RusNLP: Semantic search engine for Russian NLP conference papers

Irina Nikishina[†‡], Amir Bakarov[*†], Andrey Kutuzov[§]

[†] National Research University Higher School of Economics, Moscow, Russia

[‡] Ivannikov Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia

[*] Federal Research Center 'Computer Science and Control' of the Russian Academy of Sciences, Moscow, Russia

[§] University of Oslo, Oslo, Norway

AIST
5 July 2018

Contents

What is RusNLP ?

- RusNLP is a web service :
- search engine and recommendation system...
- ...over proceedings of three major Russian NLP conferences :
  1. **Dialogue**
  2. **AIST**
  3. **AINL**

What is RusNLP ?

- RusNLP is a web service :
- search engine and recommendation system...
- ...over proceedings of three major Russian NLP conferences :
  1. **Dialogue**
  2. **AIST**
  3. **AINL**
- 18 years of publishing ;

What is RusNLP ?

- RusNLP is a web service :
- search engine and recommendation system...
- ...over proceedings of three major Russian NLP conferences :
  1. **Dialogue**
  2. **AIST**
  3. **AINL**
- 18 years of publishing ;
- 400 academic papers in English ;

What is RusNLP ?

- RusNLP is a web service :
- search engine and recommendation system...
- ...over proceedings of three major Russian NLP conferences :
  1. **Dialogue**
  2. **AIST**
  3. **AINL**
- 18 years of publishing ;
- 400 academic papers in English ;
- metadata (titles, abstracts authors, affiliations) extracted automatically and normalized manually ;
- `http://nlp.rusvectores.org`
- Search for yourself or your institution :-)
- NB : we do not provide full texts of the papers, but there are hyperlinks to source URLs.

What is RusNLP ?

AIST

ARTIFICIAL INTELLIGENCE
& NATURAL LANGUAGE

2002-2017

`http://dialog-21.ru`

281 papers in English

Source : conference website

2014-2017

`https://aistconf.org`

45 papers in English

Source : Springer

2015-2017

`http://ainlconf.ru`

67 papers in English

Source : Springer

The dataset is described in details in [Bakarov et al., 2018].

Contents

Typical use cases

Generally, *RusNLP* can be used to :

- Discover academic knowledge you were not aware of :
  - may be someone has already done what you are doing ?

Typical use cases

Generally, *RusNLP* can be used to :

- Discover academic knowledge you were not aware of :
  - may be someone has already done what you are doing ?
- Identify ‘gaps’ in Russian NLP, where we still lack knowledge ;

Typical use cases

Generally, *RusNLP* can be used to :

- Discover academic knowledge you were not aware of :
  - may be someone has already done what you are doing ?
- Identify 'gaps' in Russian NLP, where we still lack knowledge ;
- Analyze academic communities and publishing patterns :
  - who published what and where ?
  - sort of 'who is who' for the Russian NLP community.

Typical use cases

Questions we can answer

- '*I know this paper, what other similar papers are there in Russian NLP ?*'

Typical use cases

Questions we can answer

- '*I know this paper, what other similar papers are there in Russian NLP ?*'
- '*What was published in 2008 by NLP scholars from Moscow State University ?*'

Typical use cases

Questions we can answer

- '*I know this paper, what other similar papers are there in Russian NLP ?*'
- '*What was published in 2008 by NLP scholars from Moscow State University ?*'
- '*Were there any papers about paraphrases detection at the **AINL** conference in 2015 ?*'
- etc...

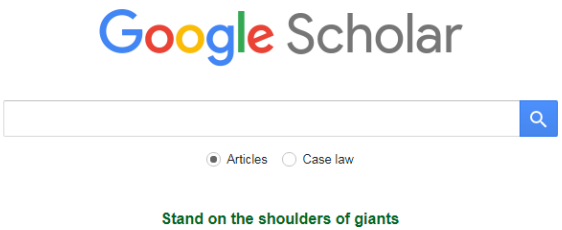
Contents

What is novel in that ?

There are many scholar search engines available :

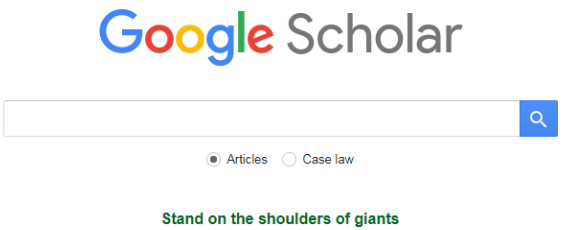
What is novel in that ?

There are many scholar search engines available :



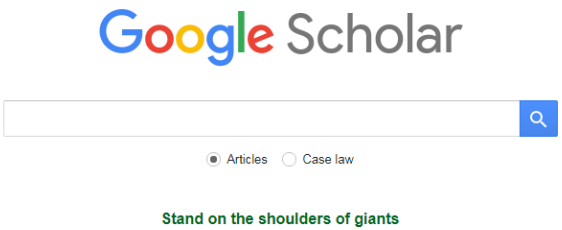
What is novel in that ?

There are many scholar search engines available :



What is novel in that ?

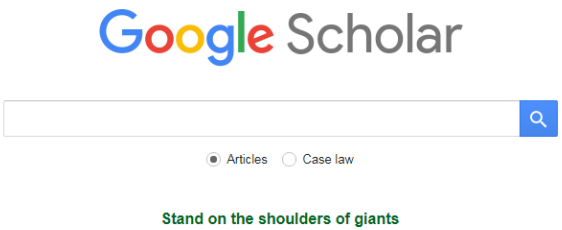
There are many scholar search engines available :



...etc

What is novel in that ?

There are many scholar search engines available :



...etc

What's different about us ?

Feature	**Google Scholar** `scholar.google.com`	**ArXiv Sanity** `arxiv-sanity.com`	**RusNLP** `nlp.rusvectores.org`
International coverage			
Recommendation system			
Manually parsed metadata			
Focus on NLP			

What is novel in that ?

Vladimir Khoroshevsky started the analysis of Russian NLP landscape [Khoroshevsky, 2012]. Compared to them, the *RusNLP* project :

1. uses much more up-to-date data and includes novel publishing venues (AIST and AINL).

What is novel in that ?

Vladimir Khoroshevsky started the analysis of Russian NLP landscape [Khoroshevsky, 2012]. Compared to them, the *RusNLP* project :

1. uses much more up-to-date data and includes novel publishing venues (AIST and AINL).
2. Implements the semantic search engine available online as a web service.

What is novel in that ?

Vladimir Khoroshevsky started the analysis of Russian NLP landscape [Khoroshevsky, 2012]. Compared to them, the *RusNLP* project :

1. uses much more up-to-date data and includes novel publishing venues (AIST and AINL).
2. Implements the semantic search engine available online as a web service.
3. Publishes the datasets and the source code.

Experiments with similar documents search

Popular approaches for document representations :

1. ***TF-IDF* (term frequency – inverted document frequency)** : a term weighting scheme from information retrieval [Sparck Jones, 1972].
2. ***LDA* (Latent Dirichlet Allocation)** : a widespread distributional topic modeling technique [Blei et al., 2003].
3. ***Paragraph Vector* (also known as *doc2vec*)** : a distributional parametric algorithm based on shallow feed-forward neural networks, extension of *word2vec* [Le and Mikolov, 2014].

Experiments with similar documents search

Popular approaches for document representations :

1. ***TF-IDF* (term frequency – inverted document frequency)** : a term weighting scheme from information retrieval [Sparck Jones, 1972].
2. ***LDA* (Latent Dirichlet Allocation)** : a widespread distributional topic modeling technique [Blei et al., 2003].
3. ***Paragraph Vector* (also known as *doc2vec*)** : a distributional parametric algorithm based on shallow feed-forward neural networks, extension of *word2vec* [Le and Mikolov, 2014].

We tested them all.

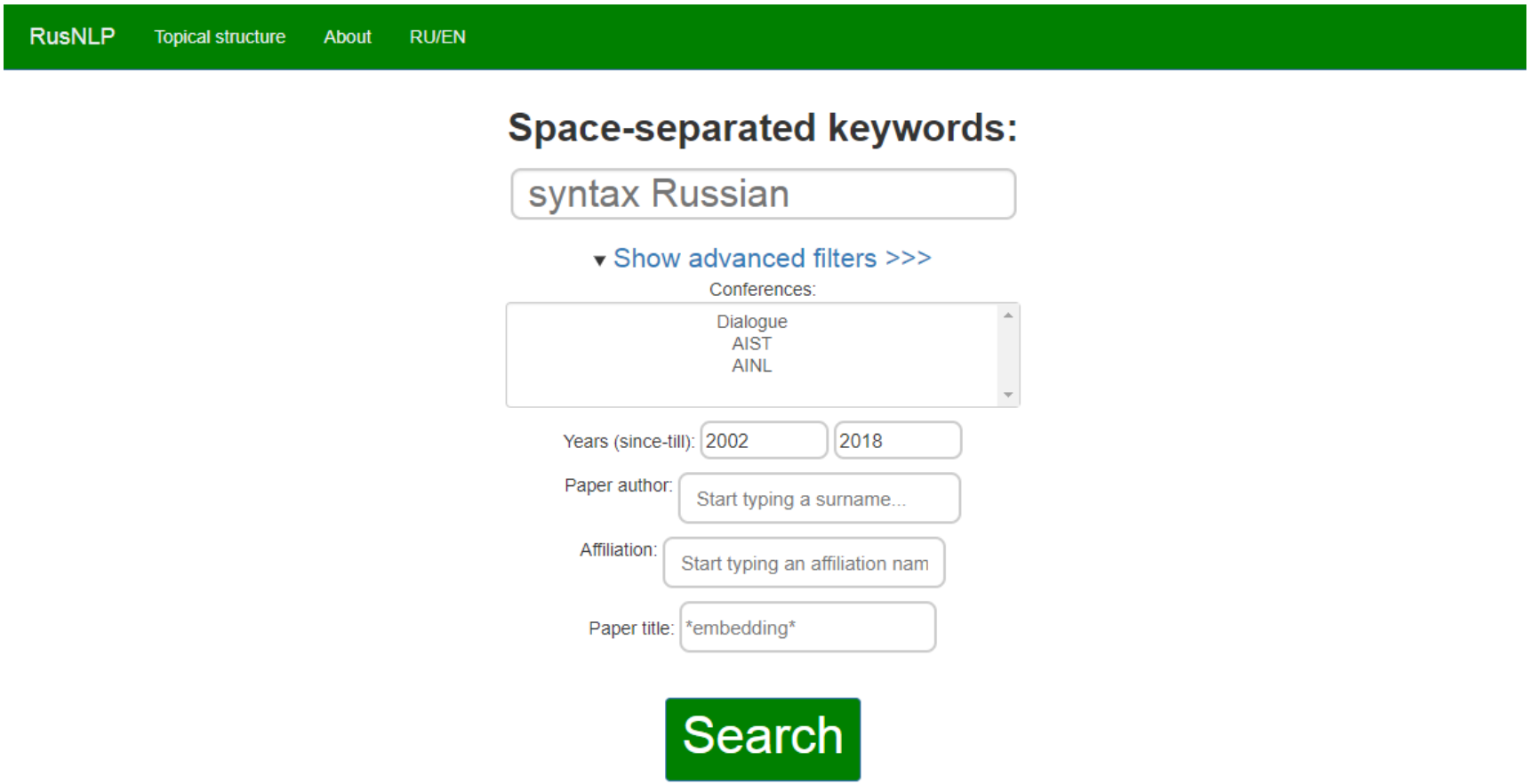
TF-IDF wins

Performance of the tested models (3 independent assessors)		
Model	**Average precision**	**Inter-rater agreement**
TF-IDF	0.64	0.73
Paragraph Vector, 40 dims	0.33	0.66
Paragraph Vector, 100 dims	0.39	0.50
LDA, 10 topics	0.20	0.66
LDA, 20 topics	0.27	0.52

Precision here is simply the ratio of the results which are at least somewhat relevant in the 10 nearest neighbors of a document. No recall evaluation was performed.

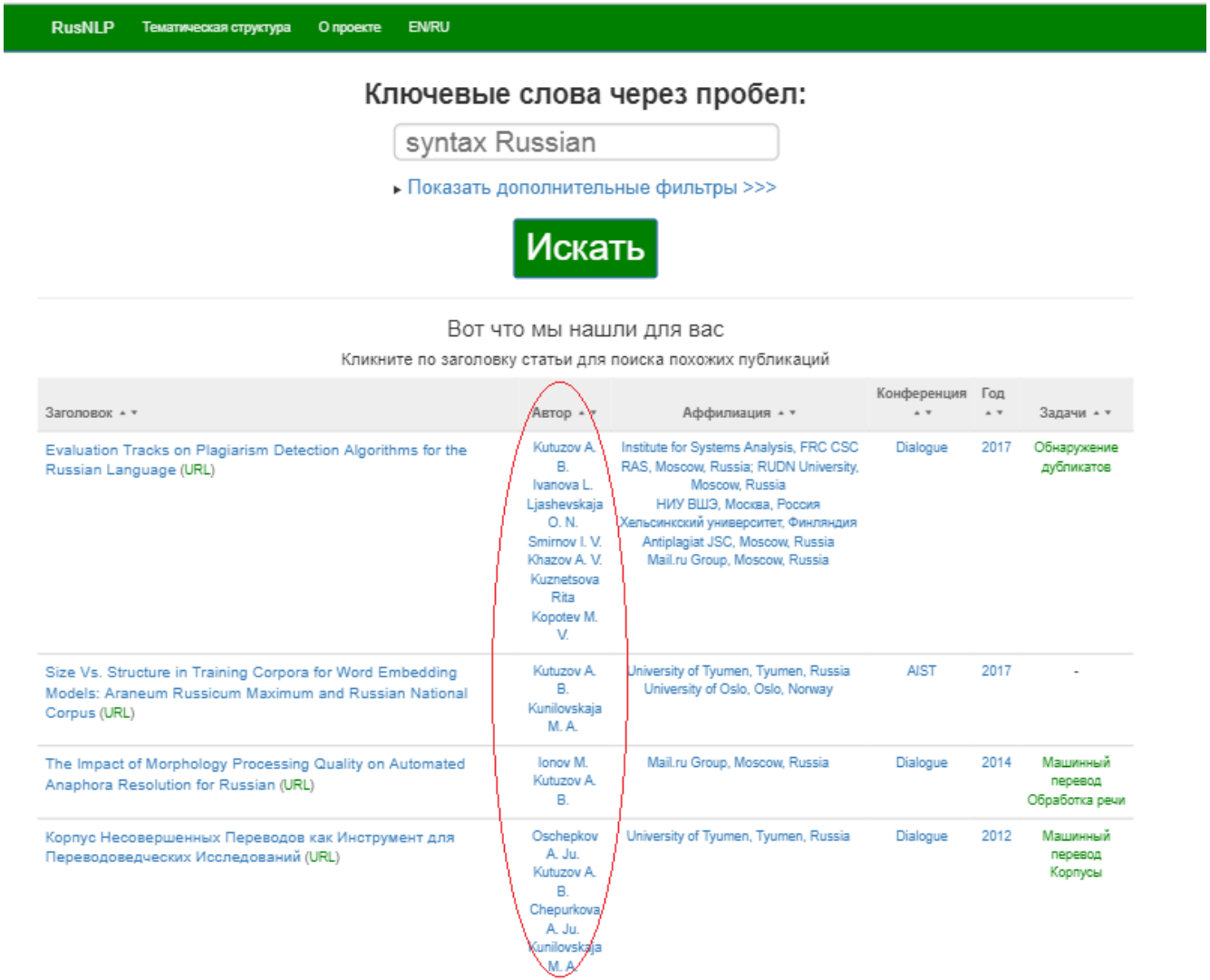
Contents

Key features of RusNLP



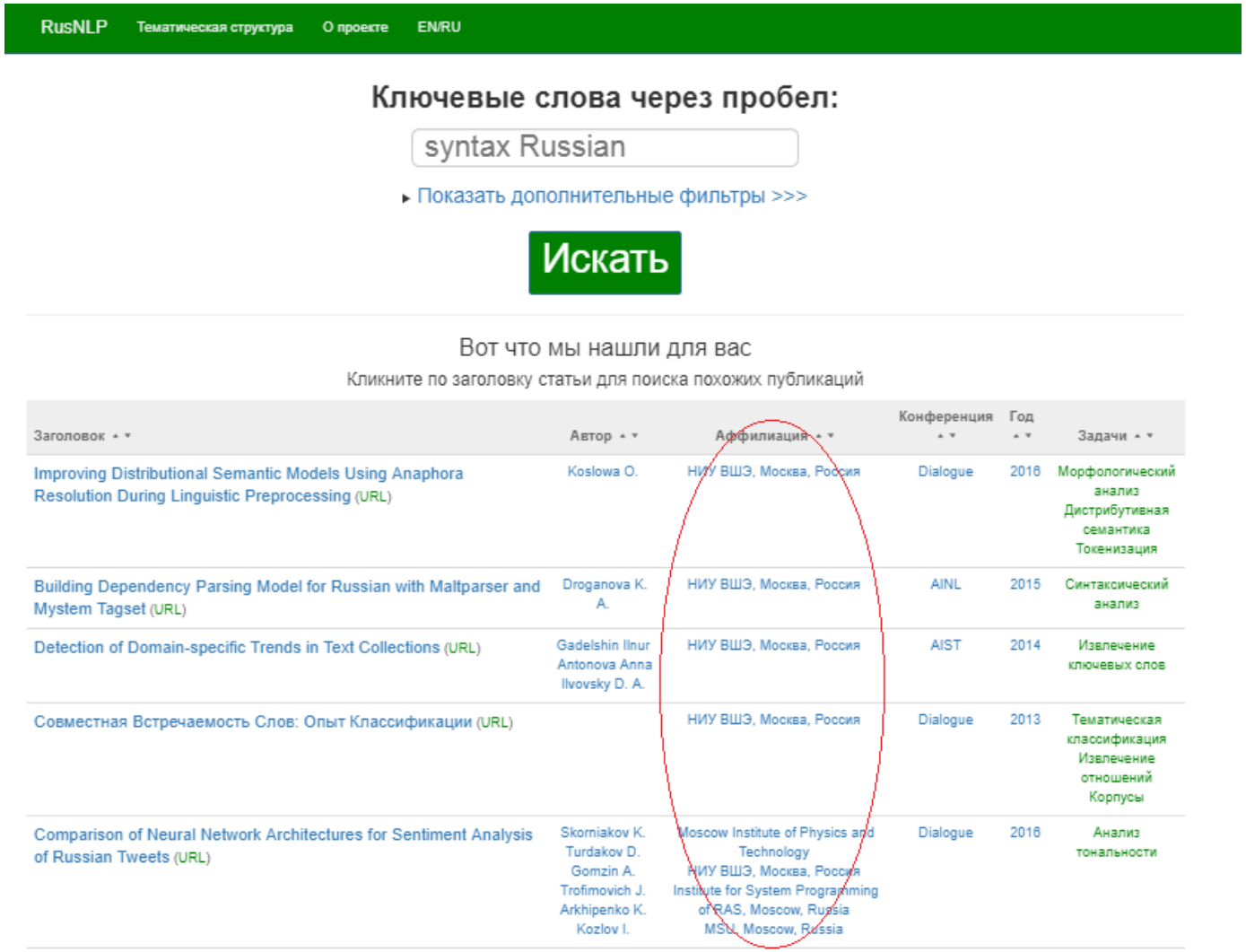
Generic search with filters

Key features of RusNLP



Interlinked authors

RusNLP web service : key features



Interlinked affiliations

Key features of RusNLP

RusNLP | Тематическая структура | О проекте | EN/RU

Entity Based Sentiment Analysis Using Syntax Patterns and Convolutional Neural Network (URL)	Kazorin V. I. Nemov N. R. Kozhevnikov M. V. Karpov I. A.		Dialogue	2016	Анализ тональности	0.031
На Входе Тексты, на Выходе Смысл: Нейронные Языковые Модели для Задач Семантической Близости (На Материале Русского Языка) (URL)	Kutuzov A. B. Andreev I.	Mail.ru Group, Moscow, Russia	Dialogue	2015	Дистрибутивная семантика	0.015
Identifying Disease-related Expressions in Reviews Using Conditional Random Fields (URL)	Miftahutdinov Z. Sh. Tropsha A. E. Tutubalina E. V.		Dialogue	2017	Извлечение именованных сущностей	0.018
Combined Feature Representation for Emotion Classification from Russian Speech (URL)	Verkholyak Oxana Karpov Alexey	ITMO university	AINL	2017	Извлечение именованных сущностей Обработка речи Тематическая классификация Извлечение отношений	0.054

Interlinking with `http://NLPub.ru` (typical NLP tasks)

Key features of RusNLP

«Morphological Analysis for Russian: Integration and Comparison of Taggers»

Kuzmenko E. A.; НИУ ВШЭ, Москва, Россия;

AIST, 2016;

Similar publications:

Number of results: 10 Search

Paper title	Paper author	Affiliation	Conference	Year	Tasks	Similarity
A Close Look at Russian Morphological Parsers: which One is the Best? (URL)	Fishcheva Irina Kotelnikov E. V. Razova E. V.		AINL	2017	Морфологический анализ	0.3375
The Beginning of a Beautiful Friendship: Rule-based and Statistical Analysis of Middle Russian (URL)	Gavrilova T. Berdichevskij A.		Dialogue	2016	Морфологический анализ	0.2115
Morphological Analyzer and Generator for Russian and Ukrainian Languages (URL)	Korobov Mikhail		AIST	2015	Морфологический анализ	0.2108
Автоматическое Извлечение Правил для Снятия Морфологической Неоднозначности (URL)	Bocharov V. V. Protopopova	СПбГУ, Санкт-Петербург, Россия	Dialogue	2013	Корпусы	0.2051

Nearest neighbors search

Contents

Conclusion

Done

- Publicly available search engine over publications from Russian computational linguistics venues ;

Conclusion

Done

- Publicly available search engine over publications from Russian computational linguistics venues ;
- search by keywords, authors, affiliations, years, conferences...

Conclusion

Done

- Publicly available search engine over publications from Russian computational linguistics venues ;
- search by keywords, authors, affiliations, years, conferences...
- manually curated metadata dataset available for downloading (*SQLite* database) ;

Conclusion

Done

- Publicly available search engine over publications from Russian computational linguistics venues ;
- search by keywords, authors, affiliations, years, conferences...
- manually curated metadata dataset available for downloading (*SQLite* database) ;
- (simplistic) evaluation of algorithms for similar documents search.

Conclusion

In the future

- we will maintain and yearly update the database ;
- analysis of citation network (coming soon !) ;
- search for papers in Russian (from the **Dialogue** proceedings) ;
- temporal topic models ;
- use multiword expressions (word n-grams) ;
- expand user queries with pre-trained word embedding models.

Conclusion

In the future

- we will maintain and yearly update the database ;
- analysis of citation network (coming soon !) ;
- search for papers in Russian (from the **Dialogue** proceedings) ;
- temporal topic models ;
- use multiword expressions (word n-grams) ;
- expand user queries with pre-trained word embedding models.

Thank you for your attention !
We welcome any questions, comments or suggestions you may have :)

`http://nlp.rusvectores.org`

References I


Bakarov, A., Kutuzov, A., and Nikishina, I. (2018).
Russian computational linguistics : topical structure in 2007-2017 conference papers.
In *Proceedings of Dialogue-2018, online papers*. ABBYY.


Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent Dirichlet allocation.
Journal of Machine Learning Research, 3(Jan) :993–1022.


Khoroshevsky, V. (2012).
Knowledge spaces in the Internet and Semantic Web, part 3 ; in Russian.
Artificial Intelligence and Decision Making, pages 3–38.


Le, Q. and Mikolov, T. (2014).
Distributed representations of sentences and documents.
In *International Conference on Machine Learning*, pages 1188–1196.


Sparck Jones, K. (1972).
A statistical interpretation of term specificity and its application in retrieval.
Journal of documentation, 28(1) :11–21.