Natural Language Processing and Topic Modelling

Ştefan Trăușan-Matu

University Politehnica of Bucharest
Romanian Academy Research Institute for Artificial Intelligence

COST Action IS1310 - *Reassembling the Republic of Letters*
23rd March St'. Anne's College University of Oxford

Natural Language Processing (NLP)

- Input:
  - Text in digital format (strings of characters)
    - document
    - corpus
    - question
    - transcription of a monologue or of a conversation
    - instant messenger log
    - discussion forum, social network
    - corpus of interlinked documents (e.g. letters)
    - dialog

Natural Language Processing

- Output:
  - Text(s) in digital format
    - translation – e.g. Google translate
    - document(s) summary - summarizers
    - answer – question answering
    - clusters of documents
  - Automatically generated annotations
  - List of topics in the text
  - Links among topics
  - Similar documents
  - Links among documents – intertextuality
  - Threads of discussion, COLLABORATION
  - Other data
    - collocations
    - structures (syntactical, discourse, rhetorical, etc.)
    - opinions In text
    - participation & collaboration degrees in conversations
    - …

NLP approaches

- Grammar-based
- Statistical (corpus-based, machine learning)
  - unsupervized (clustering, LSA, LDA)
  - Supervized

    annotated corpus

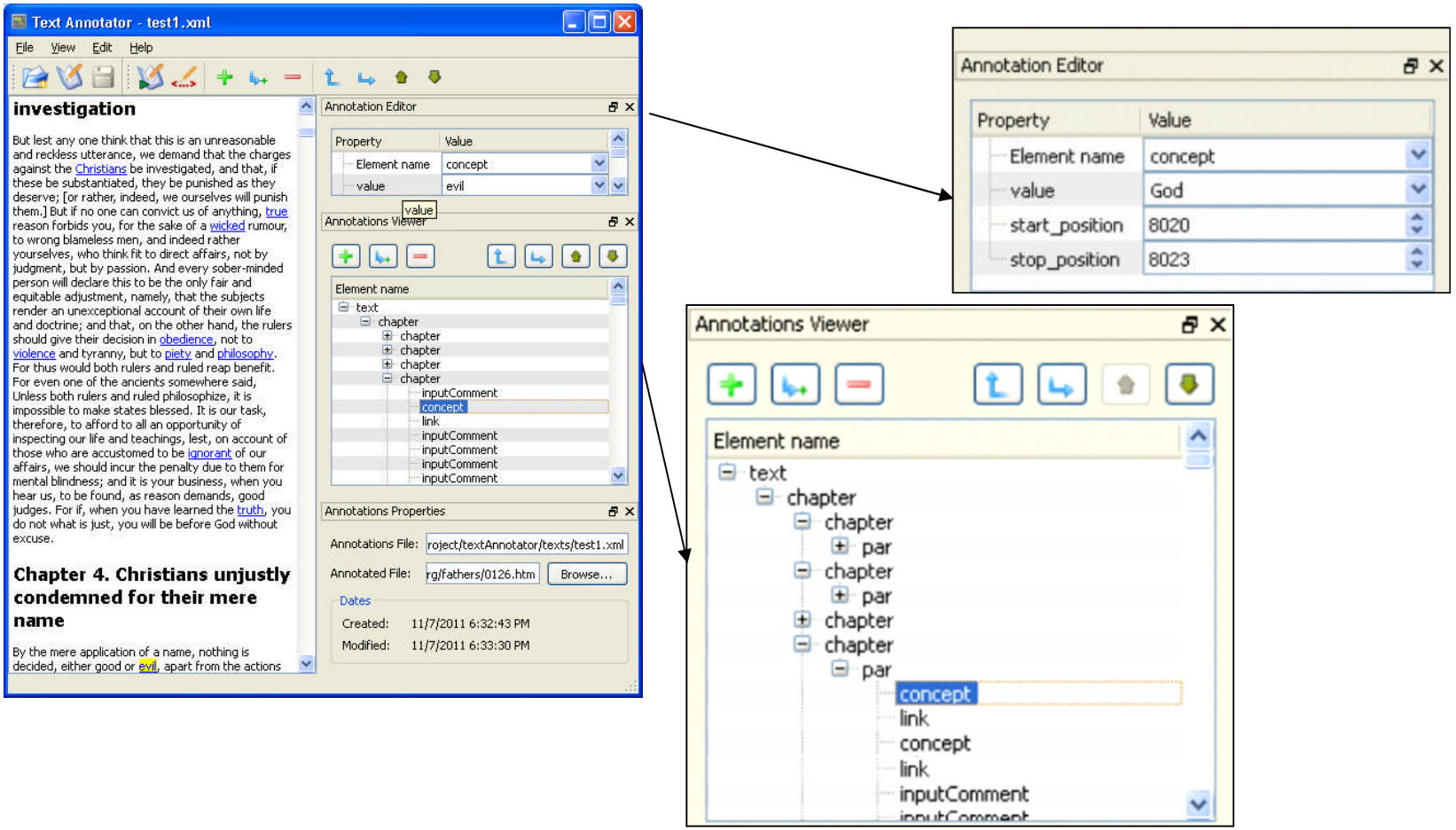
    →learned model

    → automated annotation

Text annotation

- Space
- Time
- Named Entities
- Links
- Syntactic
- Semantic
- Pragmatic
- Discourse
- Rhetoric
- …

Text Annotators



Topic Modeling

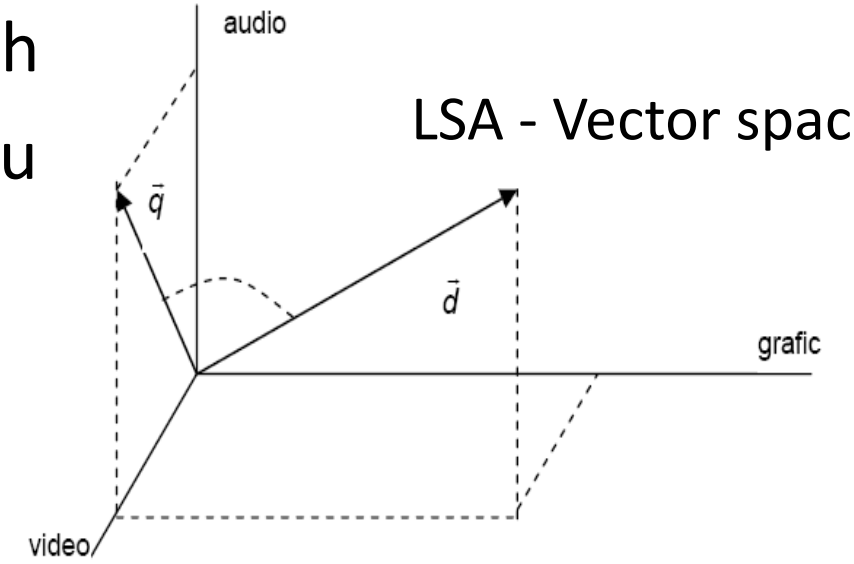
- No generally accepted definition for a “topic” in NLP
  - Document clusters
  - Abstractions based on document clusters
    - Labels;
    - Centroids, etc
  - (Word, Probability) pairs
- Bayesian statistical models
  - Topics – distributions over words
  - Documents – distributions over topics
  - Generative model
  - Topic Intertwining
  - Conceptually similar to the ideas of Mikhail Bakhtin
    - Topics and voices

Topic Modeling (2)

- LSA/pLSA/hLDA/CTM
  - Each newer version corrects some flaws of the earlier ones
- LDA
  - Readily available
    - Mallet
    - Easily reproducible experiments

The LSA idea

- Reducing the dimensionality of the vector space, similarly to the *least squares method*
- The effect is the creation of semantic spaces containing semantically related words
- Bag-of-words approach
- http://lsa.colorado.edu



LSA - Vector space model

Terms-documents array

(ex. from Manning and Schutze, 1999)

$$A = \begin{array}{c|cccccc} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \hline \text{cos } monaut & 1 & 0 & 1 & 0 & 0 & 0 \\ astronaut & 0 & 1 & 0 & 0 & 0 & 0 \\ moon & 1 & 1 & 0 & 0 & 0 & 0 \\ car & 1 & 0 & 0 & 1 & 1 & 0 \\ truck & 0 & 0 & 0 & 1 & 0 & 1 \end{array}$$

Singular value decomposition (SVD)

$$A_{txd} = T_{txn} S_{nxn} D^T_{dxn}$$ n=min(t,d)

$$T^T = \begin{array}{c|ccccc} & \text{cos } monaut & astronaut & moon & car & truck \\ \hline \text{dim}1 & -0.44 & -0.13 & -0.48 & -0.70 & -0.26 \\ \text{dim}2 & -0.30 & -0.33 & -0.51 & 0.35 & 0.65 \\ \text{dim}3 & 0.57 & -0.59 & -0.37 & 0.15 & -0.41 \\ \text{dim}4 & 0.58 & 0.00 & 0.00 & -0.58 & 0.58 \\ \text{dim}5 & 0.25 & 0.73 & -0.61 & 0.16 & -0.09 \end{array} \quad S = \begin{pmatrix} 2.16 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.59 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.28 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.39 \end{pmatrix} \quad D^T = \begin{array}{c|cccccc} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \hline \text{dim}1 & -0.75 & -0.28 & -0.20 & -0.45 & -0.33 & -0.12 \\ \text{dim}2 & -0.29 & -0.53 & -0.19 & 0.63 & 0.22 & 0.41 \\ \text{dim}3 & 0.28 & -0.75 & 0.45 & -0.20 & 0.12 & -0.33 \\ \text{dim}4 & 0.00 & 0.00 & 0.58 & 0.00 & -0.58 & 0.58 \\ \text{dim}5 & -0.53 & 0.29 & 0.63 & 0.19 & 0.41 & -0.22 \end{array}$$

Reduced A

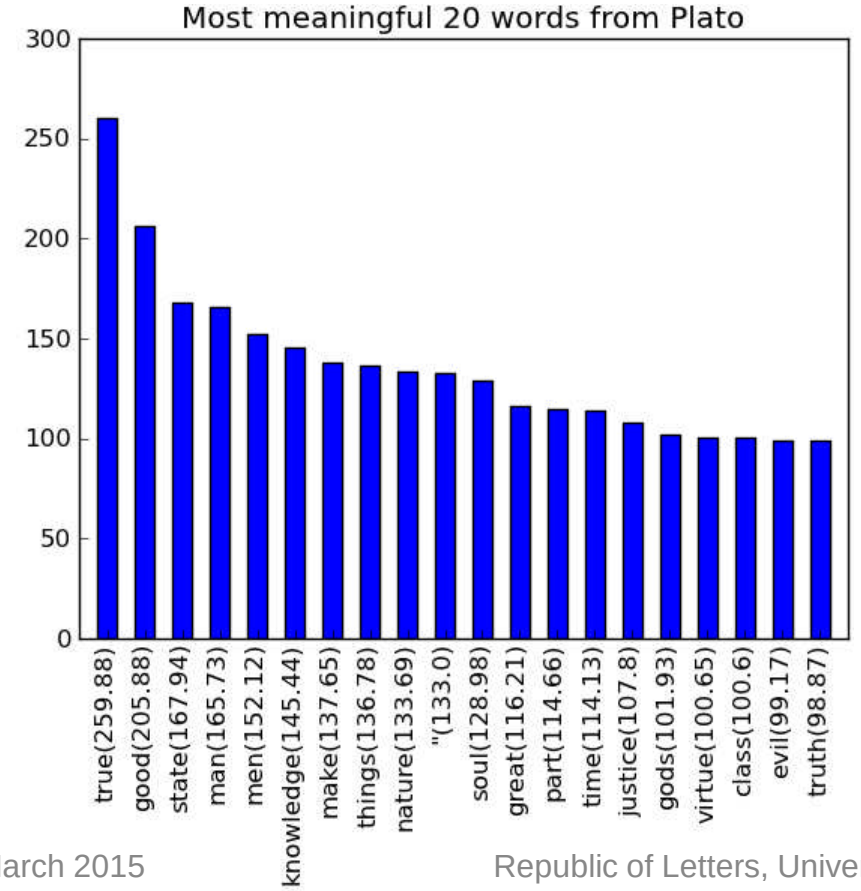
- By SVD on maps the n-dimension space on a k-dimension one, with n >>k
- Common values for k are 100 and 150.

$$B = S_{2x2} D^T_{dx2}$$

$$\Delta = \| A - \hat{A} \|_2$$

$$B = \begin{array}{c|cccccc} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \hline \text{dim}1 & -1.62 & -0.60 & -0.04 & -0.97 & -0.71 & -0.26 \\ \text{dim}2 & -0.46 & -0.84 & -0.30 & 1.00 & 0.35 & 0.65 \end{array}$$

LSA based text processing

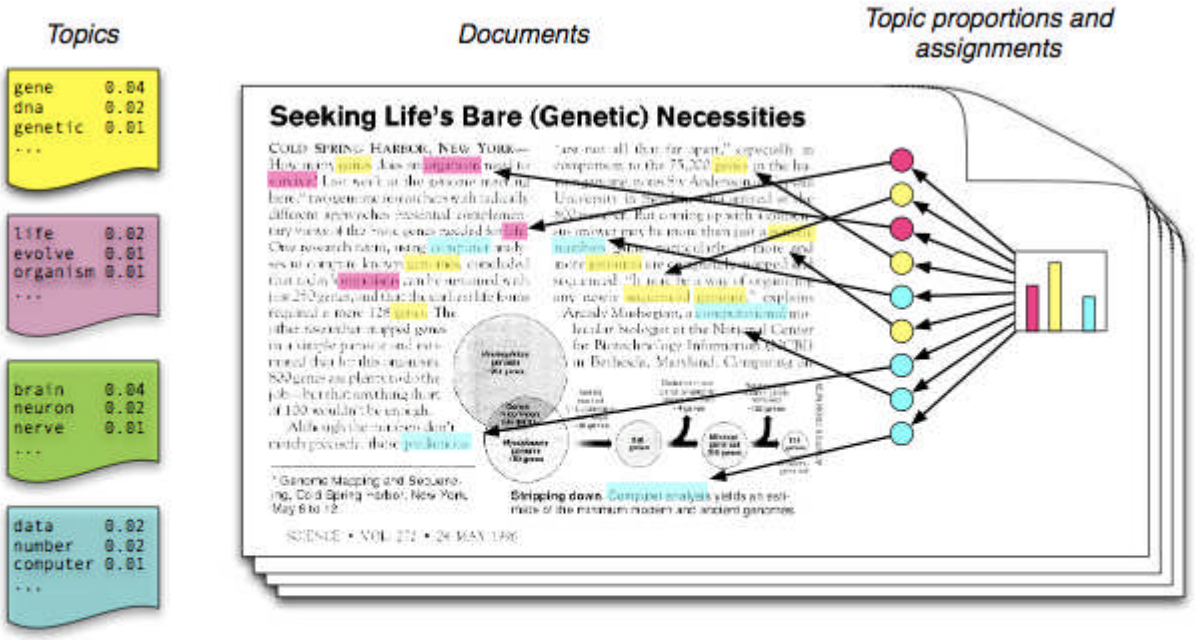


The most significant 20 words from Plato

The similarity of Plato's works with the works of other writers

```
[Plato|TheApology,
Justin|TheSecondApology-(0.6475);
Plato|TheRepublic.7,
Irenaeus|AgainstHeresies.6-
(0.6095)]
```

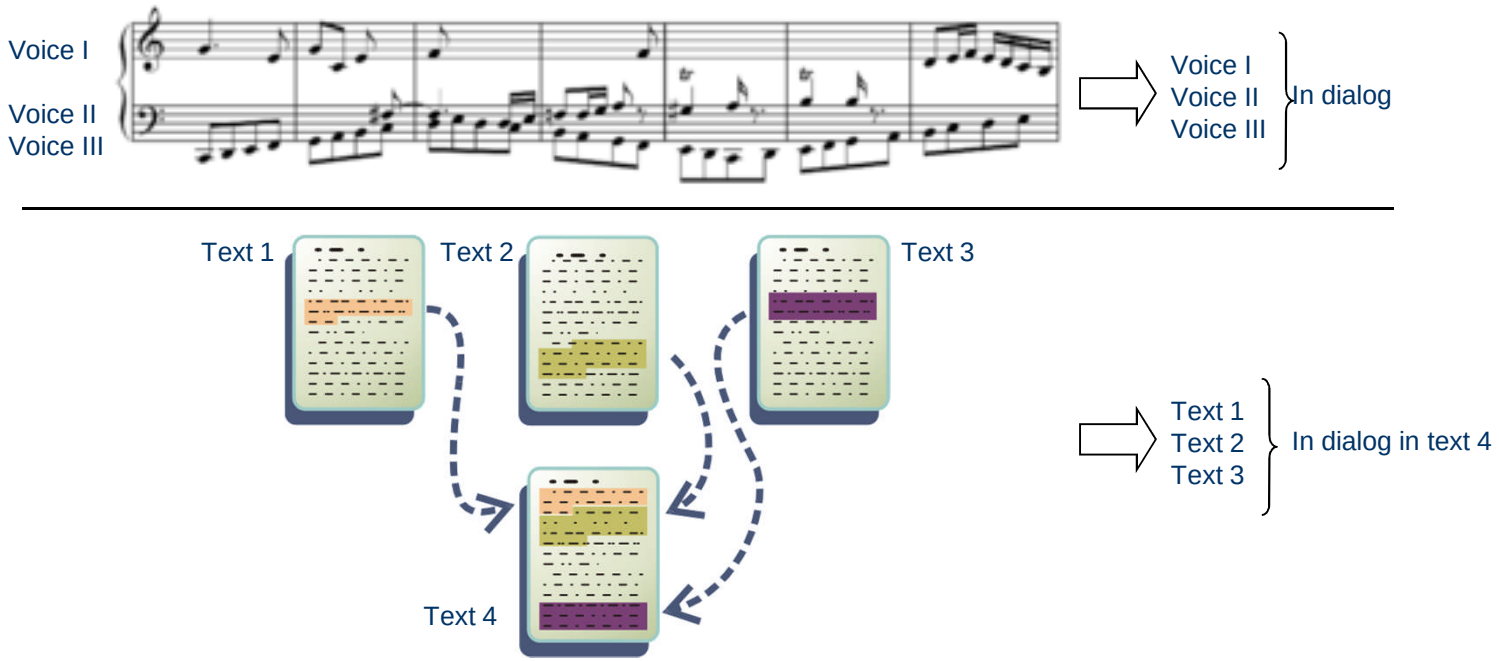
Latent Dirichlet Allocation



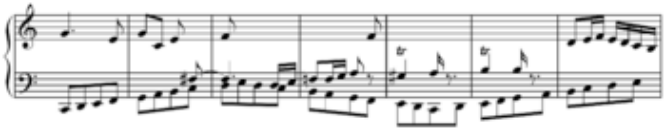
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

http://www.columbia.edu/~ih2240/dataviz/G4063-week5/images/text/LDA.png

Bakhtin’s Polyphonic Intertextuality



Polyphony

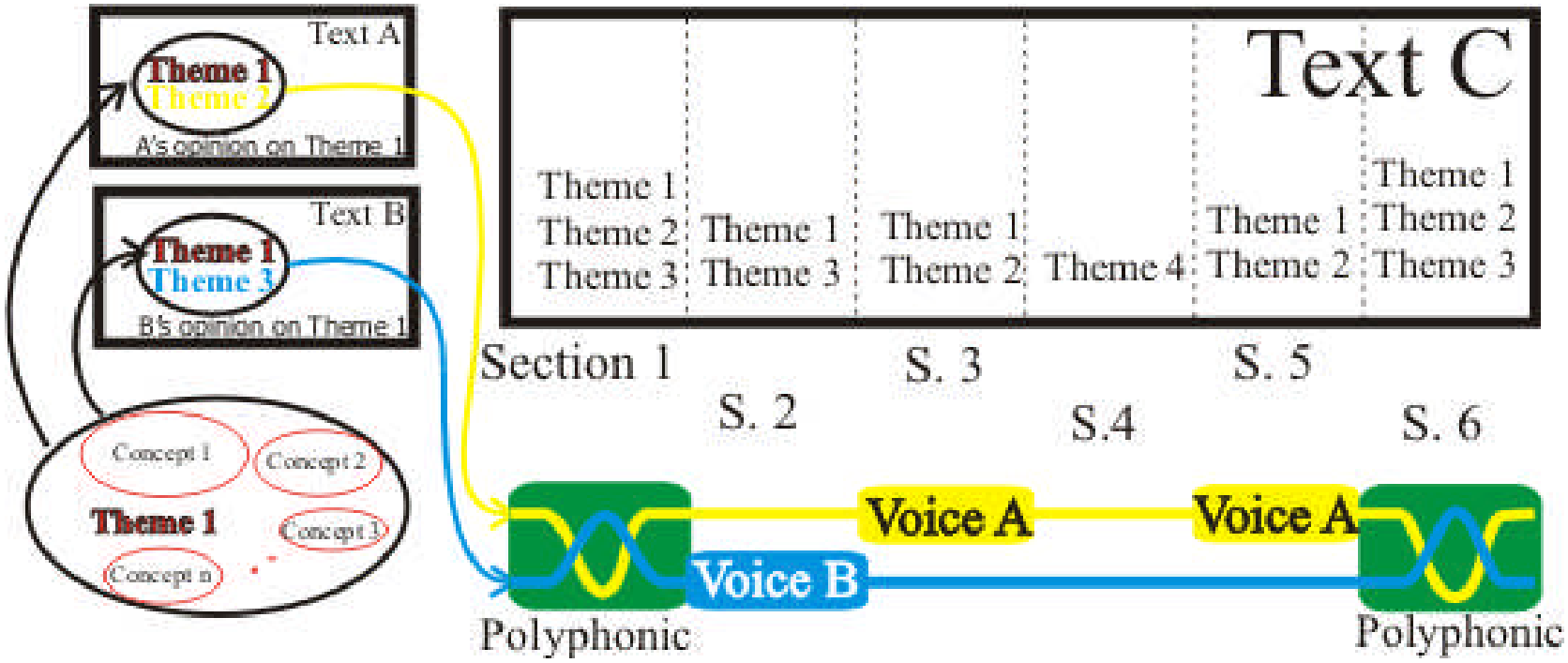


- Appears in music (e.g. J.S.Bach) and in novels (Bakhtin)
- The Polyphonic
  - *Model* (Trausan-Matu, 2005, 2010)
  - *Analysis method* (Trausan-Matu, Dascalu and Rebedea, 2010)
  - *Computer support tools* for the polyphonic analysis of conversations and networks of documents
    - The “Polyphony” system (Trausan-Matu and all, 2007)
    - ASAP (Dascalu, Chioasca and Trausan-Matu, 2008)
    - PolyCAFe (Trausan-Matu, Rebedea and Dascalu, 2011; Rebedea, Dascalu, Trausan-Matu and all, 2010)
    - Collaboration regions detection (Banica, Trausan-Matu and Rebedea, 2011)
    - Detection of the Important moments (Chiru and Trausan-Matu, 2012)
    - Intertextuality detection (Ghiban and Trausan-Matu, 2012)
    - ReaderBench (Dascălu, Trăușan-Matu and Dessus, 2013)

Intertextuality analysis

- Mikhail Bakhtin's dialogistical and polyphonic model → Intertextuality (Kristeva)
- Analyze how concepts are echoed from one text to another (intertextual networks)
- To indicate membership to a philosophical trend or influences among authors

Bakhtin's Polyphonic Intertextuality



Theme 2 and Theme 3 may have the same words but only different concepts

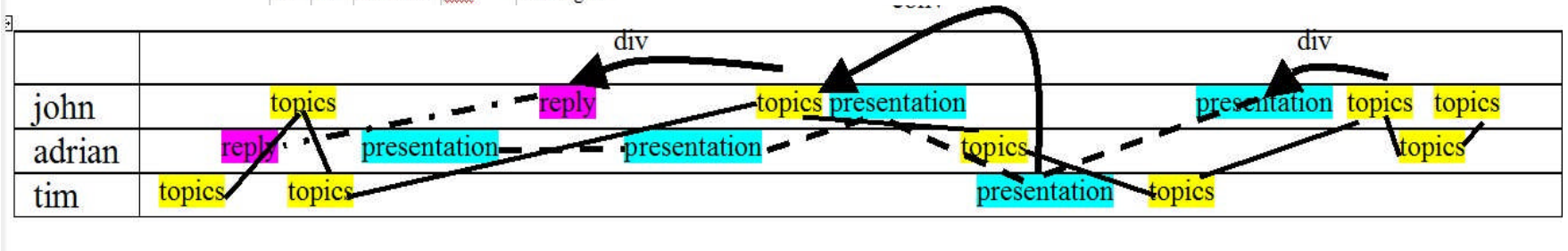
Section 1 and 6 are dialogical or polyphonical. They may present a higher force of expresivity.

PolyCAFe

(Trăușan-Matu, Dascălu and Rebedea)

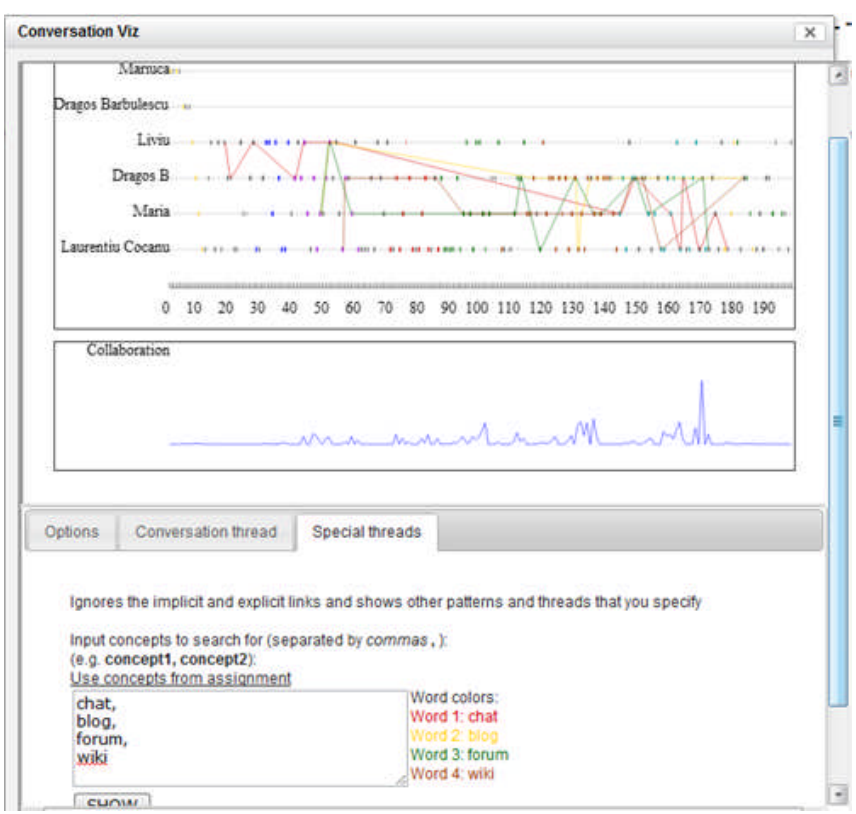
- Polyphony-based Collaboration Analysis and Feedback generation
- Developed in the “Language Technologies for Lifelong Learning” EU FP7 project (http://www.ltfll-project.org/)
- Analyses chat (instant messenger) logs with more than two participants using the polyphonic model (Trăușan-Matu)

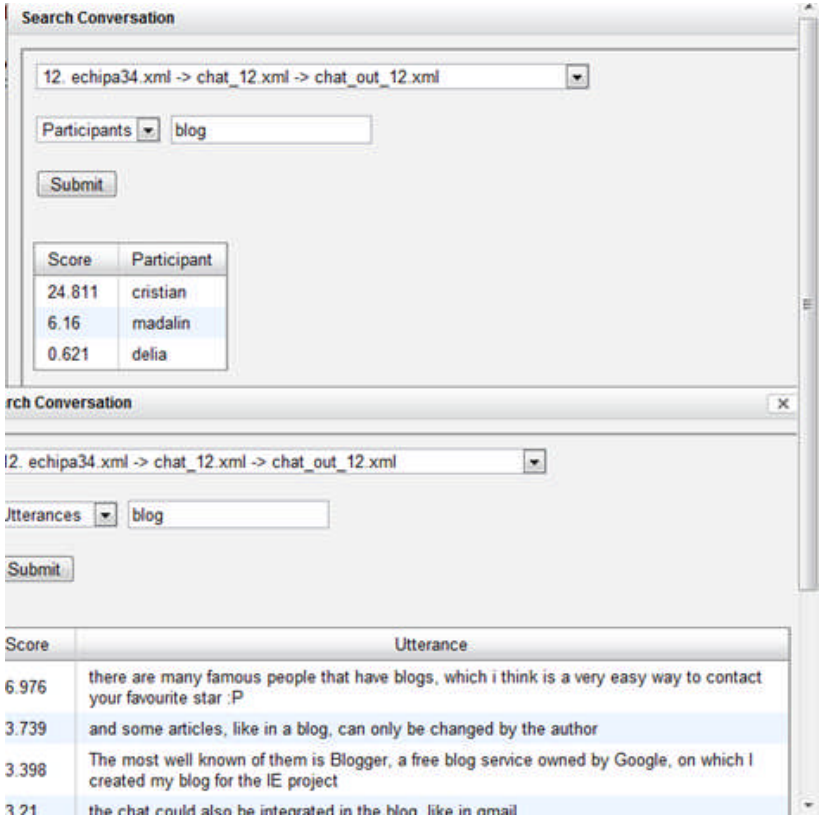
Nr	Ref	Time	User	Text
17		10.26.25	tim	You discussed about a topic separation
18	15	10.26.37	adrian	First of all, the reply method is cumbersome
19	17	10.26.50	john	yes, because we did not like the way the topics were presented in concert chat
20	18	10.26.56	john	yes !!
21	20	10.27.04	john	i hate double-clicking!
22	20	10.27.18	tim	and how can we find topics ?
23	18	10.27.26	adrian	What bothers me is the linear presentation of the discussin
24	23	10.27.43	john	Yep
25	18	10.27.46	adrian	and double-clicking too
26		10.27.54	tim	You mean u want something like a chat forum ? :D
27	24	10.27.58	john	and the reply-to facility is supposed to help you
28	18	10.28.15	adrian	i'd like a tree presentation more
29	18	10.28.38	adrian	or maybe multiple chat columns, for each chat sub-thread
30	27	10.28.58	john	but it is really difficult to use in real-time, because there are so many topics discussed which intertwine each other
31	28	10.29.18	john	i subscribe to a tree-like presentation form
32	P 30	10.29.20	adrian	yes, that's why a clear separation of topics is needed
33	31	10.29.47	adrian	this is easy to implement, no problem here :)
34	30	10.29.49	tim	You need also a clever visual representation
35	30	10.30.05	tim	you'll need also a clever visual interface
36		10.30.22	tim	Who decides the topics ?
37	33	10.30.33	john	i suppose you are refering to the visual representation, right ?
38	37	10.30.45	john	What i would like is a clever way to separate the topics :)
39	38	10.30.59	john	not just doing ot myself, manually
40	37	10.31.00	adrian	Yeah
41	39	10.31.44	adrian	When you start a new thread (a new message, non-related to other message), the app can assume a new topic
42	39	10.31.46	john	i would like the application to be able to detect w topic change all by itself
43	42	10.32.01	tim	That right



From: Trăuşan-Matu , A Polyphonic Model for Interethnic Discourse, 2013

PolyCAFe



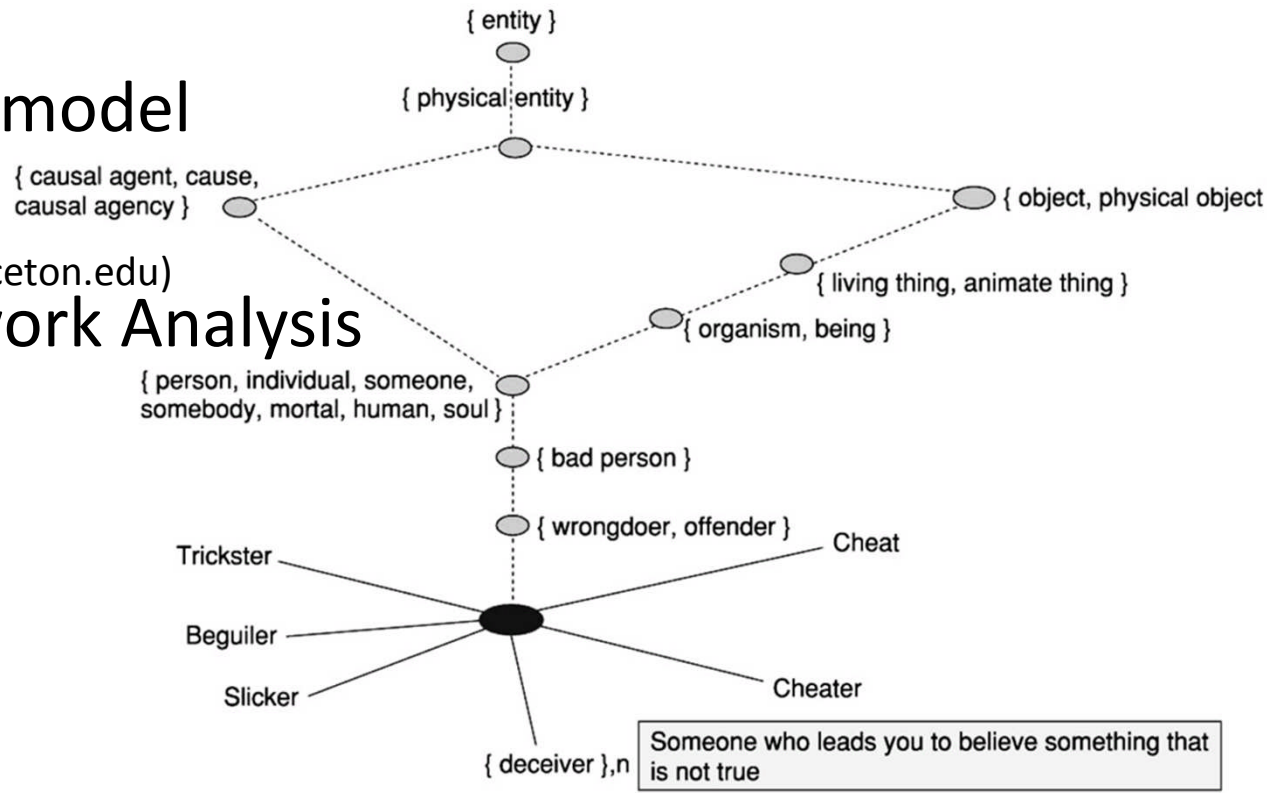


From: Trăuşan-Matu , A Polyphonic Model for Interethnic Discourse, 2013

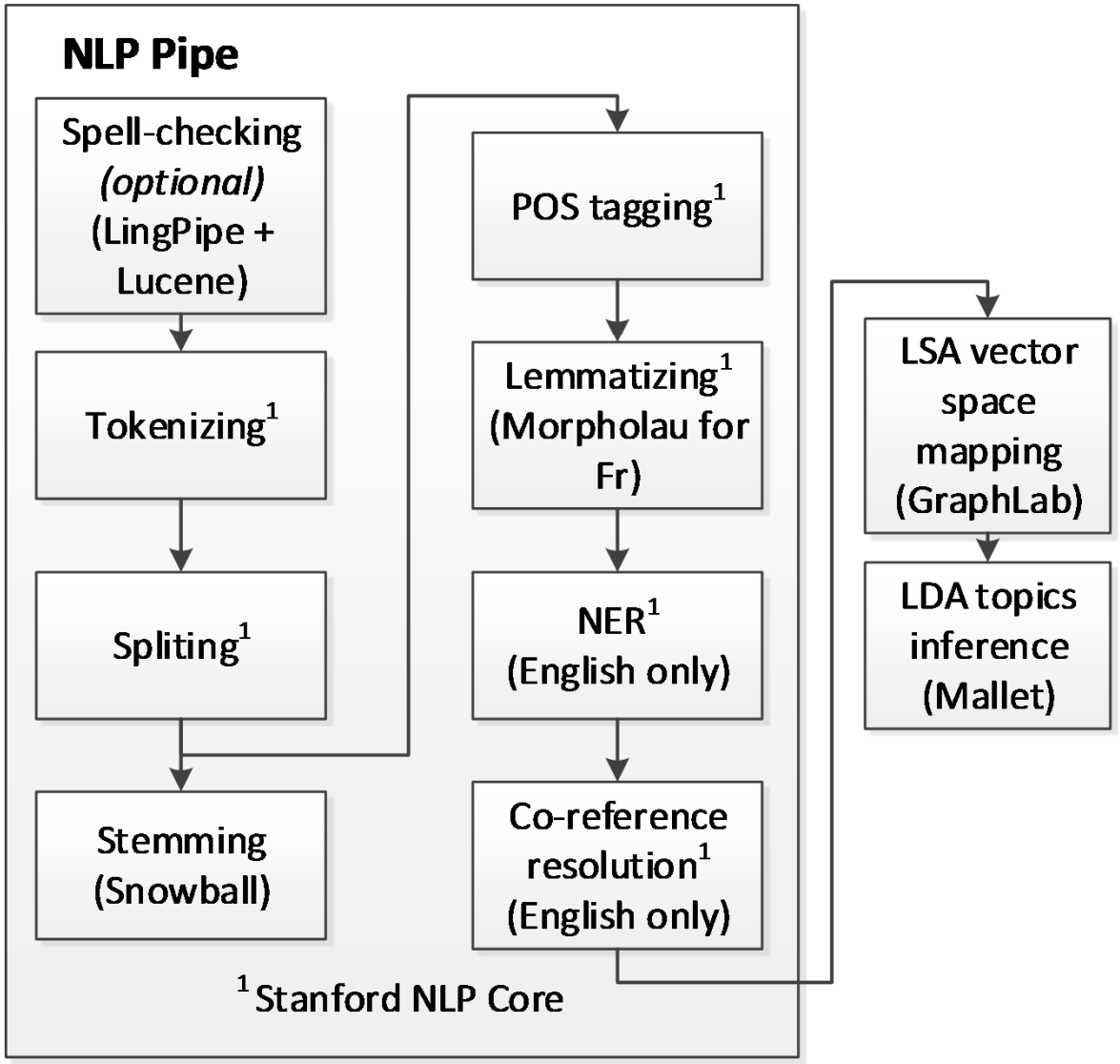
ReaderBench

(Dascalu, Trăușan-Matu and Dessus)

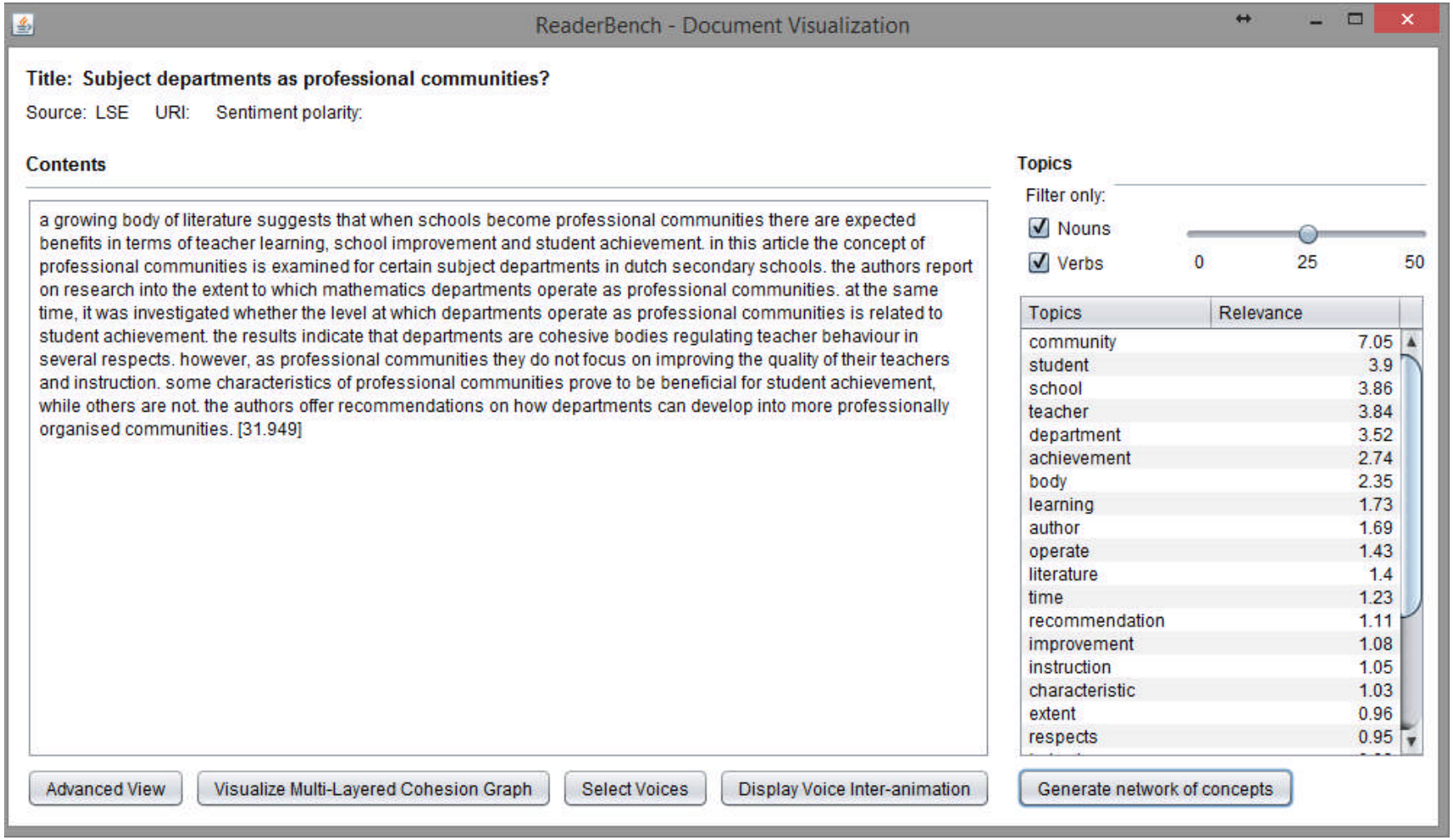
- Based on
  - LSA, LDA
  - Polyphonic model
  - WordNet (http://wordnet.princeton.edu)
  - Social Network Analysis



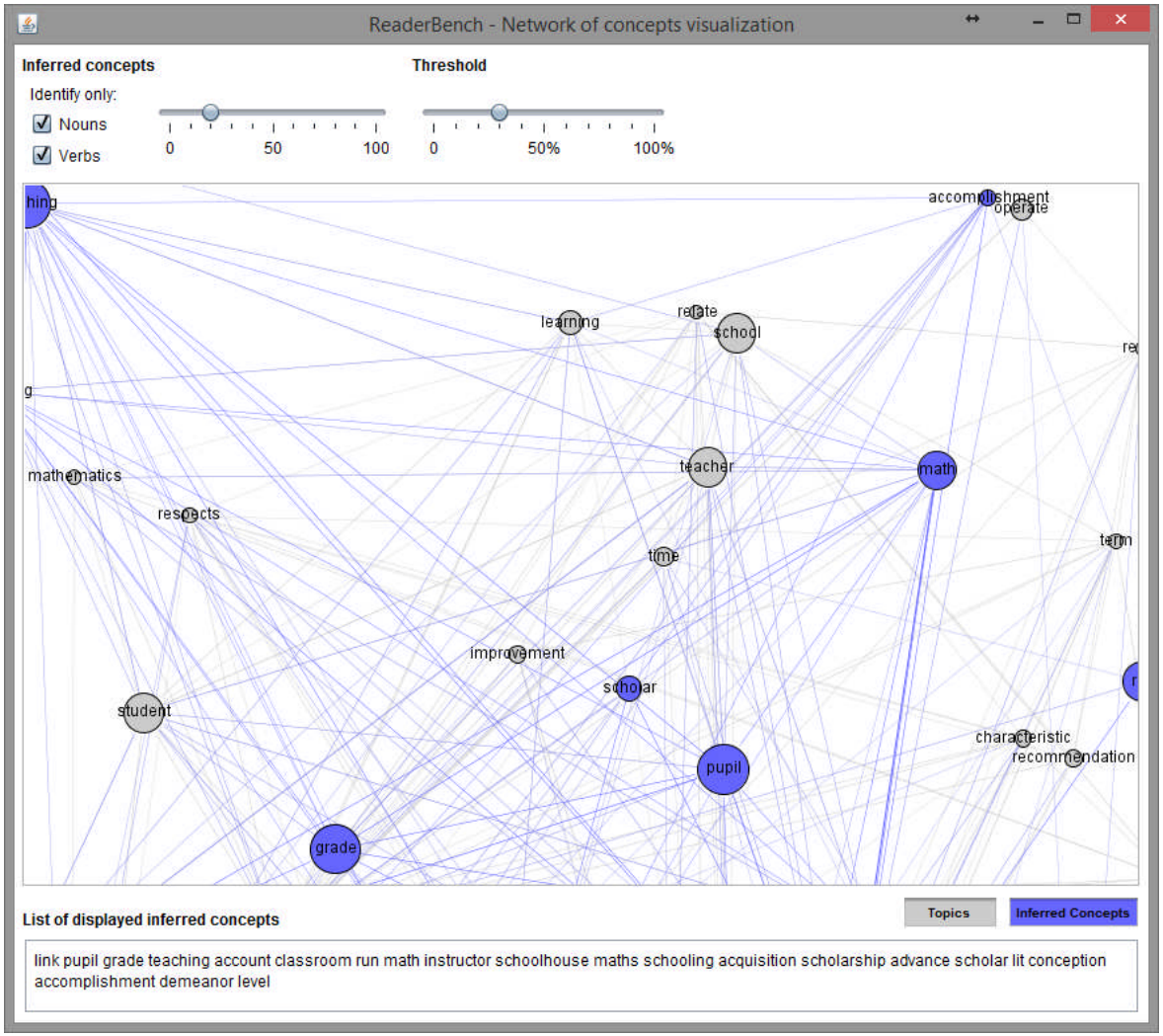
NLP Text pre-processing in PolyCAFe and ReaderBench



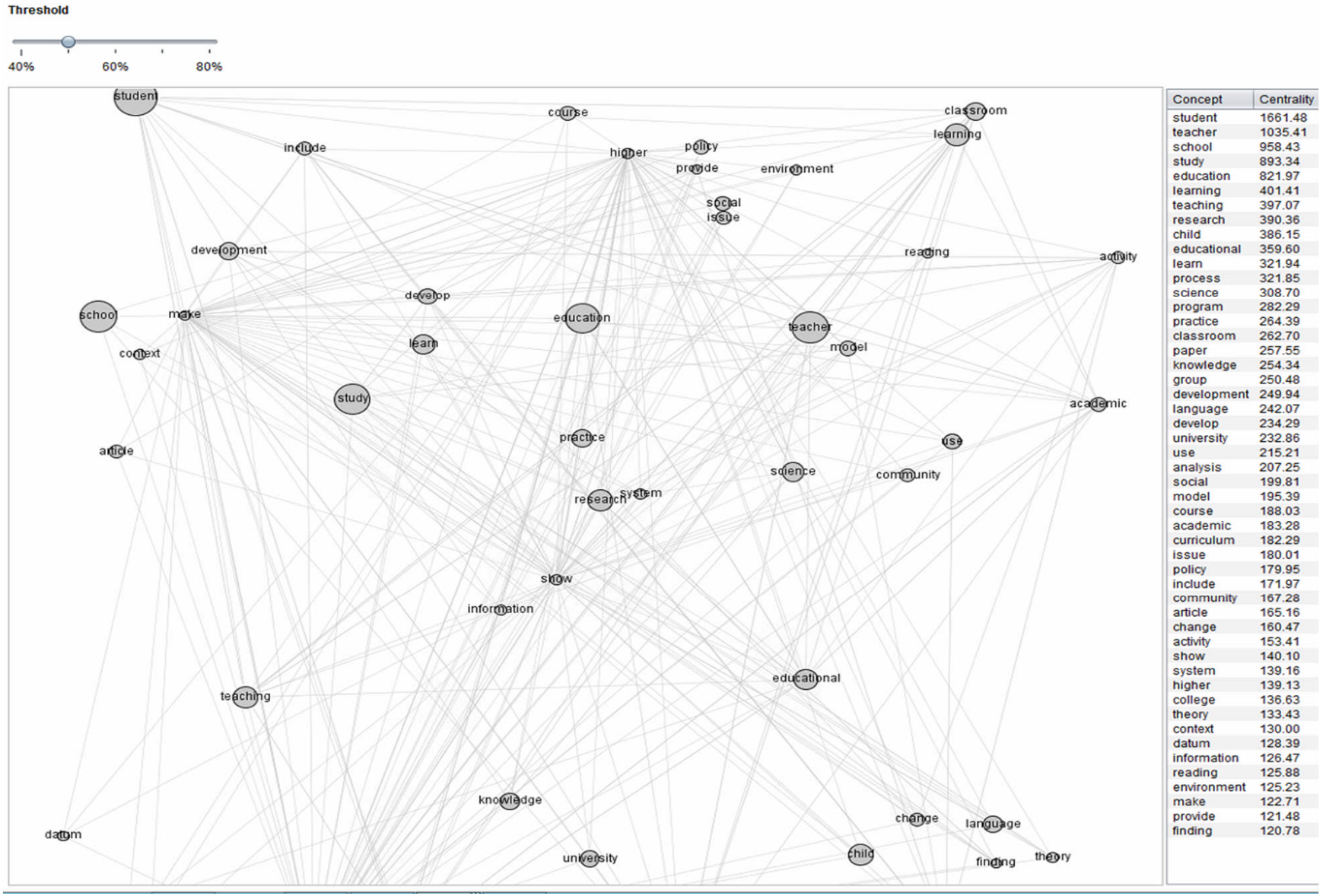
ReaderBench Document view



ReaderBench Concept View

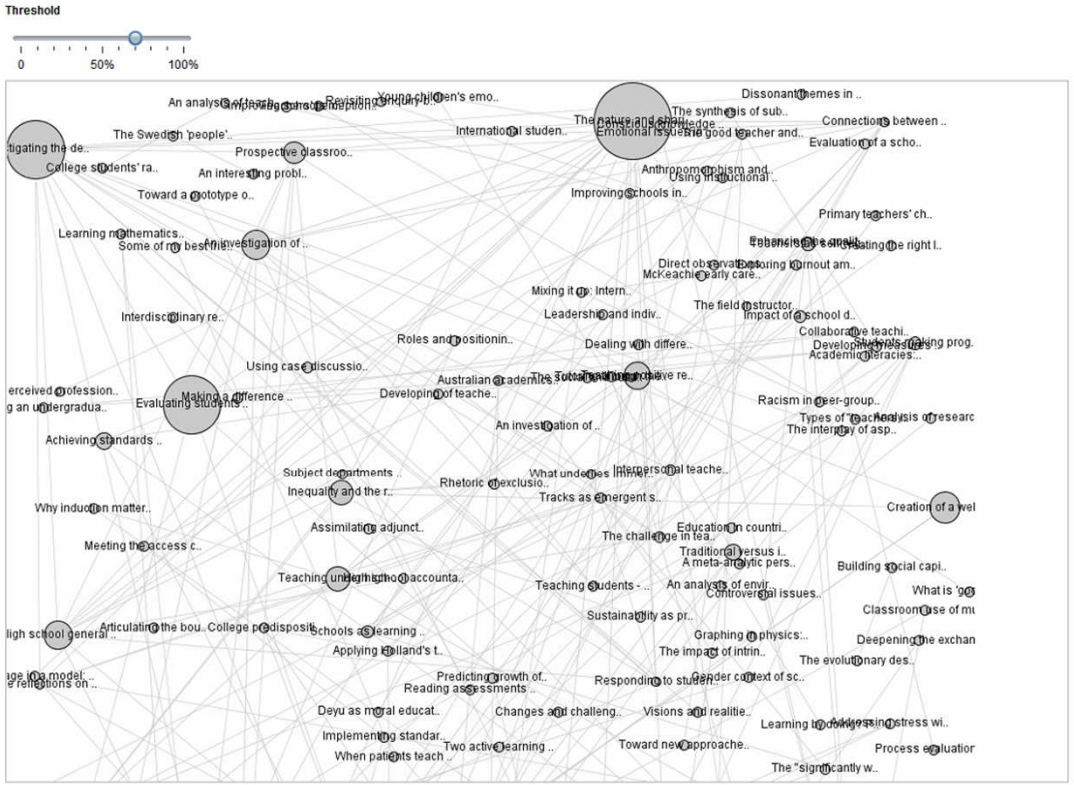


Concept View



Concept	Centrality
student	1661.48
teacher	1035.41
school	958.43
study	893.34
education	821.97
learning	401.41
teaching	397.07
research	390.36
child	386.15
educational	359.60
learn	321.94
process	321.85
science	308.70
program	282.29
practice	264.39
classroom	262.70
paper	257.55
knowledge	254.34
group	250.48
development	249.94
language	242.07
develop	234.29
university	232.86
use	215.21
analysis	207.25
social	199.81
model	195.39
course	188.03
academic	183.28
curriculum	182.29
issue	180.01
policy	179.95
include	171.97
community	167.28
article	165.16
change	160.47
activity	153.41
show	140.10
system	139.16
higher	139.13
college	136.63
theory	133.43
context	130.00
datum	128.39
information	126.47
reading	125.88
environment	125.23
make	122.71
provide	121.48
finding	120.78

ReaderBench Corpus Similarity



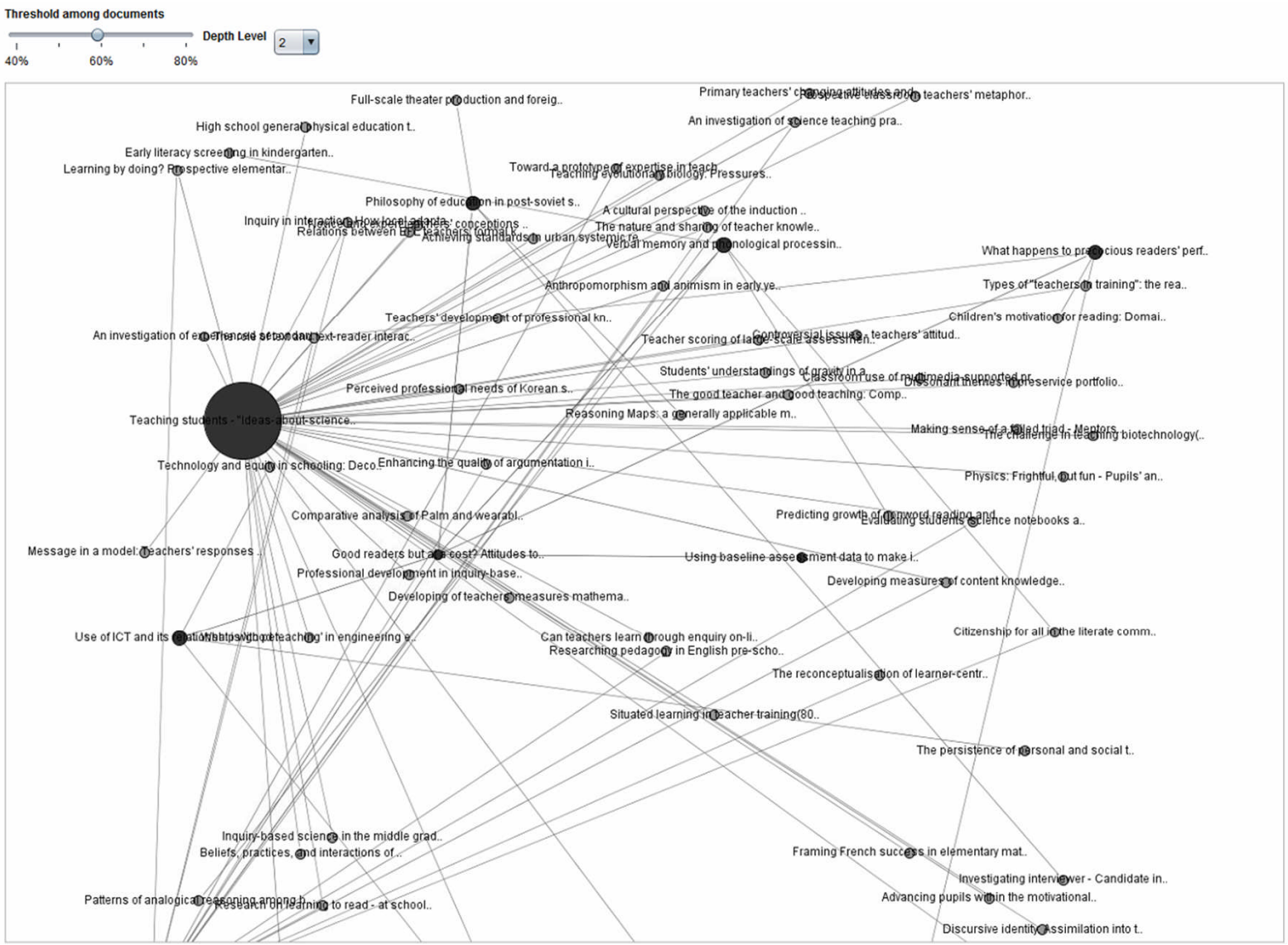
Distance to Central Document

Document	Distance
Novice and expert teachers' conceptions of learners' prior knowledge	0.77
Connections between teacher and student attitudes regarding script choic...	0.77
An investigation of science teaching practices in Indonesian rural seconda...	0.76
Message in a model: Teachers' responses to a court-ordered mandate for...	0.75
Investigating the decision-making skills of cooperating teachers and stud...	0.74
Using case discussion on the web to develop student teacher problem so...	0.73
South Korean high school English teachers' code switching: Questions an...	0.73
The good teacher and good teaching: Comparing beliefs of second-grade ...	0.72
Teaching evolutionary biology: Pressures, stress, and coping	0.72
The challenge in teaching biotechnology	0.72
High school general physical education teachers' behaviors and beliefs a...	0.71
Responding to students' views about university teaching: The experience of...	0.71
Relations between EFL teachers' formal knowledge of grammar and their ...	0.71
The field instructor as group worker: Managing trust and competition in gro...	0.71
The synthesis of subject and pedagogy for effective learning and teaching ...	0.70
Evaluating students' science notebooks as an assessment tool	0.70
Prospective classroom teachers' metaphorical images of selves and com...	0.70
Perceived professional needs of Korean science teachers majoring in che...	0.70
Some reflections on learning from early school experience	0.70
Analysing cognitive or non-cognitive factors involved in the process of phys...	0.70
Developing of teachers' measures mathematics knowledge for teaching	0.70
Developing measures of content knowledge for teaching reading	0.70
Interpersonal teacher behaviour and student outcomes	0.70
Primary teachers' changing attitudes and cognition during a two-year scie...	0.69
Impact of a school district's science reform effort on the achievement and ...	0.69
Inequality and the right to learn: Access to qualified teachers in California's...	0.69
Enhancing the quality of argumentation in school science	0.69
Controversial issues - teachers' attitudes and practices in the context of cit...	0.69
The interplay of aspirations, enjoyment, and work habits in academic ende...	0.68
Teacher scoring of large-scale assessment: professional development or...	0.68
Learning mathematics in a classroom community of inquiry	0.67
What is 'good teaching' in engineering education in India? A case study	0.67
Situated learning in teacher training	0.67
Visions and realities of Internet use in schools: Canadian perspectives	0.67
What underlies immersion students' production: The case of Avoir Besoin ...	0.67
Dissonant themes in preservice portfolio development	0.66
Achieving standards in urban systemic reform: An example of a sixth grad...	0.66
Tutoring through the internet: how students and teachers interact to constr...	0.66
Evaluation of a school-based intervention for HIV/AIDS prevention among ...	0.65
Between idealism and pragmatism: a case study of student teachers' ped...	0.65
Teaching under high-stakes testing - Dilemmas and decisions of a teach...	0.65
Toward new approaches for evaluating student field performance: Tappin...	0.65
Traditional versus integrated preservice teacher education curriculum - A c...	0.65
Teaching students - "Ideas-about-science": dimensions of effective practice	0.65

Top Similar Articles

Article 1	Article 2	Semantic Similarity Score
Developing of teachers' measures mathematics knowledge for teaching	Developing measures of content knowledge for teaching reading	0.86
Investigating the decision-making skills of cooperating teachers and student teachers of English in a Turkish context	Interpersonal teacher behaviour and student outcomes	0.79
South Korean high school English teachers' code switching: Questions and challenges in the drive for maximal use of English in tea...	Connections between teacher and student attitudes regarding script choice in first-year Japanese language classrooms	0.79
Gender context of schooling and levels of stress among early adolescent pupils	Gender organization of schooling and television viewing among early adolescents: a test of two alternative hypotheses	0.79
Teaching under high-stakes testing - Dilemmas and decisions of a teacher educator	Teachers as self-regulated learners	0.78
College students' ratings of student effort, student ability and teacher input as correlates of student performance on multiple-choice e...	An analysis of teachers' rating scales as sources of evidence for a standardised Greek reading test	0.77
The nature and sharing of teacher knowledge of technology in a student teacher/mentor teacher pair	Novice and expert teachers' conceptions of learners' prior knowledge	0.77

ReaderBench Document Centrality



Top similar articles

Article	Similarity
Verbal memory and phonological proce...	0.68
Philosophy of education in post-soviet s...	0.65
Use of ICT and its relationship with perf...	0.62
What happens to precocious readers' p...	0.61
Reading assessments in Kindergarten ...	0.61
Using baseline assessment data to ma...	0.60
Teaching students - "Ideas-about-scien...	0.59
Citizenship for all in the literate commun...	0.58
The persistence of personal and social ...	0.58
Patterns of analogical reasoning amon...	0.56
The role of text and text-reader interactio...	0.56
Early literacy screening in kindergarten: ...	0.56
Predicting growth of nonword reading a...	0.55
Framing French success in elementary ...	0.55
Developing measures of content knowl...	0.55
Research on learning to read - at schoo...	0.55
Researching pedagogy in English pre-s...	0.54
Situated learning in teacher training	0.52
Students' understandings of gravity in a...	0.51
Beyond phonological skills: broader lan...	0.51
Children's motivation for reading: Doma...	0.51
Investigating interviewer - Candidate int...	0.50
Anthropomorphism and animism in earl...	0.50
The challenge in teaching biotechnology	0.50
Can teachers learn through enquiry on-l...	0.50
Impact of a school district's science refo...	0.50
Inquiry in interaction: How local adaptati...	0.50
Teachers' development of professional ...	0.49
Enhancing the quality of argumentation i...	0.49
Relations between EFL teachers' formal...	0.49
A study of science teaching self-efficacy ...	0.49
Professional development in inquiry-ba...	0.48
Primary teachers' changing attitudes an...	0.48
Evaluation of a school-based interventio...	0.48
Prospective classroom teachers' metap...	0.48
Message in a model: Teachers' respon...	0.47
An investigation of science teaching pra...	0.47
Developing of teachers' measures math...	0.47
Beliefs, practices, and interactions of te...	0.47
Evaluating students' science notebooks...	0.47
Full-scale theater production and foreig...	0.47
Inquiry-based science in the middle gra...	0.47
Novice and expert teachers' conception...	0.47
Classroom use of multimedia-supporte...	0.46
Teacher scoring of large-scale assess...	0.46
Toward a prototype of expertise in teachi...	0.46
Learning by doing? Prospective elemen...	0.46
Achieving standards in urban systemic r...	0.45
Types of "teachers in training": the reacti...	0.45
The difficult relationship between theory ...	0.45
Physics: Frightful, but fun - Pupils' and t...	0.44
The nature and sharing of teacher know...	0.44
High school general physical education ...	0.44
Technology and equity in schooling: De...	0.44
Comparative analysis of Palm and wear...	0.44
The good teacher and good teaching: C...	0.44
Teaching evolutionary biology: Pressure...	0.43

Thank you!

Questions?

stefan.trausan@cs.pub.ro
trausan@gmail.com
http://www.racai.ro/trausan