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

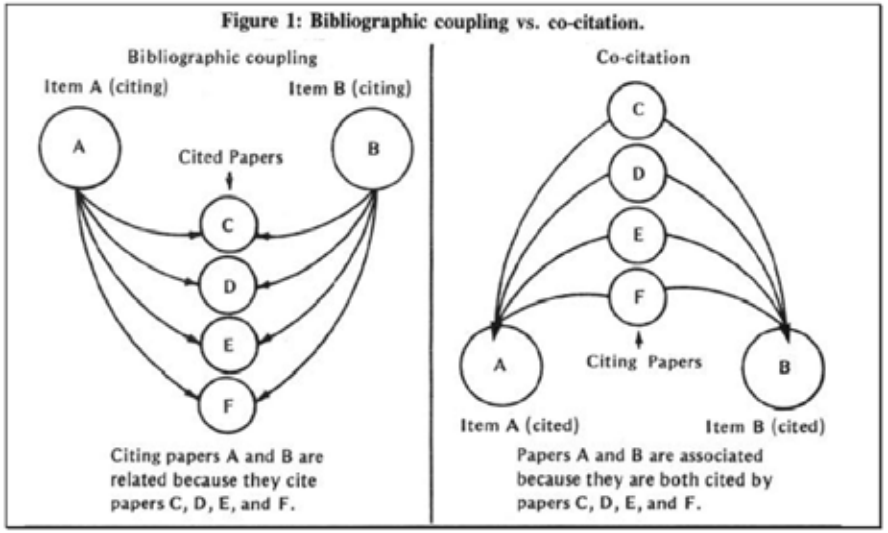
Purpose

- 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

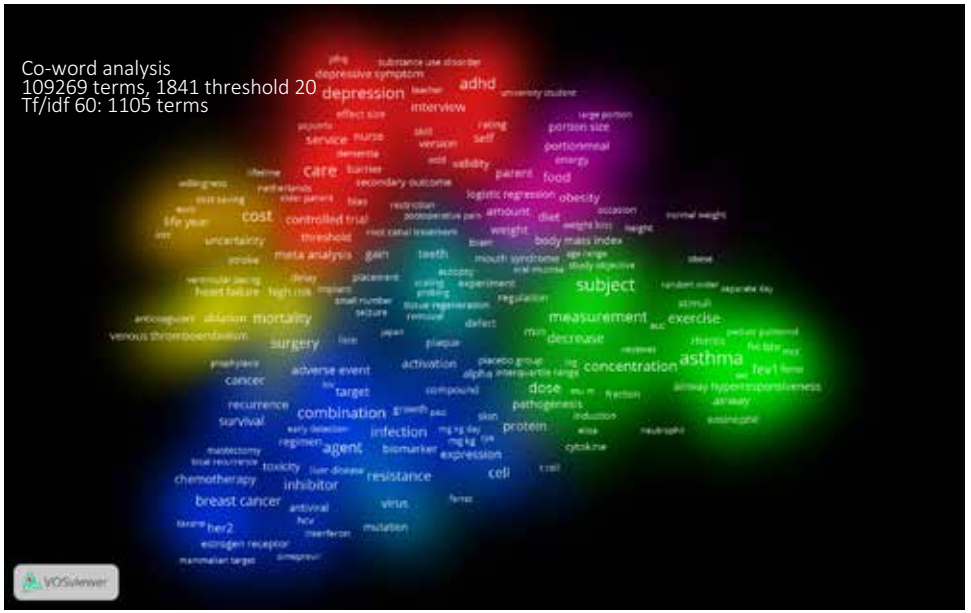
- From citation analysis to...
- Bibliographic coupling
- Co-citation analysis



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

- Aggregation level:
  - Publication,
  - Author,
  - Source title
  - Institution and Country level (B.C. only)
- (WoS data)

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”

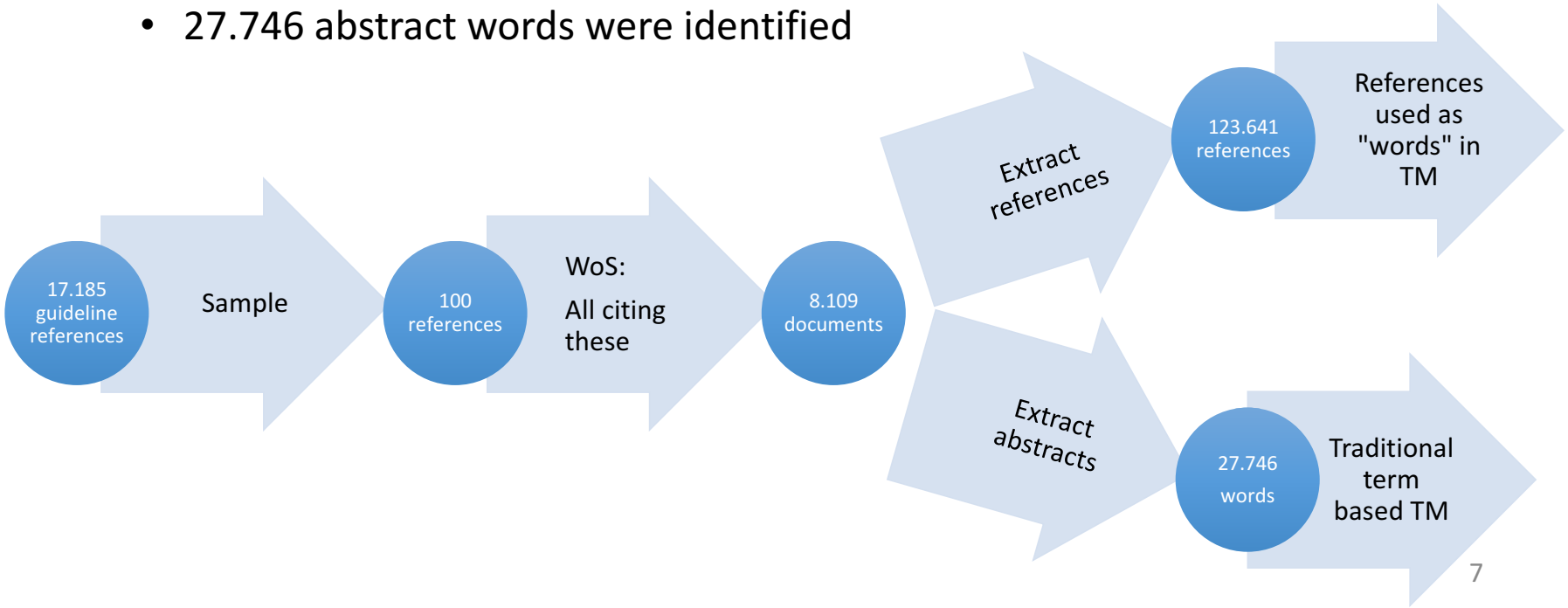
(Small, 1978)

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).

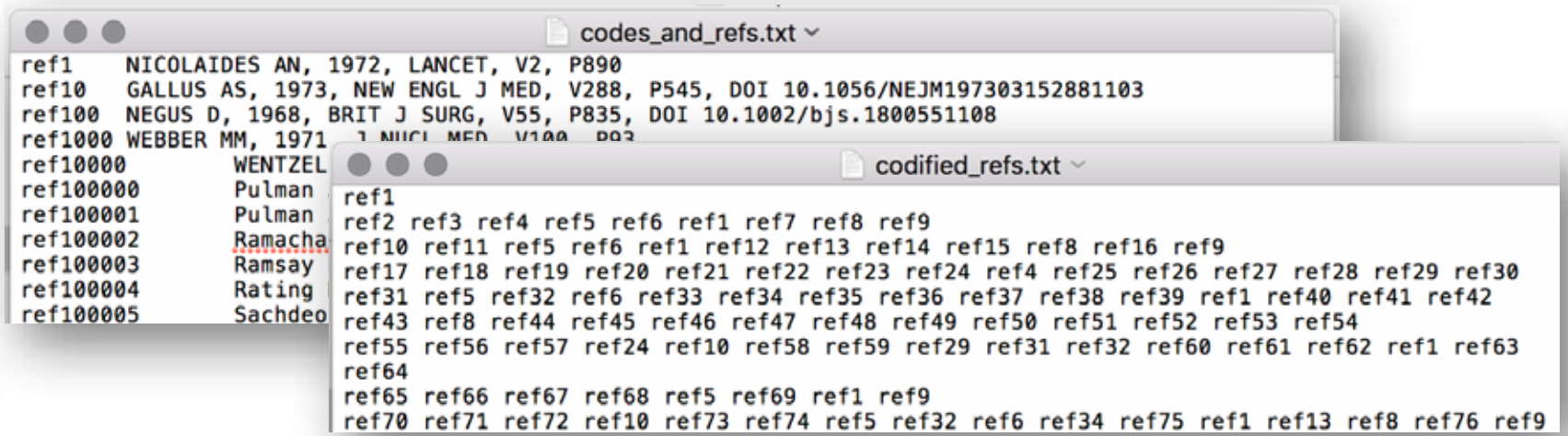
Materials and methods

- 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.



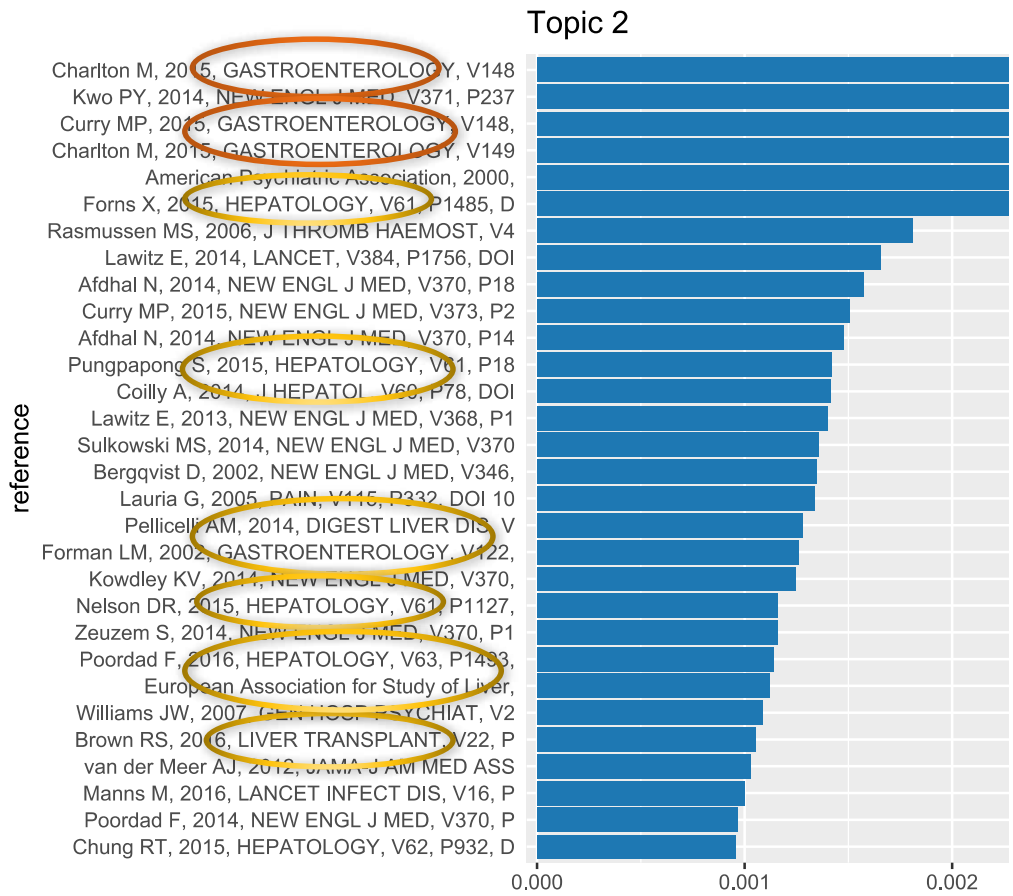
Machine learning library Gensim, created by Radim Řehůřek,
For both the analysis of the abstracts, as well as of the reference lists,

- 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.

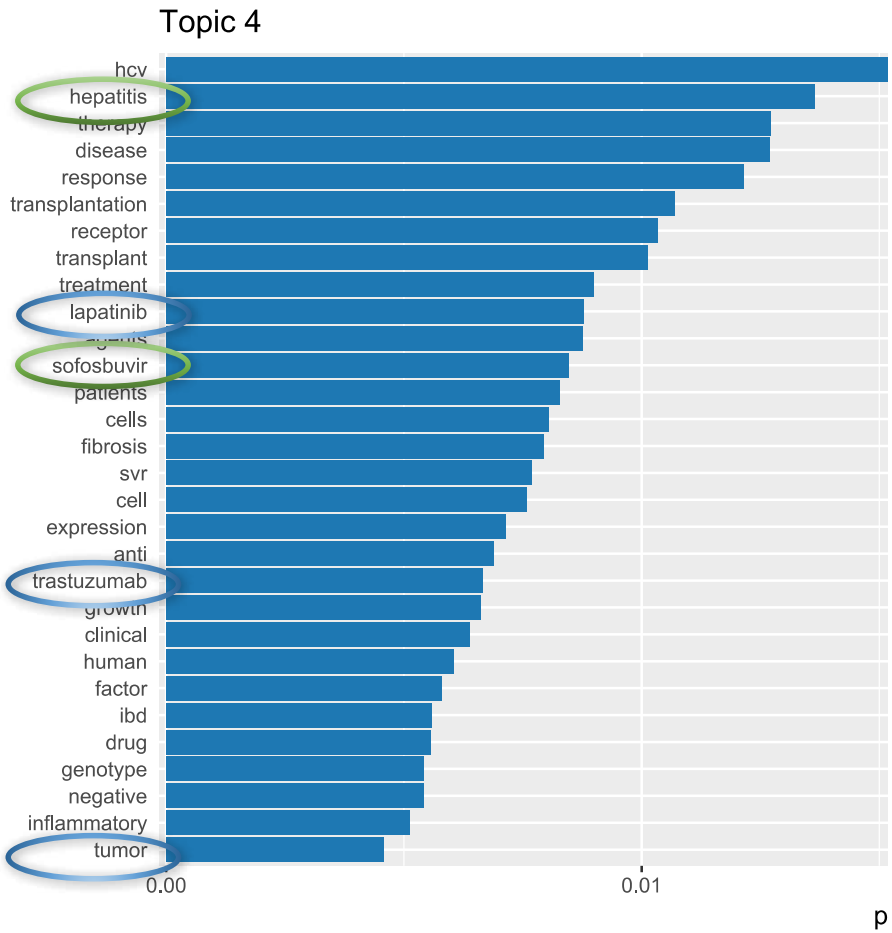
Two topics

Digestive sys.
Liver dis.
Hepatitis C
Breast cancer

References



Abstracts



Hellinger distance

Hellinger distance, which given two discrete probability distributions **p** = (p_1, p_2, …, p_n) and **q** = (q_1, q_2, …, q_n) is defined

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

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

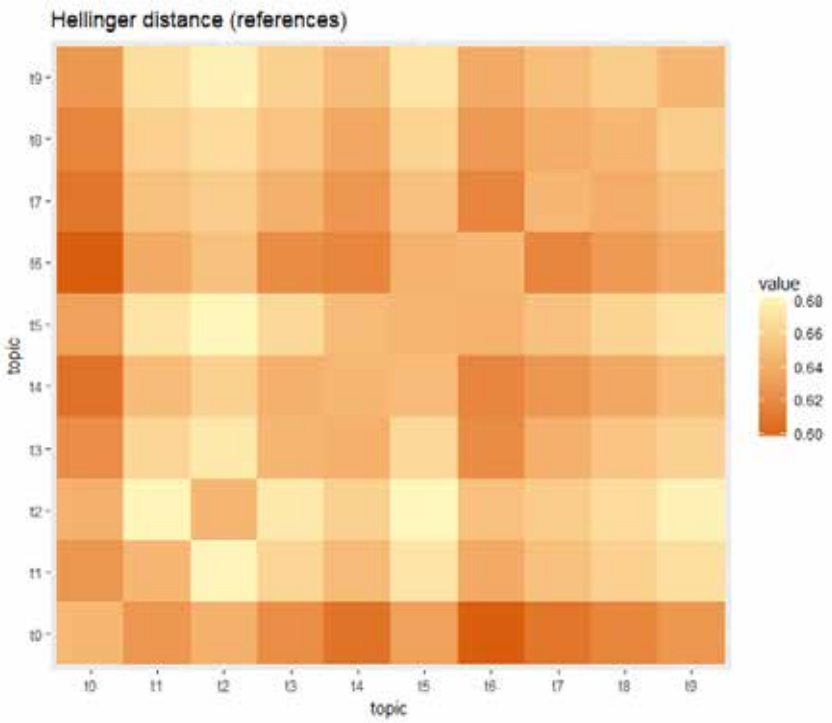


Figure 3: Hellinger distance refs

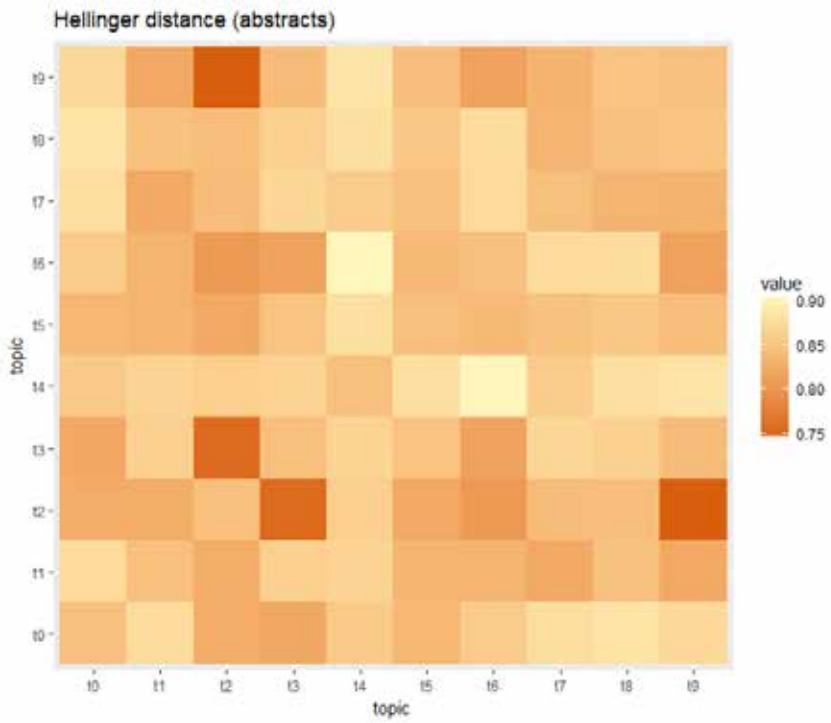


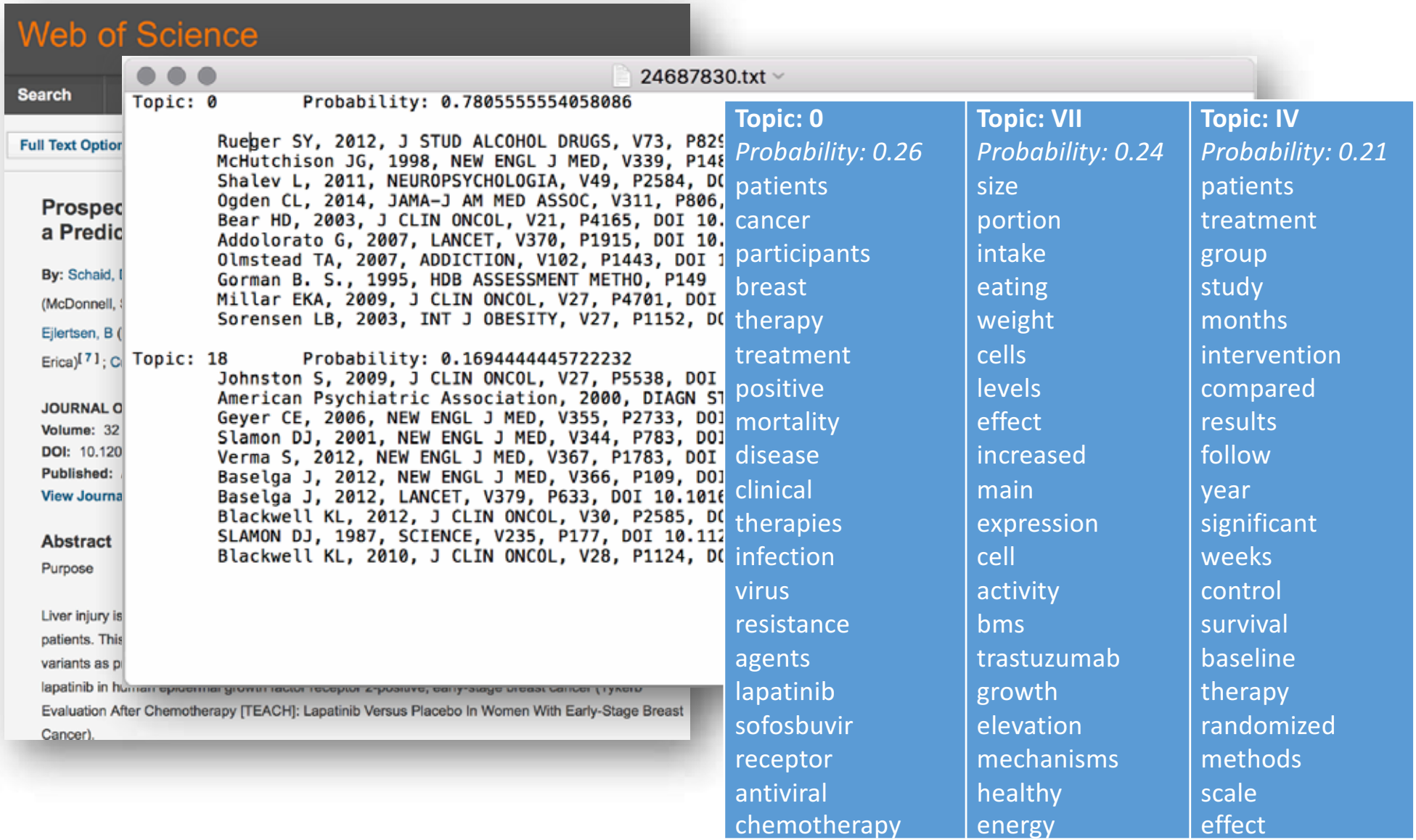
Figure 4: Hellinger Distances (abstracts)

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*



Topic: 0	Topic: VII	Topic: IV
Probability: 0.26	*Probability: 0.24*	*Probability: 0.21*
patients	size	patients
cancer	portion	treatment
participants	intake	group
breast	eating	study
therapy	weight	months
treatment	cells	intervention
positive	levels	compared
mortality	effect	results
disease	increased	follow
clinical	main	year
therapies	expression	significant
infection	cell	weeks
virus	activity	control
resistance	bms	survival
agents	trastuzumab	baseline
lapatinib	growth	therapy
sofosbuvir	elevation	randomized
receptor	mechanisms	methods
antiviral	healthy	scale
chemotherapy	energy	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(!)