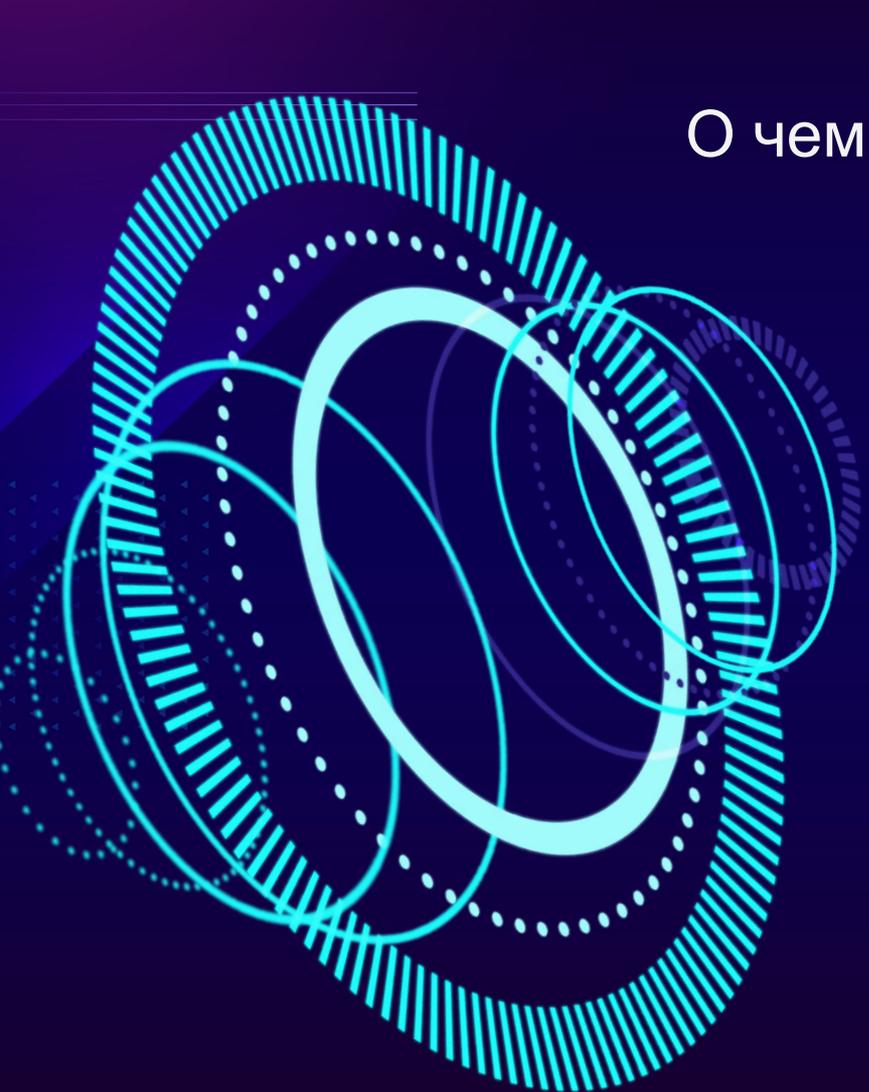


20 лет цифровых
гуманитарных
проектов:
МОТИВАЦИЯ
и риски

Борис Орехов





О чем пойдет речь

01

Таймлайн

Когда и с чего все началось

02

Мотивация

Что заставляет начинать десятки цифровых гуманитарных проектов

03

Риски

Что угрожает проектам в процессе создания и в ходе жизненного цикла

04

Предварительные (?) итоги

Мы в январе 2025



С января 2007

20 лет ровно еще не исполнилось, но мне хотелось бы
поделиться опытом, который уже не укладывается в одно
десятилетие



Фольклорный архив Башкирского государственного университета

Оглавление издания						Фильтр
ID	Название	Жанр	Район	Год записи	Ссылка	
1	Таш йың	сказка в	Салаватский район	1963	д. 1. л. 3—10	
2	Ғалия	сказка в	Салаватский район	1963	д. 1. л. 11—16	
3	Энем һары ат	сказка в	Салаватский район	1963	д. 1. л. 17—20	
4	Кеса-мора	сказка б	Салаватский район	1963	д. 1. л. 21—23	
5	Танәббер хатын	сказка б	Салаватский район	1963	д. 1. л. 27	
6	Өмәт	сказка в	Салаватский район	1963	д. 1. л. 31—33	
7	Дейә һарты	сказка б	Салаватский район	1963	д. 1. л. 24—26	
8	Васыят	сказка б	Салаватский район	1963	д. 1. л. 34—35	
9	Йеңе туган	сказка в	Не указано	неизв.	д. 1. л. 28—30	
10	Хайлакар кеше	сказка в	Оренбургская область, Кувадский район	1964	д. 1. л. 36—38	
11	Өмәт	сказка б	Оренбургская область, Кувадский район	1964	д. 1. л. 39—40	

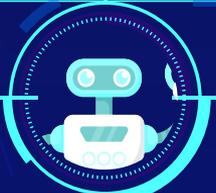
Башкирский фольклор

Мотивация: цифровизация сама по себе

Риски: в публикации остался устаревший URL

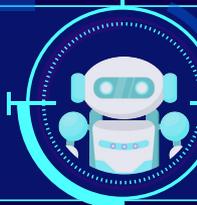
nevmenandr.net/pages/bashfolk.php ←
lcpb.bashedu.ru

Пропавшие ресурсы



91-й том

Разработчики потеряли доступ к серверу и домену `index.tolstoy.ru`



Башкирские вектора

Университет ограничил доступ к своей IT-инфраструктуре `lcpb.bashedu.ru`



Башкирский корпус

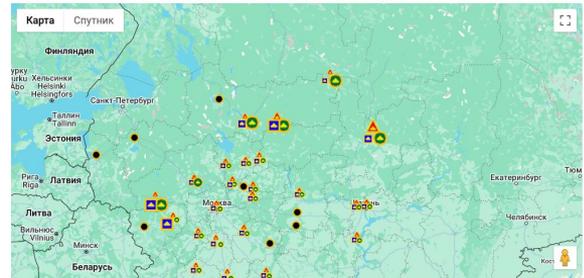
Нарушена обратная совместимость при переходе на новую версию ПО `bascorp.us.ru`

Мотивация: обновление практик работы с данными

Карта

Карта «Недоимки по подушной подати с 1724 по 1740 гг. в России (относительные показатели)»

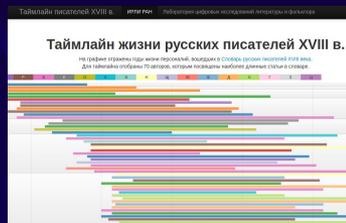
На карте показан средний размер недоимок по провинциям с 1724 по 1740 года в процентном отношении к годовому окладу по трем видам прямого налога (семигривенному, четырехгривенному, сорокоальтному).



Сеть

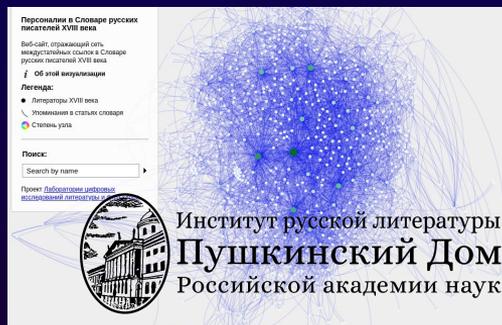
Сетевые отношения внутри словаря русских писателей XVIII века

Риски: безвестность для специалистов



Таймлайн

На основе датасета из Словаря русских писателей XVIII века



Векторные романы



Литература
Необычные и
узнаваемые
варианты
текстов

Векторные романы

Русская литература и дистрибутивная семантика

Да вспомни
Из

Компьютерная лингвистика позволяет вычислять семантическую близость, то есть автоматически находить слова, которые ближе всего другому значению. Это можно делать благодаря так называемым «векторным моделям». Подробно об этом рассказывается на сайте [RusVectores](#).

А что будет, если в хорошо знакомых нам текстах все слова заменить на близкие по смыслу? Насколько изменится такой текст? Что произойдет? Здесь вы найдёте результаты такого эксперимента.

Я взял пять классических русских романов: «Евгений Онегин», «Преступление и наказание», «Война и мир», «Отцы и дети», «Мастер и Маргарита», считая, что романы, содержащие в своем названии «и», всегда играют особую роль в истории русской литературы, так что таких романов четыре из пяти) и автоматически подобрал к ним близкие по значению слова (так называемые квазисинонимы). Для этого я воспользовался [RusVectores](#) (модель построена на текстах НКРЯ и Википедии за ноябрь 2016).

В принципе, такие эксперименты уже были, в частности, можно найти попытку сделать такие же замены в романе «Гордость и предубеждение», этот фокус проверить гораздо проще: там нет ни склонения, ни согласования по роду, да и спряжение весьма редуцированное, можно сказать русском языке если мы просто заменим одну произвольную форму слова на другую произвольную, текст распадется и станет нечитабельным, проводить хитрее, используя морфологический разбор исходного слова и автоматически порождая грамматическую форму для слова-замены. Морфологический анализатор [morphy2](#), который умеет и то, и другое: и устанавливать грамматическую форму слова, и генерировать новую, делает все замены, доступен на [GitHub](#).

Замены подвергаются только самостоятельные части речи (существительные, прилагательные, глаголы и наречия). Имена собственные без таковыми, как в исходном тексте (но не всегда). Если в векторной модели для слова не находится квазисинонимов, то оно не заменяется.

Я сделал три версии замен: в первом случае из векторной модели извлекается ближайшее по значению слово той же части речи, в другом (о существительном, то ищется слово того же рода, что и исходное. Это делает текст более связным грамматически, хотя и лишает нас замены, что вынесен в эпиграф. В третьем случае использовались эффицированные тексты романов (тексты получены с использованием алгоритма эффикации, разработанного в рамках проекта [Карта слов](#)) и глаголы дополнительно фильтруются по залогу (замена для возвратного глагола возвратных).



Технологии

Возможность
совместить
технологии и
культурно
значимые объекты

Персидский поэтический корпус

Риски: Вышка в любой момент может отключить сервер

linghub.ru/persian_poet_corpus

Персидский поэтический корпус
Persian poetic corpus

Поиск в корпусе

Описание
Метаданные
Метры
Грамматика и позиция
Грамматика
Позиция
Авторы и контакты

Описание

Персидский поэтический корпус включает в себя тексты классической персидской поэзии IX-XVII веков в объеме 4,3 млн. словоупотреблений (16 842 произведения, 330 723 бейта). Тексты морфологически размечены, доступен поиск по словам в позиции редифа и рифмы, часть текстов размечена метрически.

Корпус создан на персидском материале как продолжение концепции Поэтического корпуса НКРЯ, Башкирского поэтического корпуса, Корпуса чешского стиха (в порядке появления). Такие корпуса обычно включают морфологическую разметку, доступную в традиционном лингвистическом корпусе, но кроме нее содержат также и специальную разметку, характеризующую уровень стиха.

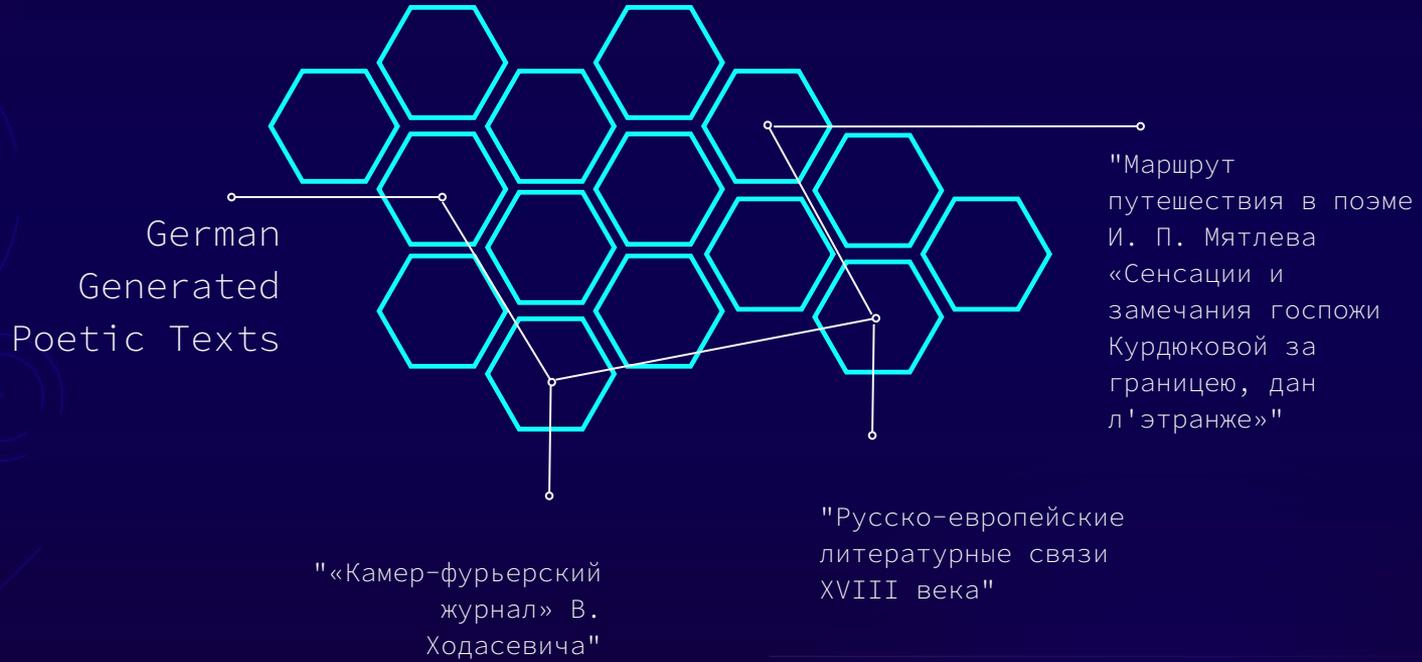
Метаданные

Тексты охарактеризованы следующими метаданными (если они известны):

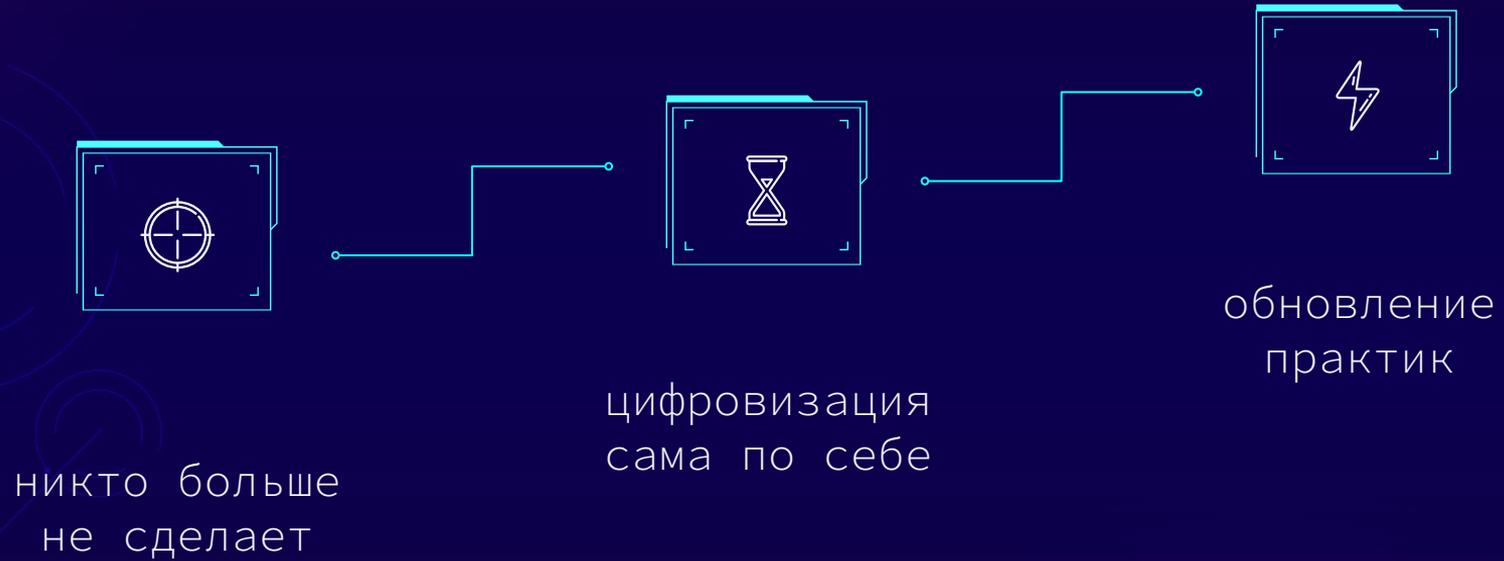
- Автор.
- Латинская транслитерация имени автора.
- Название.
- Жанр.
- Век.
- Метр.

Метры

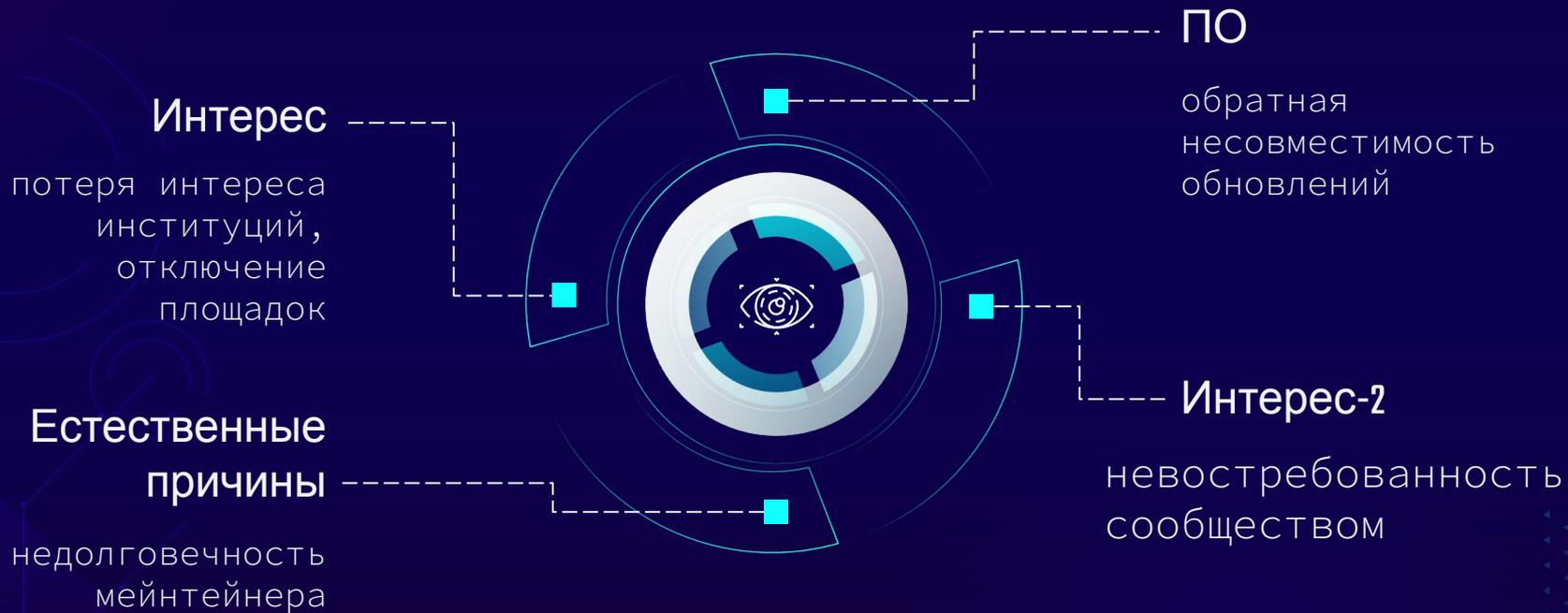
Датасеты



Мотивация



Риски



Спасибо!

Подписывайтесь на
просветительский проект
«Демонтаж красноречия»



youtube.com/@schonenrede



t.me/schonenrede

ССЫЛКИ:

- ◀ nevmenandr.net/slovo
- ◀ nevmenandr.net/pages/bashfolk.php
- ◀ nevmenandr.net/pages/fiscmap.php
- ◀ [nevmenandr.github.io/rus-dict18-persons](https://github.com/nevmenandr/rus-dict18-persons)
- ◀ [nevmenandr.github.io/18cent-timeline](https://github.com/nevmenandr/18cent-timeline)

- ◀ [nevmenandr.github.io/novel2vec](https://github.com/nevmenandr/novel2vec)
- ◀ linghub.ru/persian_poet_corpus