

Topic modelling

approaches to aggregated citation data

Johan Eklund & Gustaf Nelhans
Swedish School of Library and Information Science
STI 2017, Paris 2017-09-07



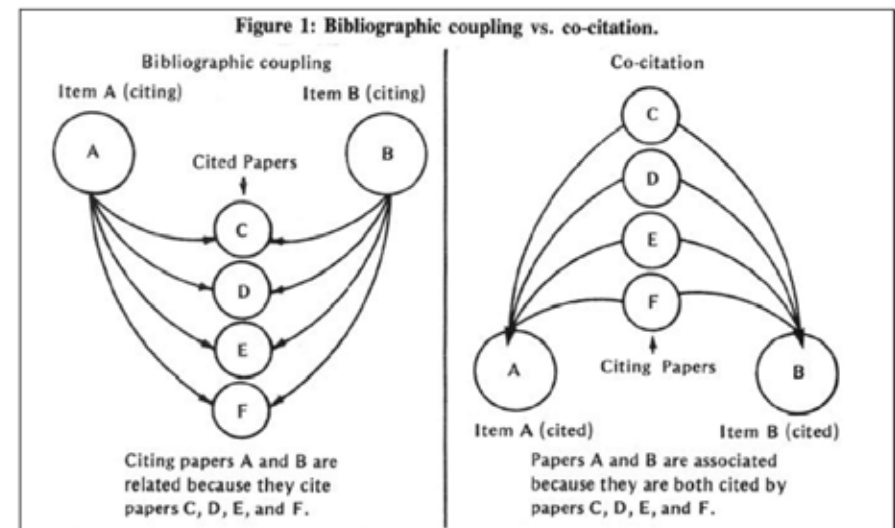
UNIVERSITY OF BORÅS
UTBILDNING FÖR EN BÄTTRE FÖRKLARING

Purpose

- To explore new methods of **combining citation analysis with semantic methods** of elucidating topics from text, so called ***Topic modelling***.
- *The goal is to identify latent structures in the collection of cited references that corresponds to meaningful descriptions of the data.*

Aggregated citation metrics

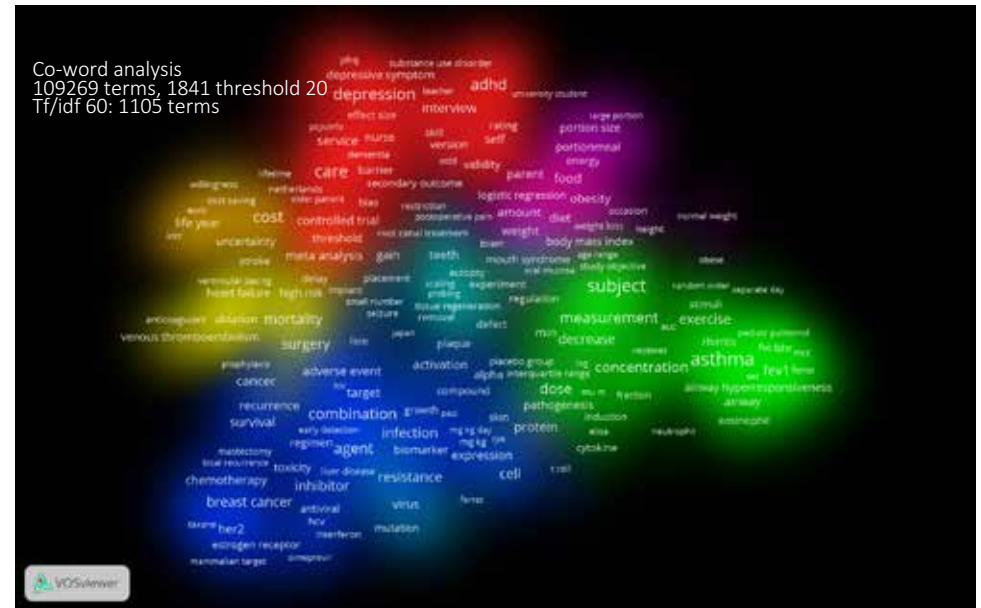
- From citation analysis to...
- Bibliographic coupling
- Co-citation analysis
- Aggregation level:
 - Publication,
 - Author,
 - Source title
 - Institution and Country level (B.C. only)
- (WoS data)



Source: Garfield 2001:
<http://garfield.library.upenn.edu/papers/drexelbelvergriffith92001.pdf>

Text based analysis in scientometrics

- Mainly *co-word analysis*
 - Callon, Law, & Rip (1986).
- *Topic modelling*
 - Few publications yet
 - Some hesitation about
 - Low degree of correlation between co-word analysis and topic modelling on small and medium sized data sets (Leydesdorff & Nerghes, 2017).



Conceptual idea

“The citation as a concept symbol”

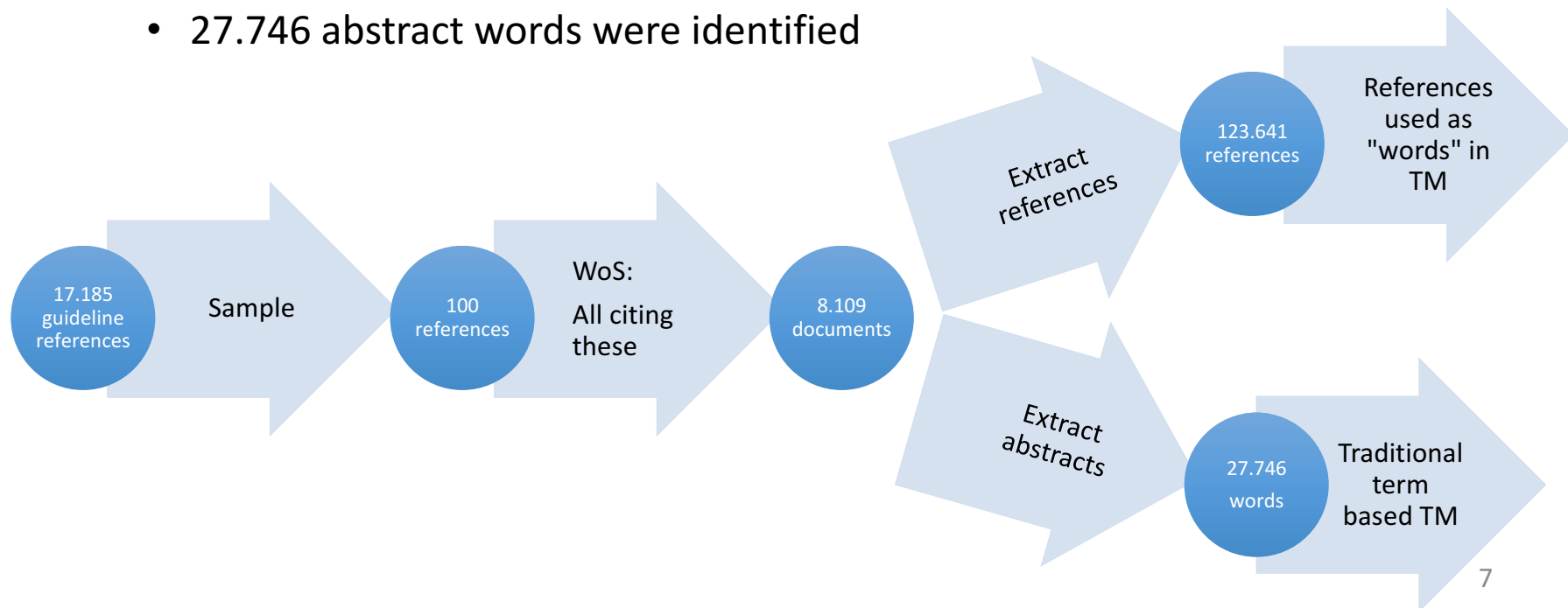
(Small, 1978)

LDA

- ***Latent Dirichlet Allocation (LDA)*** procedure to identify ***latent*** topics "explaining" the patterns of term cooccurrence in a set of texts.
 - The ***underlying assumption*** is that documents contain a ***mixture of topics*** which in turn can be ***expressed as probability distributions of terms***.
- Two different sets:
 - **8.109 Reference lists** of analyzed articles for which each reference is regarded as a word, i.e. a manifest unit of a language (n=123.631 references).
 - **7.178 Abstract texts:** traditional, for which each word in the abstracts are regarded as manifest units corresponding to the underlying topics (n=27.746 terms).

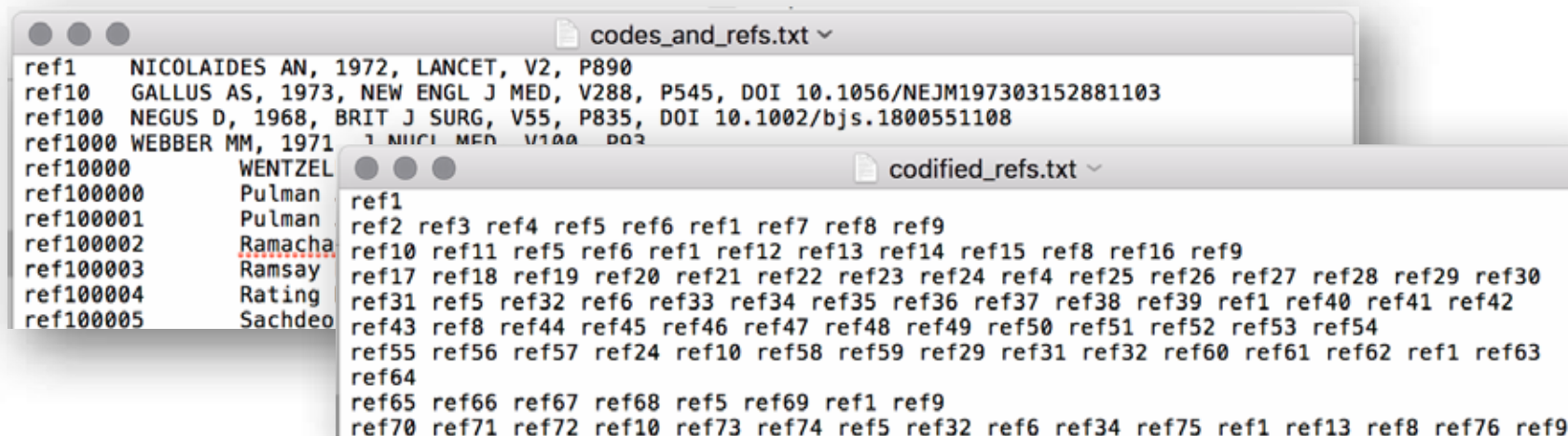
Materials and methods

- 17.185 references in Swedish national clinical guidelines, treatment recommendations and systematic reviews identified in WoS.
- a randomized sample of 100 was made.
- 8.109 papers that cited these were used.
 - 123.641 references were treated as “words” for Topic modelling analysis.
 - 27.746 abstract words were identified



LDA on references

- *Each reference* has been **converted to the code** on the form **ref0123** to facilitate the topical analysis and the generated codes have been combined to form "sentences" according to reference lists in which they appear.



The image shows two overlapping text windows. The top window, titled 'codes_and_refs.txt', contains a list of references with codes: ref1, ref10, ref100, ref1000, ref10000, ref100000, ref100001, ref100002, ref100003, ref100004, and ref100005. The bottom window, titled 'codified_refs.txt', shows a list of reference codes grouped together, representing 'sentences' for LDA analysis. The codes are: ref1, ref2 ref3 ref4 ref5 ref6 ref1 ref7 ref8 ref9, ref10 ref11 ref5 ref6 ref1 ref12 ref13 ref14 ref15 ref8 ref16 ref9, ref17 ref18 ref19 ref20 ref21 ref22 ref23 ref24 ref4 ref25 ref26 ref27 ref28 ref29 ref30, ref31 ref5 ref32 ref6 ref33 ref34 ref35 ref36 ref37 ref38 ref39 ref1 ref40 ref41 ref42, ref43 ref8 ref44 ref45 ref46 ref47 ref48 ref49 ref50 ref51 ref52 ref53 ref54, ref55 ref56 ref57 ref24 ref10 ref58 ref59 ref29 ref31 ref32 ref60 ref61 ref62 ref1 ref63, ref64, ref65 ref66 ref67 ref68 ref5 ref69 ref1 ref9, and ref70 ref71 ref72 ref10 ref73 ref74 ref5 ref32 ref6 ref34 ref75 ref1 ref13 ref8 ref76 ref9.

```
codes_and_refs.txt
ref1  NICOLAIDES AN, 1972, LANCET, V2, P890
ref10 GALLUS AS, 1973, NEW ENGL J MED, V288, P545, DOI 10.1056/NEJM197303152881103
ref100 NEGUS D, 1968, BRIT J SURG, V55, P835, DOI 10.1002/bjs.1800551108
ref1000 WEBBER MM, 1971, J NUCL MED, V100, P03
ref10000 WENTZEL
ref100000 Pulman
ref100001 Pulman
ref100002 Ramacha
ref100003 Ramsay
ref100004 Rating
ref100005 Sachdeo

codified_refs.txt
ref1
ref2 ref3 ref4 ref5 ref6 ref1 ref7 ref8 ref9
ref10 ref11 ref5 ref6 ref1 ref12 ref13 ref14 ref15 ref8 ref16 ref9
ref17 ref18 ref19 ref20 ref21 ref22 ref23 ref24 ref4 ref25 ref26 ref27 ref28 ref29 ref30
ref31 ref5 ref32 ref6 ref33 ref34 ref35 ref36 ref37 ref38 ref39 ref1 ref40 ref41 ref42
ref43 ref8 ref44 ref45 ref46 ref47 ref48 ref49 ref50 ref51 ref52 ref53 ref54
ref55 ref56 ref57 ref24 ref10 ref58 ref59 ref29 ref31 ref32 ref60 ref61 ref62 ref1 ref63
ref64
ref65 ref66 ref67 ref68 ref5 ref69 ref1 ref9
ref70 ref71 ref72 ref10 ref73 ref74 ref5 ref32 ref6 ref34 ref75 ref1 ref13 ref8 ref76 ref9
```

Machine learning library Gensim, created by Radim Řehůřek,

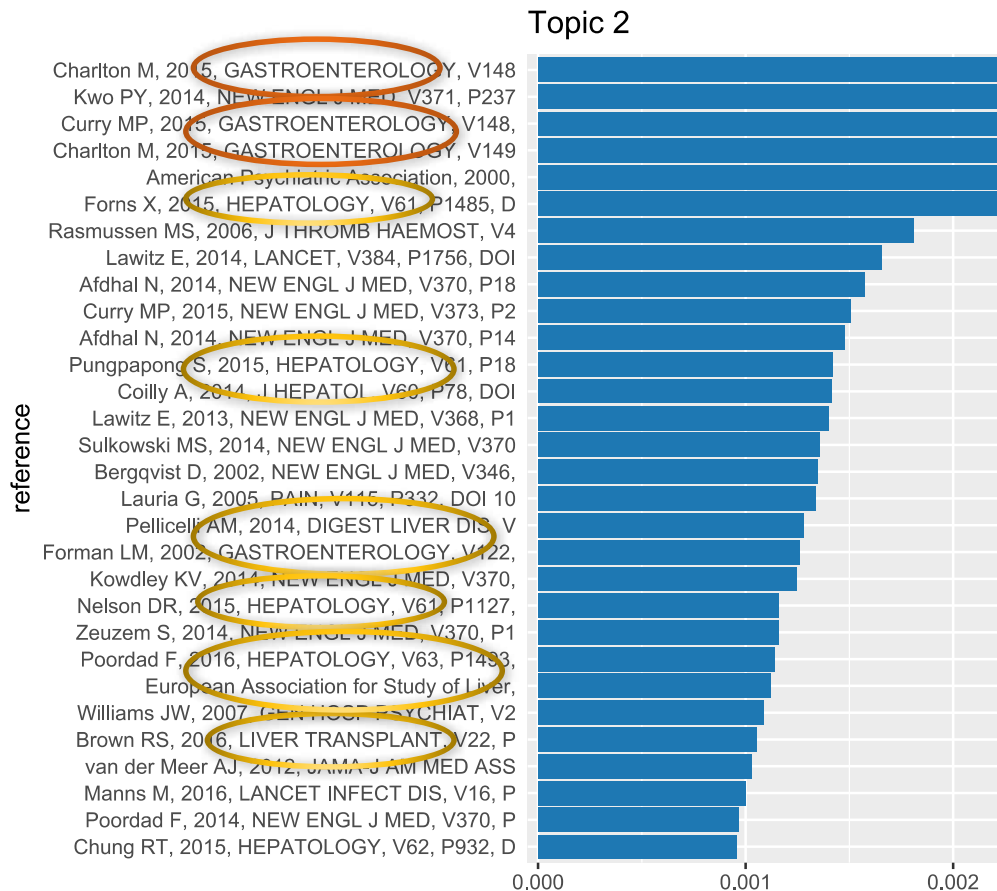
For both the analysis of the abstracts, as well as of the reference lists,

- 5 training passes, chunk size of 100 documents has been used
- 10 topics for each document type (abstracts and references respectively).
- For each topic selected the 30 terms having the highest probability of appearing in the topic.

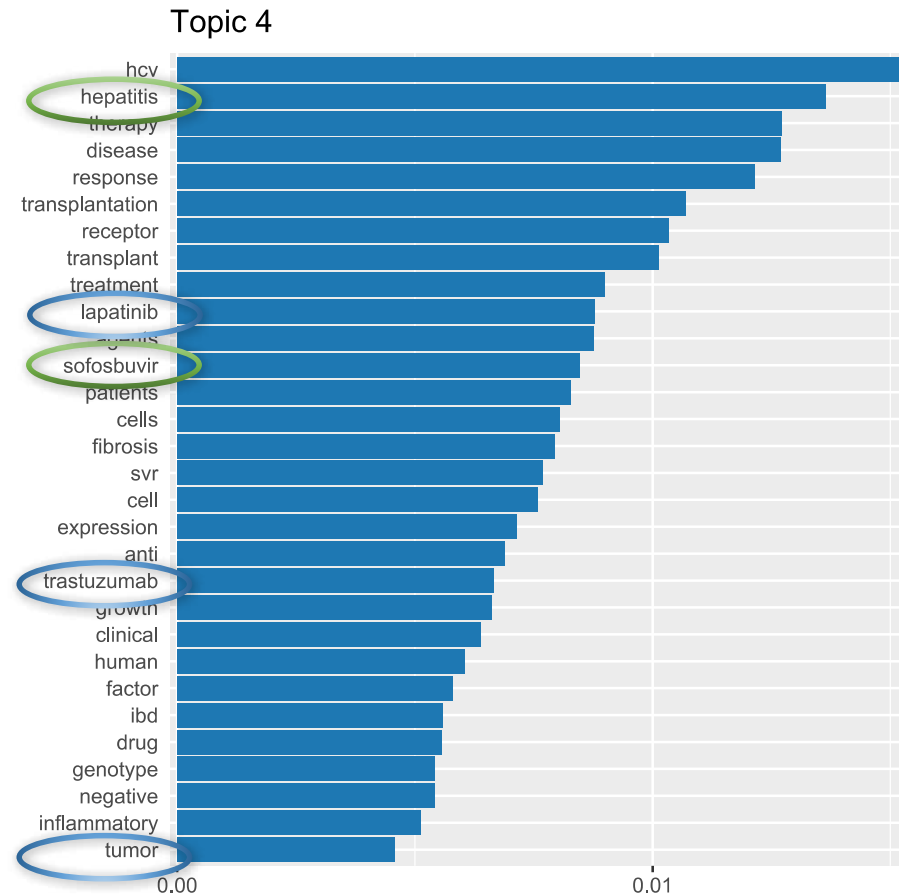
Two topics

-  Digestive sys.
-  Liver dis.
-  Hepatitis C
-  Breast cancer

References



Abstracts



Hellinger distance, which given two discrete probability distributions $\mathbf{p} = (p_1, p_2, \dots, p_n)$ and $\mathbf{q} = (q_1, q_2, \dots, q_n)$ is defined

Hellinger distance

$$\frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^n (\sqrt{p_i} - \sqrt{q_i})^2}$$

This measure yields a value between 0 (complete similarity) and 1 (complete dissimilarity).

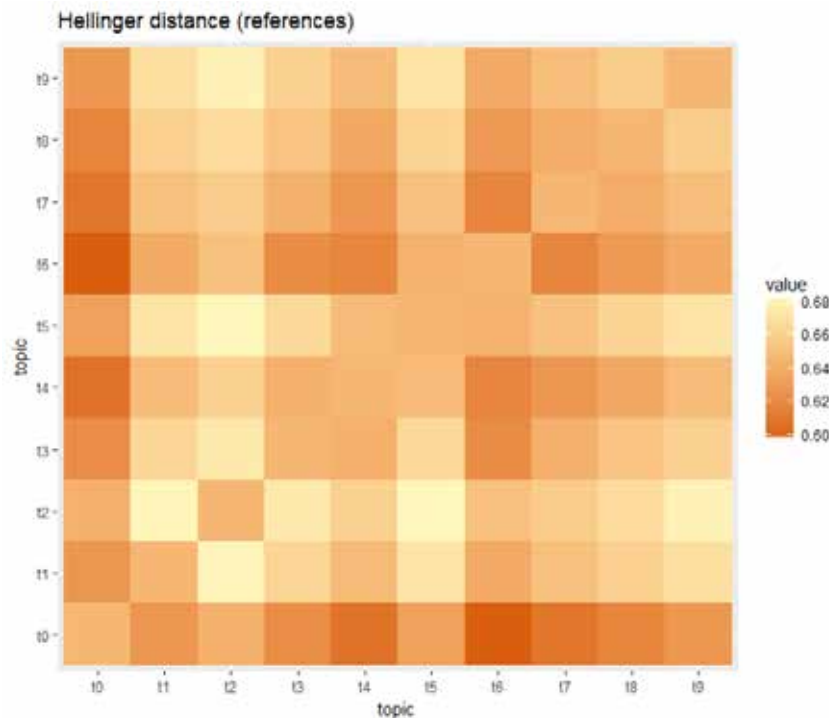


Figure 3: Hellinger distance refs

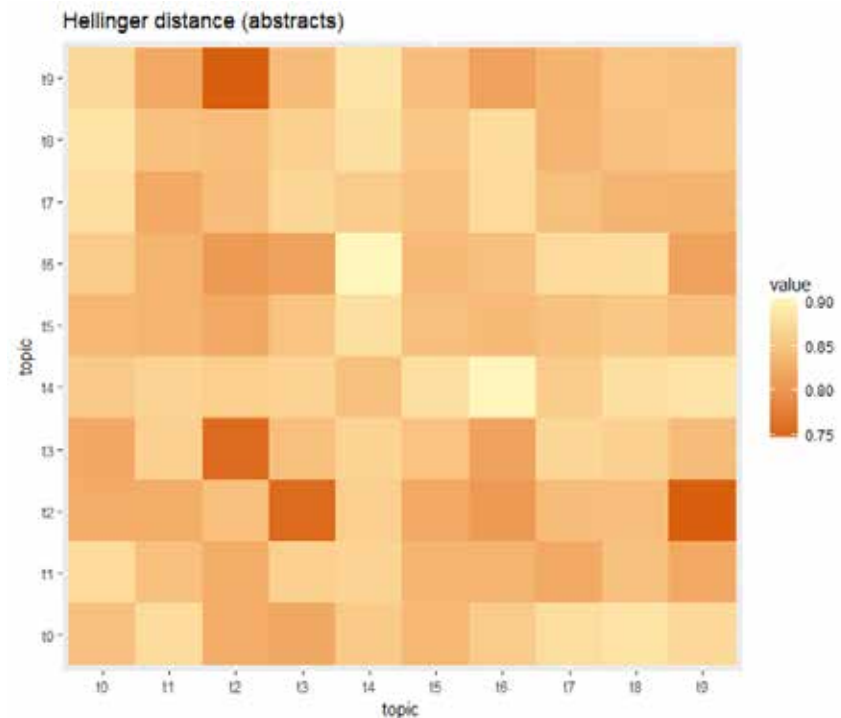


Figure 4: Hellinger Distances (abstracts)

Dissimilarity is generally higher for the abstracts than for the reference lists.

Example use:

doc24687830 SCHAID ET AL 2014 J CLIN ONCOL VOL 32 P 2296

*Prospective Validation of HLA-DRB1*07:01 Allele Carriage As a Predictive Risk Factor for Lapatinib-Induced Liver Injury*

Web of Science

Search

Full Text Option

Prospect
a Predic

By: Schaid, I
(McDonnell, S
Ejlertsen, B (C
Erica) [71]; C

JOURNAL O
Volume: 32
DOI: 10.120
Published:
View Journa

Abstract
Purpose

Liver injury is
patients. This
variants as p
lapatinib in human epidermal growth factor receptor 2-positive, early-stage breast cancer (Tykerb
Evaluation After Chemotherapy [TEACH]: Lapatinib Versus Placebo In Women With Early-Stage Breast
Cancer).

24687830.txt

Topic: 0 Probability: 0.7805555554058086

Rueger SY, 2012, J STUD ALCOHOL DRUGS, V73, P829
McHutchison JG, 1998, NEW ENGL J MED, V339, P148
Shalev L, 2011, NEUROPSYCHOLOGIA, V49, P2584, DOI 10.1016/j.neuropsychologia.2011.05.011
Ogden CL, 2014, JAMA-J AM MED ASSOC, V311, P806, DOI 10.1001/jama.2014.10000
Bear HD, 2003, J CLIN ONCOL, V21, P4165, DOI 10.1200/JCO.2003.09.1450
Addolorato G, 2007, LANCET, V370, P1915, DOI 10.1016/S0140-6736(07)61000-0
Olmstead TA, 2007, ADDICTION, V102, P1443, DOI 10.1111/j.1360-0443.2007.01443.x
Gorman B. S., 1995, HDB ASSESSMENT METHO, P149
Millar EKA, 2009, J CLIN ONCOL, V27, P4701, DOI 10.1200/JCO.2009.09.1450
Sorensen LB, 2003, INT J OBESITY, V27, P1152, DOI 10.1038/sj.ijo.0825000

Topic: 18 Probability: 0.1694444445722232

Johnston S, 2009, J CLIN ONCOL, V27, P5538, DOI 10.1200/JCO.2009.09.1450
American Psychiatric Association, 2000, DIAGN ST
Geyer CE, 2006, NEW ENGL J MED, V355, P2733, DOI 10.1056/NEJMoa060885
Slamon DJ, 2001, NEW ENGL J MED, V344, P783, DOI 10.1056/NEJMoa012618
Verma S, 2012, NEW ENGL J MED, V367, P1783, DOI 10.1056/NEJMoa1200000
Baselga J, 2012, NEW ENGL J MED, V366, P109, DOI 10.1056/NEJMoa1200000
Baselga J, 2012, LANCET, V379, P633, DOI 10.1016/S0140-6736(12)60000-0
Blackwell KL, 2012, J CLIN ONCOL, V30, P2585, DOI 10.1200/JCO.2012.01.0000
SLAMON DJ, 1987, SCIENCE, V235, P177, DOI 10.1126/science.2806321
Blackwell KL, 2010, J CLIN ONCOL, V28, P1124, DOI 10.1200/JCO.2010.01.0000

Topic: 0
Probability: 0.26
patients
cancer
participants
breast
therapy
treatment
positive
mortality
disease
clinical
therapies
infection
virus
resistance
agents
lapatinib
sofosbuvir
receptor
antiviral
chemotherapy

Topic: VII
Probability: 0.24
size
portion
intake
eating
weight
cells
levels
effect
increased
main
expression
cell
activity
bms
trastuzumab
growth
elevation
mechanisms
healthy
energy

Topic: IV
Probability: 0.21
patients
treatment
group
study
months
intervention
compared
results
follow
year
significant
weeks
control
survival
baseline
therapy
randomized
methods
scale
effect

Conclusions

- Novel ways of combining text based information science approaches with established scientometric methods.
 - ***Complement existing*** text based and citation based techniques for clustering of research
 - **Bridging** the two approaches
 - Embodying the idea of citations as ***concept symbols*** (Small, 1978)
- Usefulness:
 - Provides methods to ***classify document sets based on their references*** (such as clinical guidelines).
 - Perspective shift: identifying ***latent references*** in a paper(!)