrane database syst rev.	8
24	forearm muscle	7.9375
25	clinical research excellence	7.924812

Yale MeSH Analyzer

- Creates a grid which displays the ways articles are indexed in Medline
- For each article, MeSH Terms are sorted and grouped alphabetically
- Choice to include author keywords, titles and abstracts in the analysis grid

Yale MeSH Analyzer



27610702
27281378
27179317
26130104
25481709

Subheadings: ● Full ○ Two-Letter Code ○ None

Article Titles: ● Full ○ Truncated ○ None

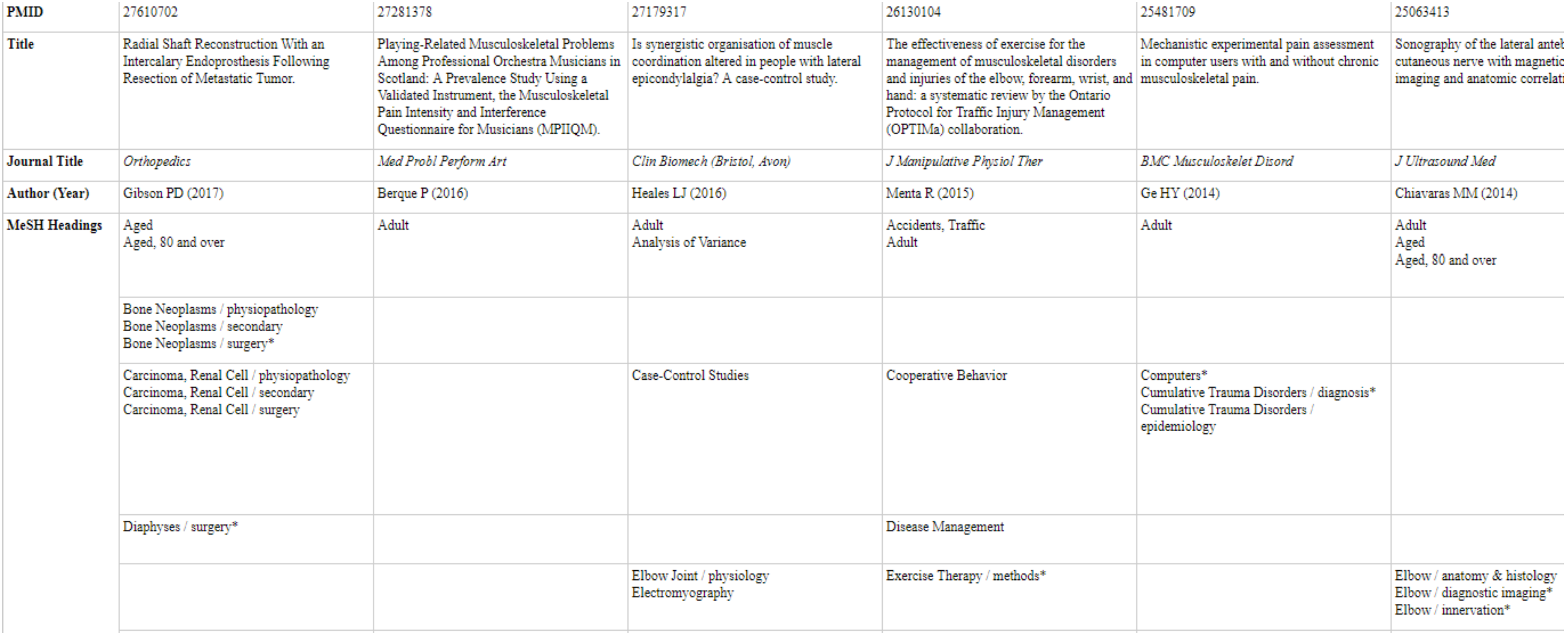
Journal Titles: ○ Full ● Abbreviated ○ None

Show: □ Abstracts ☑ Author Keywords ☑ Field Names ☑ Major Topic Indicators

Introduce list of PMID

Only 20 articles at a time !

Yale MeSH Analyzer



PMID	27610702	27281378	27179317	26130104	25481709	25063413
Title	Radial Shaft Reconstruction With an Intercalary Endoprosthesis Following Resection of Metastatic Tumor.	Playing-Related Musculoskeletal Problems Among Professional Orchestra Musicians in Scotland: A Prevalence Study Using a Validated Instrument, the Musculoskeletal Pain Intensity and Interference Questionnaire for Musicians (MPIIQM).	Is synergistic organisation of muscle coordination altered in people with lateral epicondylalgia? A case-control study.	The effectiveness of exercise for the management of musculoskeletal disorders and injuries of the elbow, forearm, wrist, and hand: a systematic review by the Ontario Protocol for Traffic Injury Management (OPTIMa) collaboration.	Mechanistic experimental pain assessment in computer users with and without chronic musculoskeletal pain.	Sonography of the lateral anteb cutaneous nerve with magnetic imaging and anatomic correlati
Journal Title	*Orthopedics*	*Med Probl Perform Art*	*Clin Biomech (Bristol, Avon)*	*J Manipulative Physiol Ther*	*BMC Musculoskelet Disord*	*J Ultrasound Med*
Author (Year)	Gibson PD (2017)	Berque P (2016)	Heales LJ (2016)	Menta R (2015)	Ge HY (2014)	Chiavaras MM (2014)
MeSH Headings	Aged Aged, 80 and over	Adult	Adult Analysis of Variance	Accidents, Traffic Adult	Adult	Adult Aged Aged, 80 and over
	Bone Neoplasms / physiopathology Bone Neoplasms / secondary Bone Neoplasms / surgery*					
	Carcinoma, Renal Cell / physiopathology Carcinoma, Renal Cell / secondary Carcinoma, Renal Cell / surgery		Case-Control Studies	Cooperative Behavior	Computers* Cumulative Trauma Disorders / diagnosis* Cumulative Trauma Disorders / epidemiology	
	Diaphyses / surgery*			Disease Management		
			Elbow Joint / physiology Electromyography	Exercise Therapy / methods*		Elbow / anatomy & histology Elbow / diagnostic imaging* Elbow / innervation*

Hands-on session

- Work in groups of 2
- Links and text analysis package on http://bit.ly/2xGknrl
- 30'

Discussion

How would you implement the use of a text mining tool today in your own systematic research?

- What are the advantages/disadvantages?
- Which functions do you find useful for your systematic researches?

Our point of view

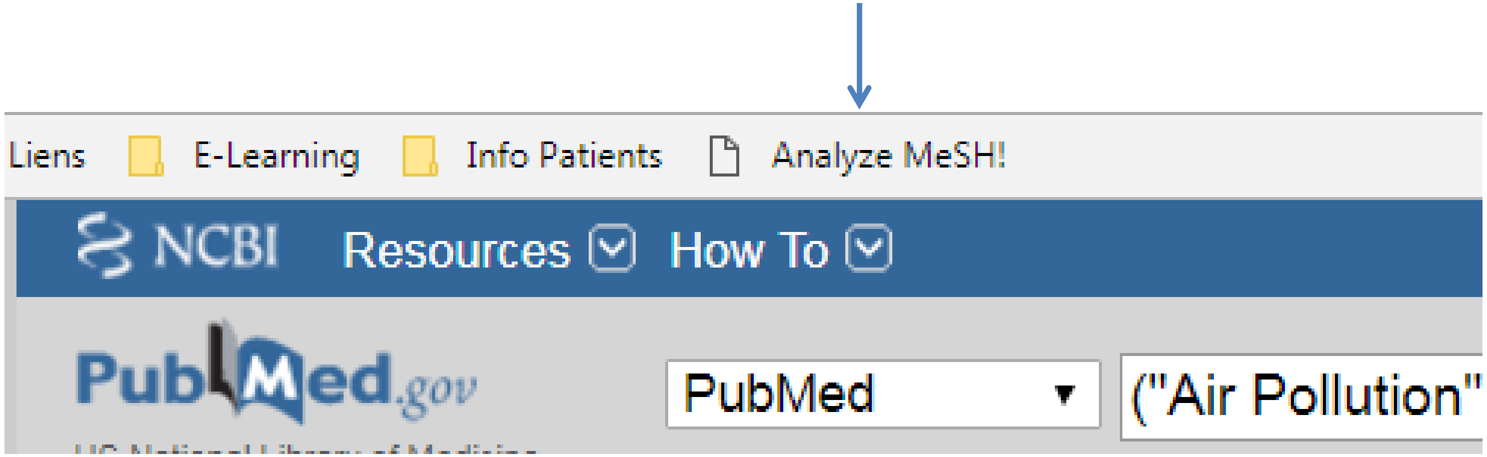
	Useful for	Advantages	Disadvantages
PubReminer	Term identification (ti, ab, MeSH) in gold standard articles provided by the researcher Explore simple PubMed query	Large amount of references Frequency count	
Yale Analyzer	MeSH term identification in gold standard articles Starting point for the MeSH analysis with researcher	Easy to use, to read.	Only 20 PMID No frequency count
Termine	Phrase identification in references text (ti, ab k eywords) useful when the number of references is quite large	References from other databases	For large amounts of text, necessity to register

Tips & Tricks - PubReminer

- can help choosing journals for the handsearching process
- Be aware of the Merge function
  - Select No to see all the term variations
  - Select Yes to have a view of the importance of the term root

Tips & Tricks – Yale analyzer

- can be used directly in PubMed



("Air Pollution"

Tips & Tricks - Termine

- Batch Service: for processing documents larger than 2MB

- create a text file by import / export in Endnote with an export style that keep only relevant text (ti, ab, keywords)

Report results to researcher

- PubReminer : save as a text file
- Yale MeSH Analyzer : output option - excel
- Termine : create an excel file by copying and pasting the table

Bibliography

1. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. Res Syn Meth. 2011;2(1):1-14. doi: 10.1002/jrsm.27
2. Belter CW. Citation analysis as a literature search method for systematic reviews. 1 nov 2016;2766-77. doi:10.1002/asi.23605
3. EPC Methods [online]: An Exploration of the Use of Text-Mining Software in Systematic Reviews - NCBI Bookshelf; [Accessed 15 August 2017]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK362044/
4. Tsuruoka Y, Tateishi Y, Kim J-D, Ohta T, McNaught J, Ananiadou S, et al. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: Advances in Informatics [Internet]. Springer, Berlin, Heidelberg; 2005 [cited 2017 Aug 27]. p. 382–92. (Lecture Notes in Computer Science). Available from: https://link.springer.com/chapter/10.1007/11573036_36
5. 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. Systematic reviews. 2015;4(1):5. doi: 10.1186/2046-4053-4-5

Cecile.jaques@chuv.ch
jolanda.elmers@chuv.ch

Bibliothèque universitaire de médecine

http://bium.ch