

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

 \rightarrow learned model

 \rightarrow automated annotation

Text annotation

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



Text Annotators

			A rest of the second second second second		
🖻 🏹 🗎 🔰 🏑 + 🖬 🗕	1_ 4 🕈 🖗		Annotation Editor		8×
investigation discussion discus	Annotation Editor		Property	Value	Company of
But lest any one think that this is an unreasonable and reckless utterance, we demand that the charges	Property Value		Element name	concept	~
against the <u>Christians</u> be investigated, and that, if these be substantiated, they be punished as they deserve: [or rather, indeed, we ourselves will punish	value evil		value	God	~
them.] But if no one can convict us of anything, <u>true</u> reason forbids you, for the sake of a <u>wicked</u> rumour,	Annotations Viewer & ×	-		8020	\$
to wrong blameless men, and indeed rather yourselves, who think fit to direct affairs, not by judgment, but by passion. And every sober-minded	÷		stop_position	8023	\$
person will declare this to be the only fair and equitable adjustment, namely, that the subjects	Element name				
render an unexceptional account of their own life and doctrine; and that, on the other hand, the rulers should give their decision in <u>obedience</u> , not to	⊟ chapter ⊕ chapter	Annotations Viewer		₽×	
violence and tyranny, but to <u>piety</u> and <u>philosophy</u> . For thus would both rulers and ruled reap benefit. For even one of the ancients somewhere said, Unless both rulers and ruled philosophize, it is		+ 🖬 🗕	1. L. 🖻		
impossible to make states blessed. It is our task, therefore, to afford to all an opportunity of inspecting our life and teachings, lest, on account of those who are accustomed to be ignorant of our	ink inputComment inputComment	Element name		^	
affairs, we should incur the penalty due to them for mental blindness; and it is your business, when you hear us, to be found, as reason demands, good	inputComment	😑 text		_	
judges. For if, when you have learned the <u>truth</u> , you do not what is just, you will be before God without	Annotations Properties 🗗 🛪	chapter chapter			
excuse.	Annotations File: roject/textAnnotator/texts/test1.xml	🕀 par			
Chapter 4. Christians unjustly	Annotated File: rg/fathers/0126.htm Browse	= chapter			
condemned for their mere	Dates	🛨 par			
name	Created: 11/7/2011 6:32:43 PM	🕀 chapter			
By the mere application of a name, nothing is decided, either good or evil, apart from the actions	Modified: 11/7/2011 6:33:30 PM	🖻 chapter 🖨 par			

concept link

inputComment

innutComment

Republic of Letters, University of Oxford

~

Topic Modeling

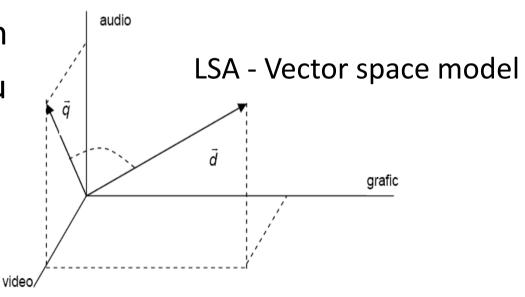
- No generally accepted definition for a "topic" in NLP O Document clusters
 - O Abstractions based on document clusters
 - × Labels;
 - Centroids, etc
 - O (Word, Probability) pairs
- Bayesian statistical models
 - O Topics distributions over words
 - O Documents distributions over topics
 - O Generative model
 - O Topic Intertwining
 - O 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



Terms-documents array

(ex. from Manning and Schutze, 1999)

	(d_1	d_2	d_3	d_4	d_5	d_6
	cos monaut	1	0	1	0	0	0
<u> </u>	astronaut	0	1	0	0	0	0
A –	moon	1	1	0	0	0	0
	car	1	0	0	1	1	0
	truck	0	0	0	1	0	1)

Singular value decomposition (SVD)

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

	(cos monaut	astronaut	moon	car	truck)	(2.16	0.00	0.00	0.00	(0.00)	(d_1	d_2	d_3	d_4	d_5	d_6
	dim1	- 0.44	-0.13	-0.48	-0.70	-0.26		0.00	1.59	0.00	0.00	0.00		dim1	-0.75	-0.28	-0.20	-0.45	-0.33	-0.12
T^T	dim 2	- 0.30	-0.33	-0.51	0.35	0.65	S =	0.00	0.00	1.28	0.00		$D^T =$	dim 2	-0.29	-0.53	-0.19	0.63	0.22	0.41
$I^{*} =$	dim 3	0.57	-0.59	-0.37	0.15	-0.41	5 -					0.00		dim 3	0.28	-0.75	0.45	-0.20	0.12	-0.33
	dim4	0.58	0.00	0.00	-0.58	0.58		0.00	0.00	0.00	1.00	0.00		dim 4	0.00	0.00	0.58	0.00	-0.58	0.58
	dim 5	0.25	0.73	-0.61	0.16	-0.09)	0.00	0.00	0.00	0.00	0.39		dim 5	-0.53	0.29	0.63	0.19	0.41	-0.22)

Reduced A

 By SVD on maps the n-dimension space on a k-dimension one, with n >>k

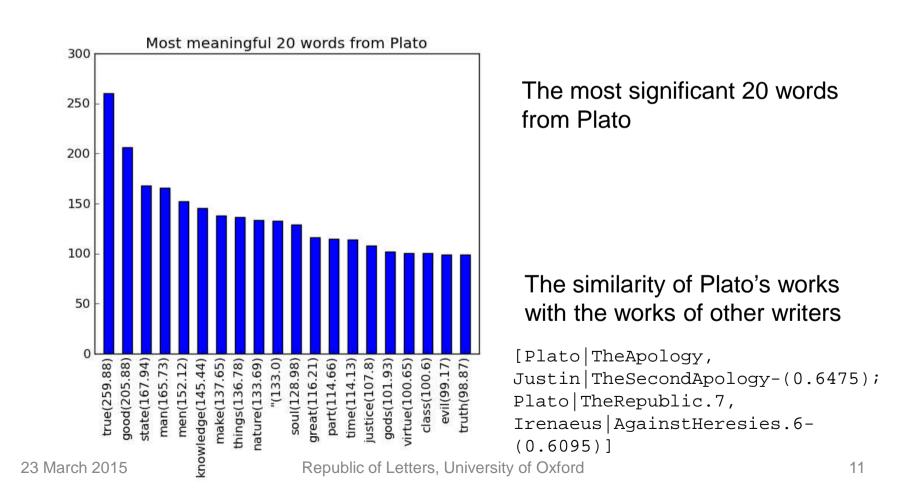
$$B = S_{2x2} D_{dx2}^T$$

• Common values for k are 100 and 150.

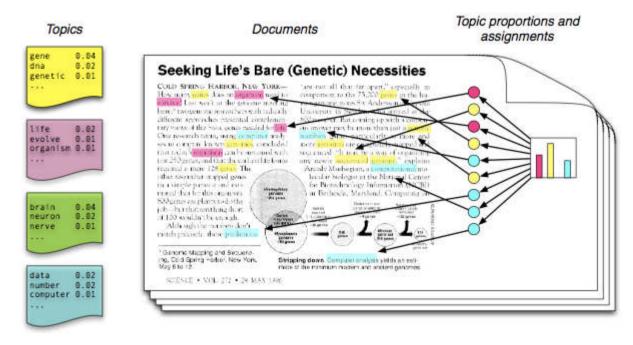
$$\Delta = \parallel A - \widehat{A} \parallel_2$$

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

LSA based text processing



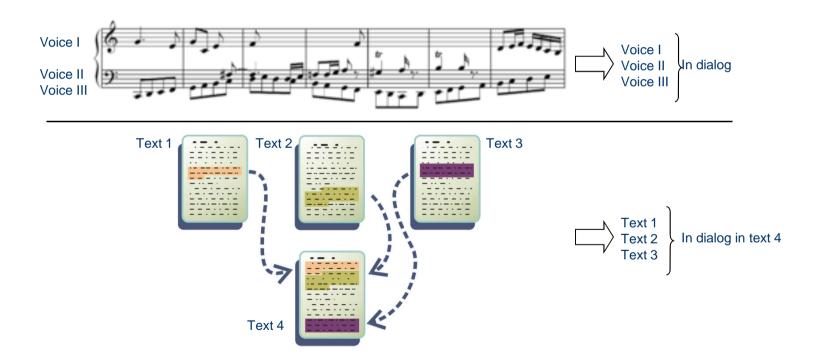
Latent Dirichlet Allocation



- 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/G4063week5/images/text/LDA.png

Bakhtin's Polyphonic Intertextuality



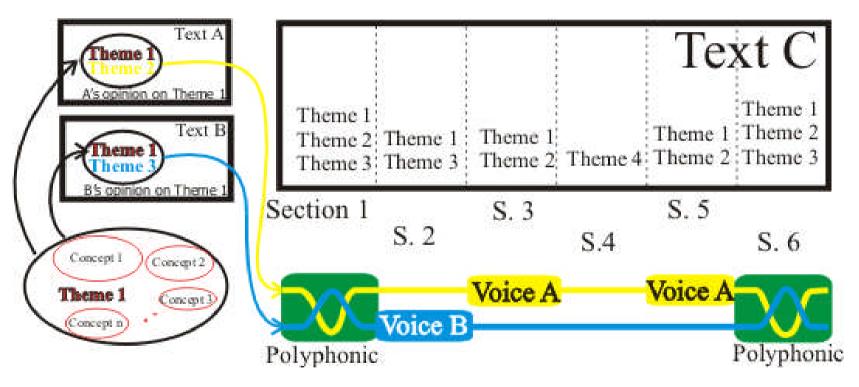
ترد المراجع الم

- 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

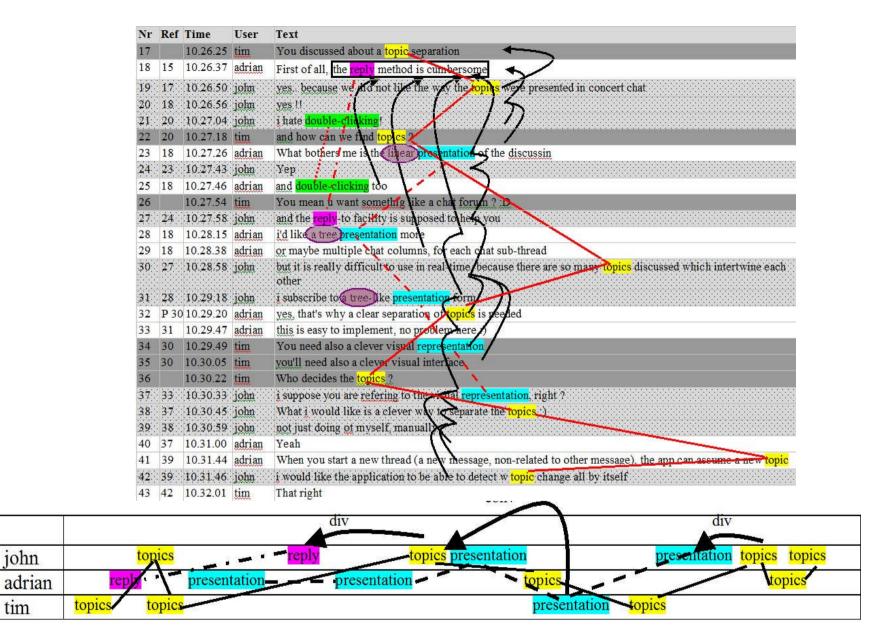


Theme 2 and Theme 3 may have the same words but only different concepts Section 1 and 6 are dialogical or polyphonical. They may present a higher force of expresivity.

PolyCAFe

(Trăușan-Matu, Dascălu and Rebedea)

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



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

tim

PolyCAFe

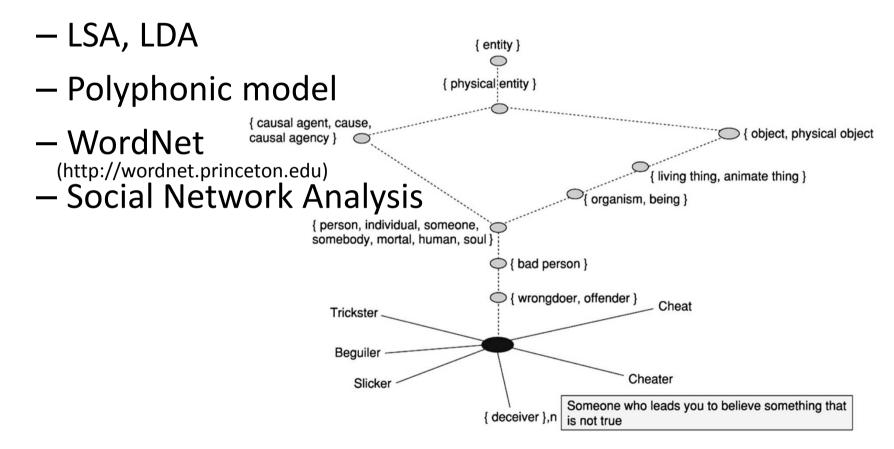
versation Viz		_ Search Conversation
Marinea Dragos Barbulescu Liviu Dragos B Maria		Participants blog Submit
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190		Score Participant 24.811 cristian 6.16 madalin
Collaboration		0.621 delia rch Conversation 12. echipa34.xml -> chat_12.xml -> chat_out_12.xml
Options Conversation thread Special threads	Ī)	Jtterances 💌 blog
Input concepts to search for (separated by commas,):		Score Utterance
(e.g. concept1, concept2): Use concepts from assignment chat, Word colors:		6.976 there are many famous people that have blogs, which i think is a very easy way to contact your favourite star :P
blog, Word 1: chat forum, Word 2: blog		3.739 and some articles, like in a blog, can only be changed by the author
Wiki Word 3: forum Word 4: wiki		3.398 The most well known of them is Blogger, a free blog service owned by Google, on which I created my blog for the IE project
cunw	121	3.21 the chat could also be integrated in the blog like in smail

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

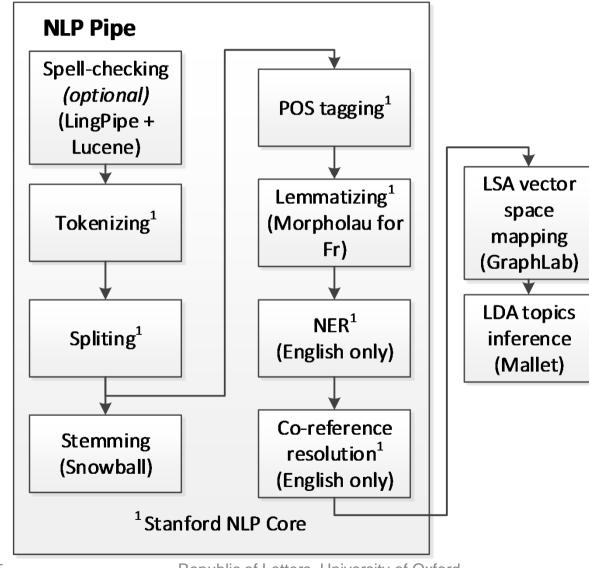
ReaderBench

(Dascalu, Trăușan-Matu and Dessus)

Based on



NLP Text pre-processing in PolyCAFe and ReaderBench

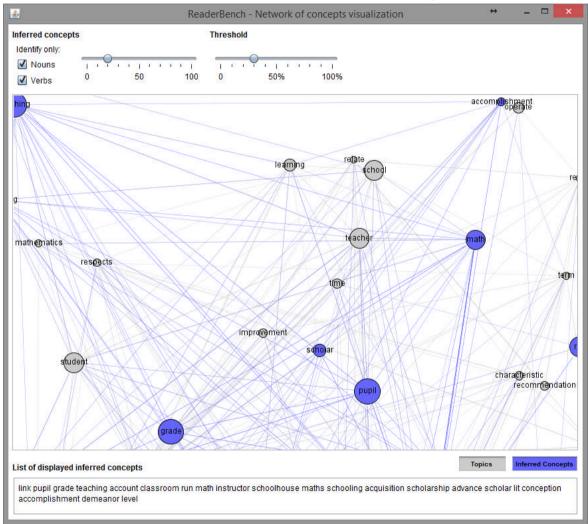


Republic of Letters, University of Oxford

ReaderBench Document view

ReaderBench - Document Visua		
itle: Subject departments as professional communities?		
ource: LSE URI: Sentiment polarity:		
ontents	Topics	
ontents	100000000	
a growing bady of literature suggests that when exhering became professional communities there are	Filter only:	
a growing body of literature suggests that when schools become professional communities there are benefits in terms of teacher learning, school improvement and student achievement. in this article the	V INUUIS	<u> </u>
professional communities is examined for certain subject departments in dutch secondary schools. on research into the extent to which mathematics departments operate as professional communities	the authors report 🗹 Verbs 0	25
time, it was investigated whether the level at which departments operate as professional communities		e
student achievement. the results indicate that departments are cohesive bodies regulating teacher b	Community	7.05
several respects. however, as professional communities they do not focus on improving the quality o	otodonit	3.9
and instruction, some characteristics of professional communities prove to be beneficial for student	School	3.86
while others are not, the authors offer recommendations on how departments can develop into more	professionally teacher	3.84
organised communities. [31.949]	department	3.52
	achievement	2.74
	body	2.35
	learning	1.73
	author	1.69
	operate	1.43
	literature	1.4
	time	1.23
	recommendation	1.11
	improvement	1.08
	instruction	1.05
	characteristic	1.03
	extent	0.96

ReaderBench Concept View

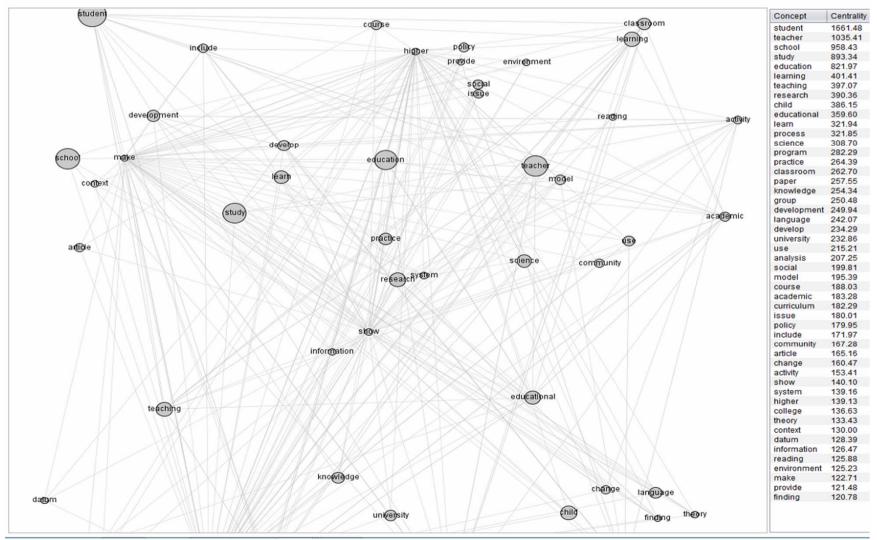


Concept View

Threshold

· · · · ·

40% 60% 80%



ReaderBench Corpus Similarity

hreshold			Distance to Central Document	
An analysize - firest	- Revisiting Children's emo	Dissonant@nemes in	Document	Distance
The Swedigh people'. pating the de Prospective College sOcients' ra An interesOng proble Toward a goototype o)assroo 	synthesis of sub. Connections between Be bood teacher andEvaluations f a scho mg Philipping and Primary tagchers' ch	Novice and expert teachers' conceptions of learners' prior knowledge Connections between teacher and student attitudes regarding script choic An investigation of science teaching practices in indnoesian rural seconda Message in a model. Teachers' responses to a court-ordered mandate for, Investigation the decision-making stills of cooperating teachers and stud Using case discussion on the web to develop student teacher problem so South Korean high school English teachers' code switching; Questions an The good teacher and good teaching; Comparing beliefs of second-grade	0.77 0.77 0.76 0.75 0.74 0.73
Some of no best nisive tigation of	Direct McKeach Mixing it () : Intern.	Enbeneis@isseneitating@e right I ob @rvatbestoring t@rnout am @safty care	Teaching evolutionary biology: Pressures, stress, and coping The challenge in teaching biotechnology. High school general physical education teachers' behaviors and beliefs a Responding to students' views about university teaching: the experience OL. Relations between EFL teachers' formal knowledge of grammar and their	0.72 0.71 0.71
Interdisc@jinary.re Using.cas	Roles and position in Dealing with differe	The field @structor. Impact of @school d. Collaborative teachi. Developited profile and a school of the school of	The field instructor as group worker. Managing trust and competition in gro The synthesis of subject and pedagogy for effective learning and teaching Evaluating students' science notebooks as an assessment tool Prospective classroom teachers' metaphorical images of selves and com	0.71 0.70 0.70 0.70
ceived golession an uguergradua Achieving Spindards	Australian @rademithe 3dtbaitendbaarcheve re. Developin@of teache An invest@ation of	Racism in geer-group Types of "gac Hooly sis of gesearc The intergy of asp.	Perceived professional needs of Korean science teachers majoring in che Some reflections on learning from early school experience Analysing cognitive or non-cognitive factors involved in the process of phys Developing of teachers' measures mathematics knowledge for teaching Developing measures of content knowledge for teaching reading Interpersonal teacher behaviour and student outcomes	0.70
	oject demantments . What underDes [APERP2/Sorgal to equal(v and the r. Rhetoric mexclusio. Tracks as mergent s.	Creation of a wel	Primary teachers' changing attitudes and cognition during a two-year scie Impact of a school district's science reform effort on the achievement and Inequality and the right to learn: Access to qualified teachers in California's Enhancing the quality of argumentation in school science	0.69 0.69 0.69 0.69
Meeting th@access c Tear	Tr. A	ucation (Dn countri. 182 meta-andivite pers Building social capi Controversigni issues What is 'goo	Controversial issues - teachers' attituides and practices in the context of ct The interplay of aspirations, enjoyment, and work habits in academic ende Teacher scoring of large-scale assessment professional development or Learning mathematics in a classroom community of inquiry What is 'good teaching' in engineering education in funda? A case study	0.68
h school generarticulating the bouCollege p@dispo	sitschools @earning	Classroom@se of mi Graphing @physics: Deepening@se exchan npage of infrin The evolut@nary des.	What is good learning in explorer many education in molar A case study Situated learning in teacher training Visions and realities of Internet use in schools: Canadian perspectives What underlies immersion students' production: The case of Avoir Besoin Dissonant themes in preservice portfolio development	0.67 0.67 0.67 0.67 0.66
Pen Binonseth	Deyu as mogral educat. Changes and challeng. Visions a Implementing standar. Two active caring Toward new Dates and the standard toward new Dates and toward new Dates and the standard toward new Dates and the standard toward new Dates and the standard toward new Dates and	Gender cantext of sc Ord realitie Learning top:dwikkgeqsing/stress wi	Achieving standards in urban systemic reform: An example of a sixth grad Tutoring through the internet how students and teachers interact to constru- Evaluation of a school-based intervention for HIV/ADS prevention among Between idealism and pragmatism: a case study of student teachers' ped Teaching under high-stakes testing - Dilemmas and decisions of a teach Toward new approaches for evaluating student field performance: Tappin	0.65 0.65
an Similar Articlas	When patients teach.	The "sign@cantly w	Traditional versus integrated preservice teacher education curriculum - A c Teaching students"Ideas-about-science": dimensions of effective practice	

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 sharinn of teacher knowledge of technology in a student teacher/mentor teacher nair	Novice and exnert teachers' concentions of learners' nrior knowledge	0.77

Republic of Letters, University of Oxford

ReaderBench Document Centrality

Id among documents	
Depth Level 2	
60% 80%	
	rtion and forein Primary teachers' c ងលុស្ក្ រស្នាល់ផ្លាំមារមិនទូលិប៉ូត្រូ teachers' metaphor
Full-scale theater p@du	
High school genera@hysical education t.	An investigation of gence teaching pra
Early literacy screeding in kindergarten	
Learning by doing? Respective elementar	Toward a prototype of expertise in tage Pressures
Philosophy of educ	
Relations between HEL Achieving	centions The nature and shareborn the induction Staffabros within system of the memory and person of the model of the staffabros o
	Verbal memory and performance of the second
	Anthropomorphism and animism in early ye Types of "feachers of training": the rea
	pment of professional kn
An investigation of emeriescele setendan mext-reader interac.	Teacher scoring of lager scales set and the score attitud.
	Students' understantions of gravity in a
Perceived profession	students' understate ings of grass of more of more and the provided the provided the service portfolio
	Reasoning Maps: a generally applicable m
Teaching students - "Ideas-about-science	Making senser Marting and Mathematic mark
Technology and equip in schooling: Deco Enhancing the qua	lim of argumentation i Physics: Frightful, @ut fun - Pupils' an
Comparative analysis Palm and	wearabl Predicting growth of the start of
	Evaluating students @Lience notebooks a
age in a model: Deachers' responses Good readers but a cost	Attitudes to Using baseline asse ment data to make i
Professional development in indu	Developing measures of content knowledge
Developing of te	acher@measures mathema
Use of ICT and its mation betrowy book teaching 'in engineering e.	Can teachers learn @rough enquiry on-li Citizenship for all i@the literate comm Researching pedago@ in English pre-scho
	The reconceptualis (t) on of learner-centr
	Situated learning ingeacher training (80
	Sindate deaning ingesting adminigiou.
	The persistence of personal and social t
	$\times / / \times \mathbb{V}$
Inquin-based sciente in the middle grad.	
Beliefs, practices, and interactions of	Framing French success in elementary mat.
	Investigating interviewer - Candidate in
Patterns of analogicatreasessing aronnearing to read - at school	Advancing pupils within the motivational
Research of realing to read at School.	
	Discursive identition into t.

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	0.52
Beyond phonological skills: broader lan	0.51
	0.51
Children's motivation for reading: Doma	
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-I	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	0.44

Thank you!

Questions?

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