Языки России из Интернета: какие, сколько и где взять?

Л.Зайдельман, И.Крылова, Б.Орехов

НИУ Высшая школа экономики

ru-lang@yandex.ru

25 марта 2016

Содержание

Содержание

Языки России и Интернет

- Прагматическое (сбор текстов для корпусов):
  - вместо дорогой оцифровки дёшево взять готовое;
  - собрать «живой» язык, на котором общаются в сети;
- Исследовательское:
  - узнать, насколько представлены в Сети разные языки;
  - сравнить полученное с другими параметрами и вычислить закономерности;

Как мы это делали?

- Ищем специальные слова (слова-маркеры) для каждого языка, а потом вбиваем их в поиск Яндекса (Яндекс.XML);
- Полученную выдачу просеиваем и получаем сайты и сообщества для выкачки;
- Выкачиваем;
- Складируем;
- Считаем.

Что сделано?

- Собраны текстовые коллекции для большинства языков, которые есть шанс обнаружить в Сети;
- Сделаны наборы поисковых слов-маркеров;
- Отстроена технологическая цепочка поиска и просева сайтов на языках России;
- Создан веб-сервис для скачивания текстовых коллекций;
- Кое-что посчитано.

Сотрудничество

☺ **ФКН** Яндекс-факультет делает для нас специальный сервис выгрузки кастомизированных текстовых коллекций

☺ **Чукотский проект** Даша Игнатенко нашла маркеры чукотского языка, а мы передали чукотским людям всё, что по маркерам нашлось в Интернете

☹ **Лаборатория языков Кавказа**

Содержание

Сколько языков в России?

Составлен список релевантных языков

- Мы считаем только те языки, которые не являются титульными в других государствах (не казахский, не абхазский);
- Но не считаем и идиш: на нём пишут в Интернете почти исключительно за границей;
- Про некоторые не вполне понятно, жив ли ещё хоть один носитель (алеутский);
- а также язык это или диалект;
- После консультаций с Коряковым мы остановились на цифре 96.

Лексические маркеры

- Ищутся слова, которые
  - уникальны для данного языка
  - частотны в данном языке
  - желательно не содержат символов, отличных от стандартной кириллицы

Бытовая система письма (первоначально Зализняк в ДНД):
татшӧм $\longrightarrow$ татшэм, татшом (коми-зырянский)

Грязный интернет, шумные маркеры

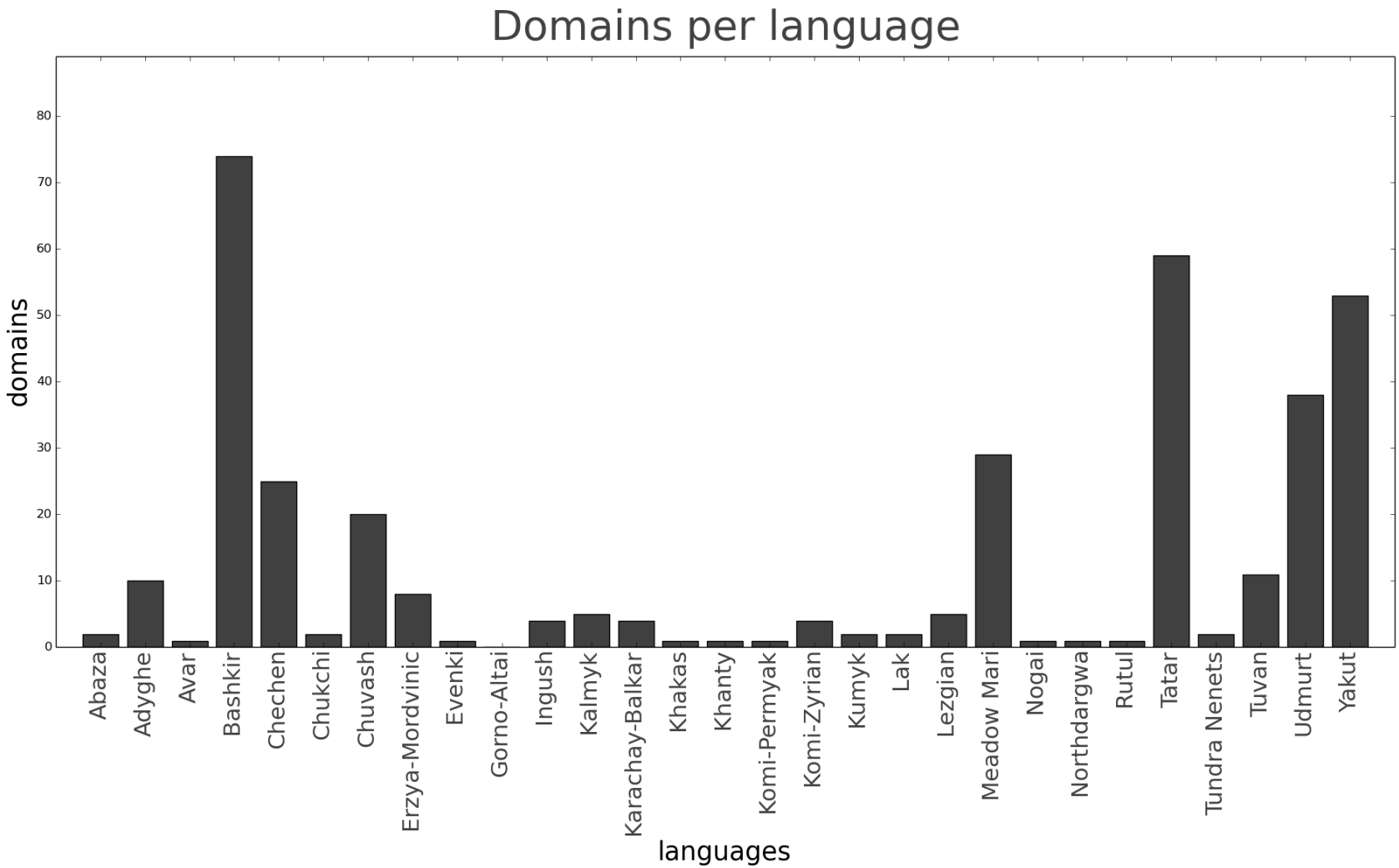
- эмиэ (якутский) & ытла (чувашский) $\longrightarrow$ www.nbchr.ru;
  - бывают сайты регионов: forum.kukmara-rayon.ru
  - бывают сайты языковых групп: finugor.ru
- зеркала и дубли;
- рефераты и научные статьи;
- музыкальные сайты (только названия песен).

Определение языка

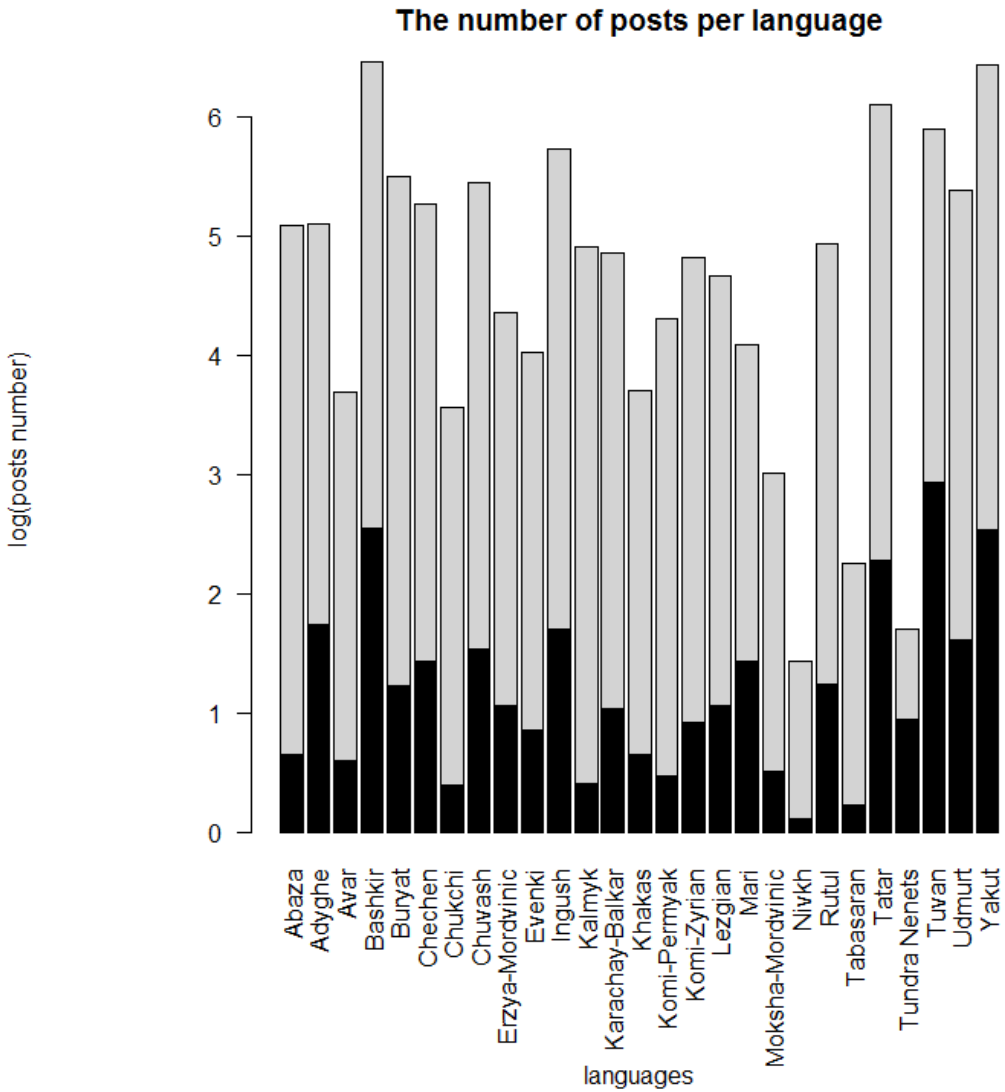
- На найденных страницах может быть текст на разных языках;
- Нас интересует всё, что отлично от русского;
- Обычная определялка должна установить язык;
- Наша определялка устанавливает, текст на русском языке или нет.

Содержание

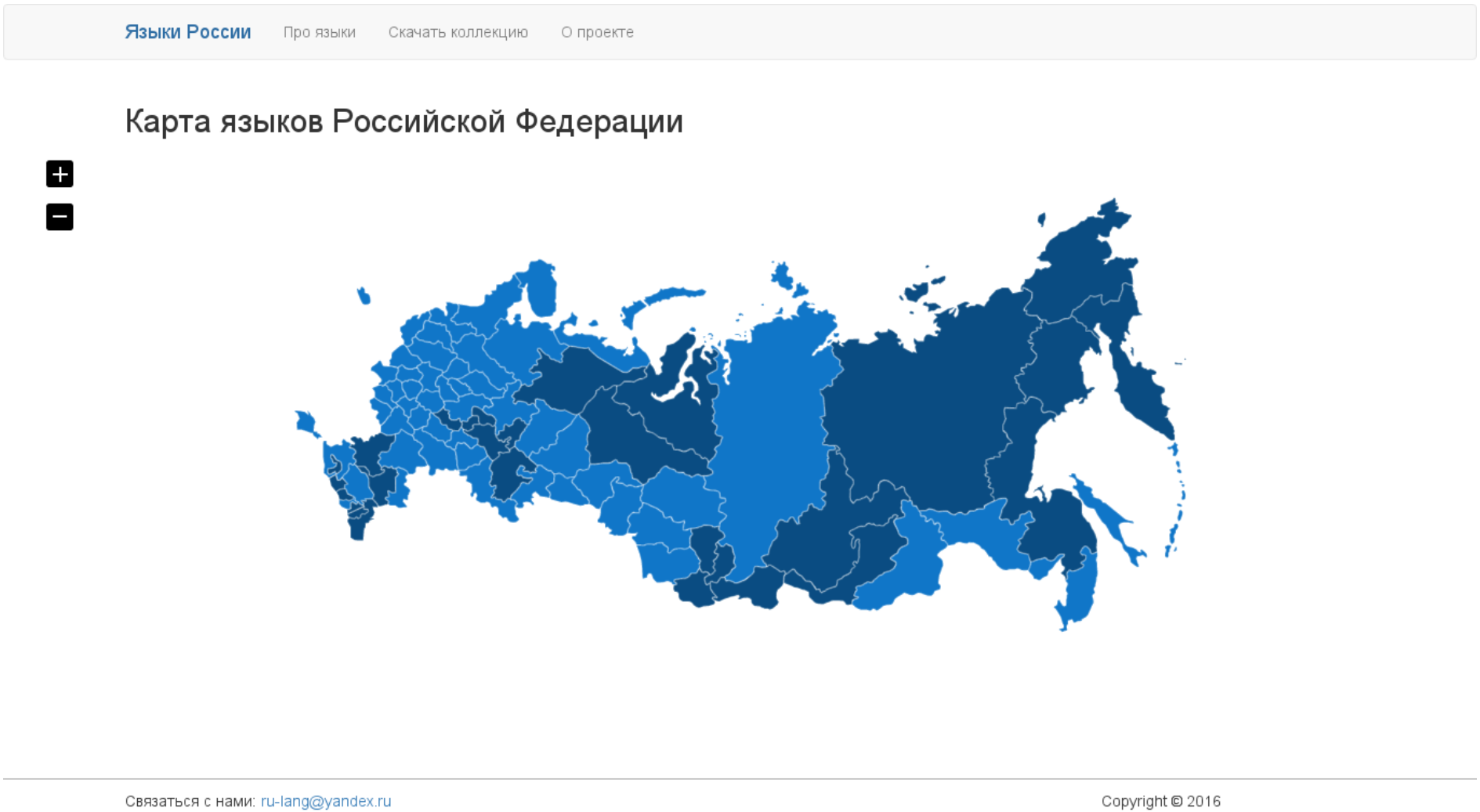
Сколько сайтов нашлось?



Записи в социальных сетях

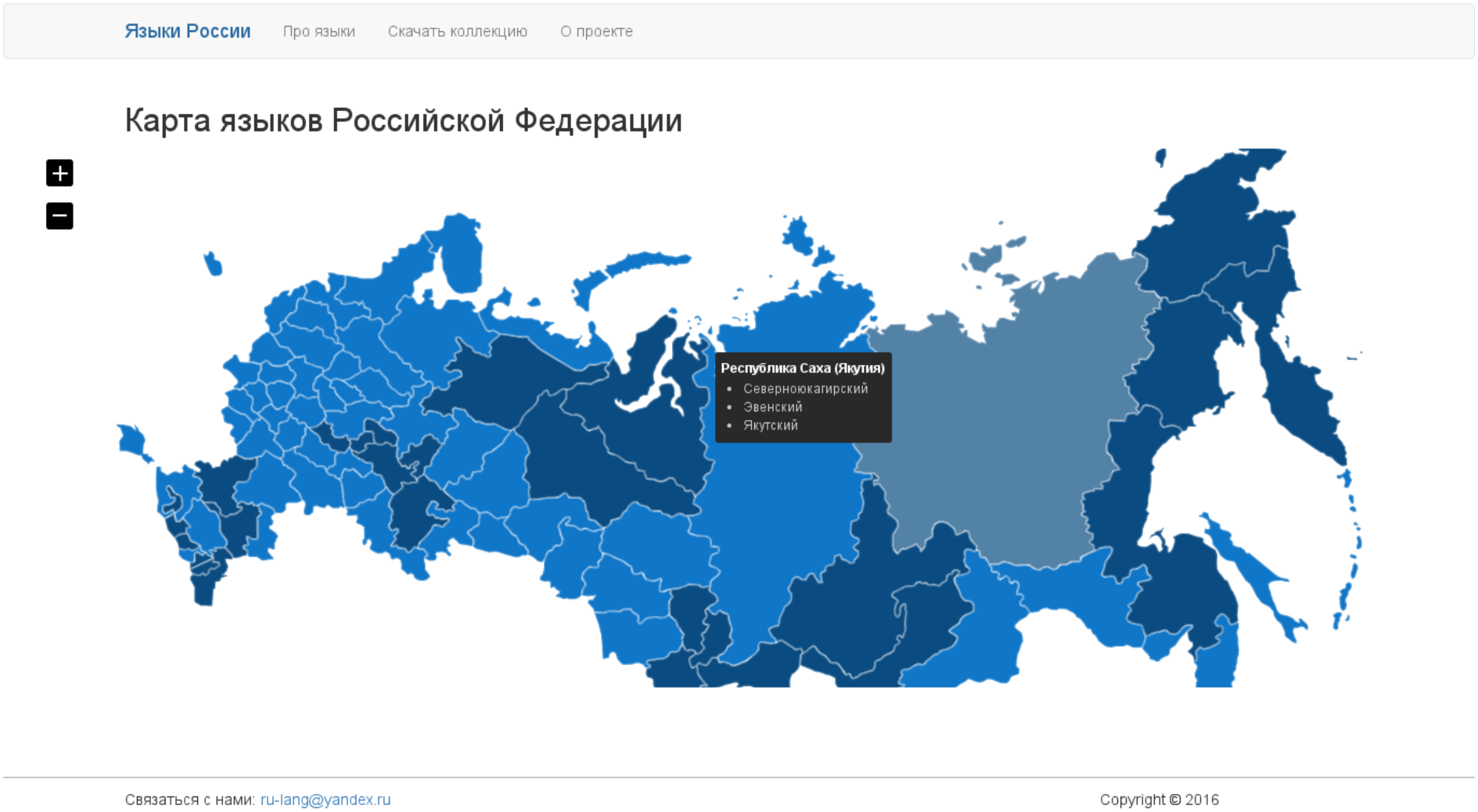


Веб-сервис

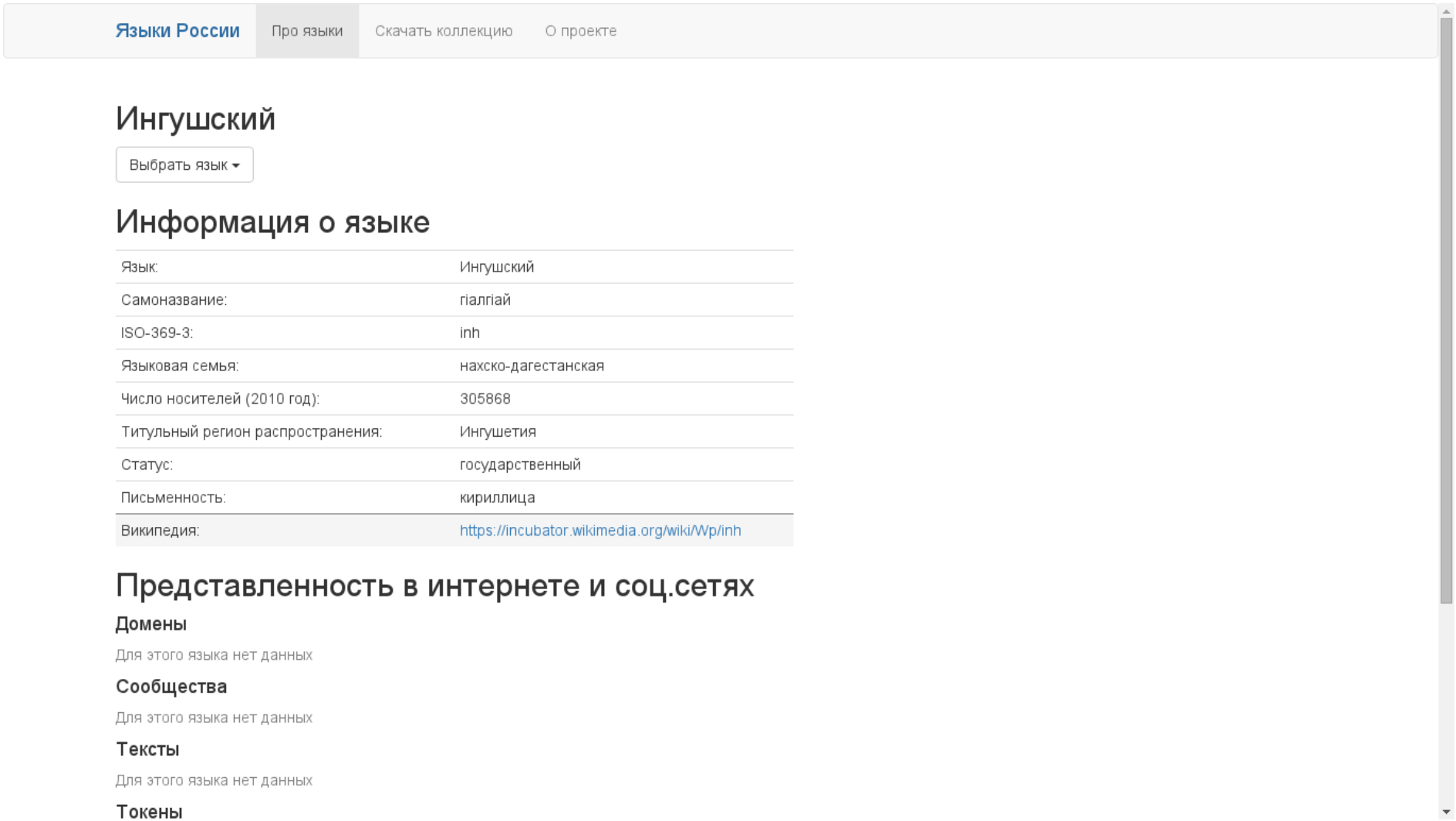


Временный адрес: http://web-corpora.net/wsgi3/irinfox/

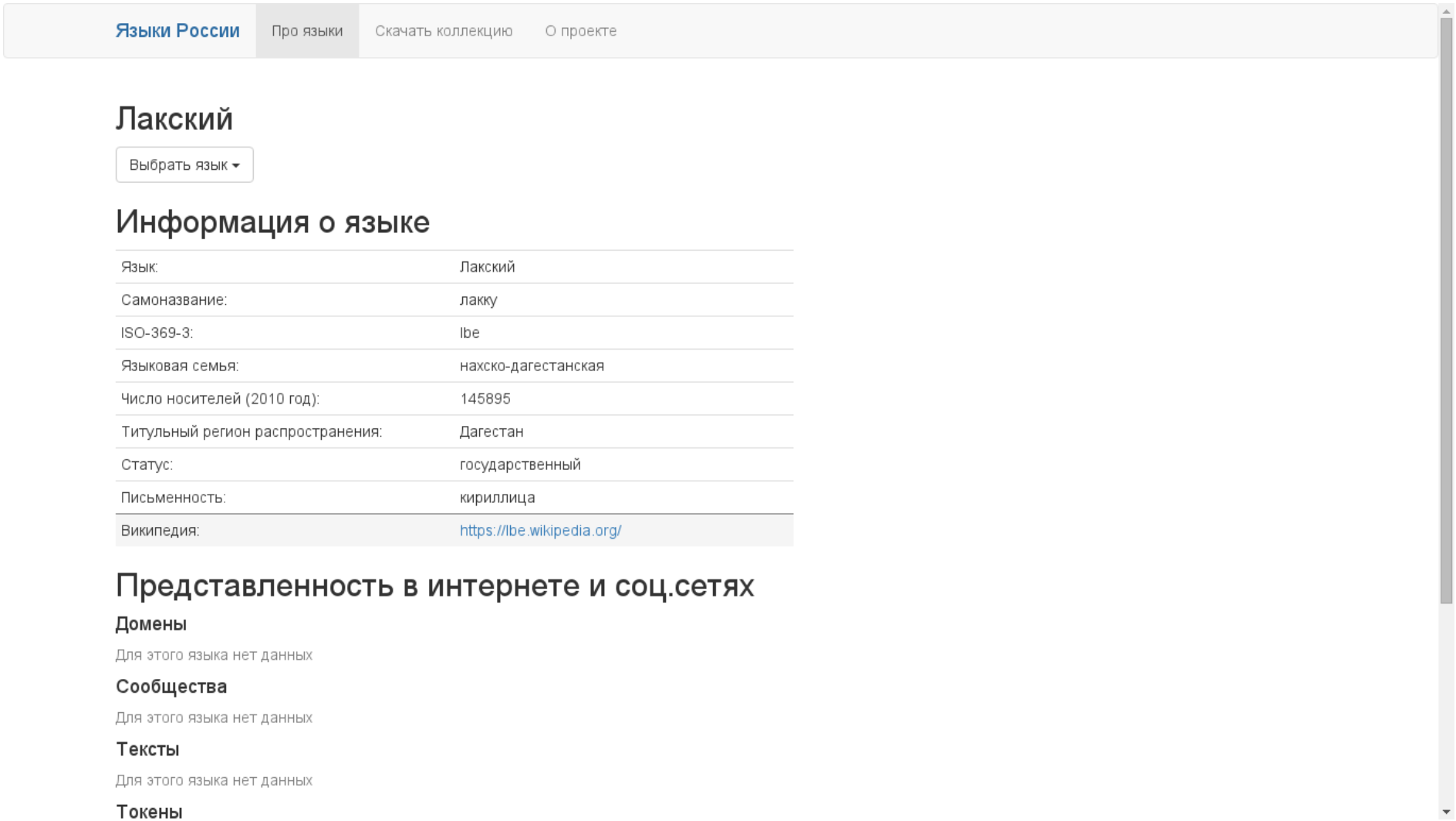
Интерактивная карта



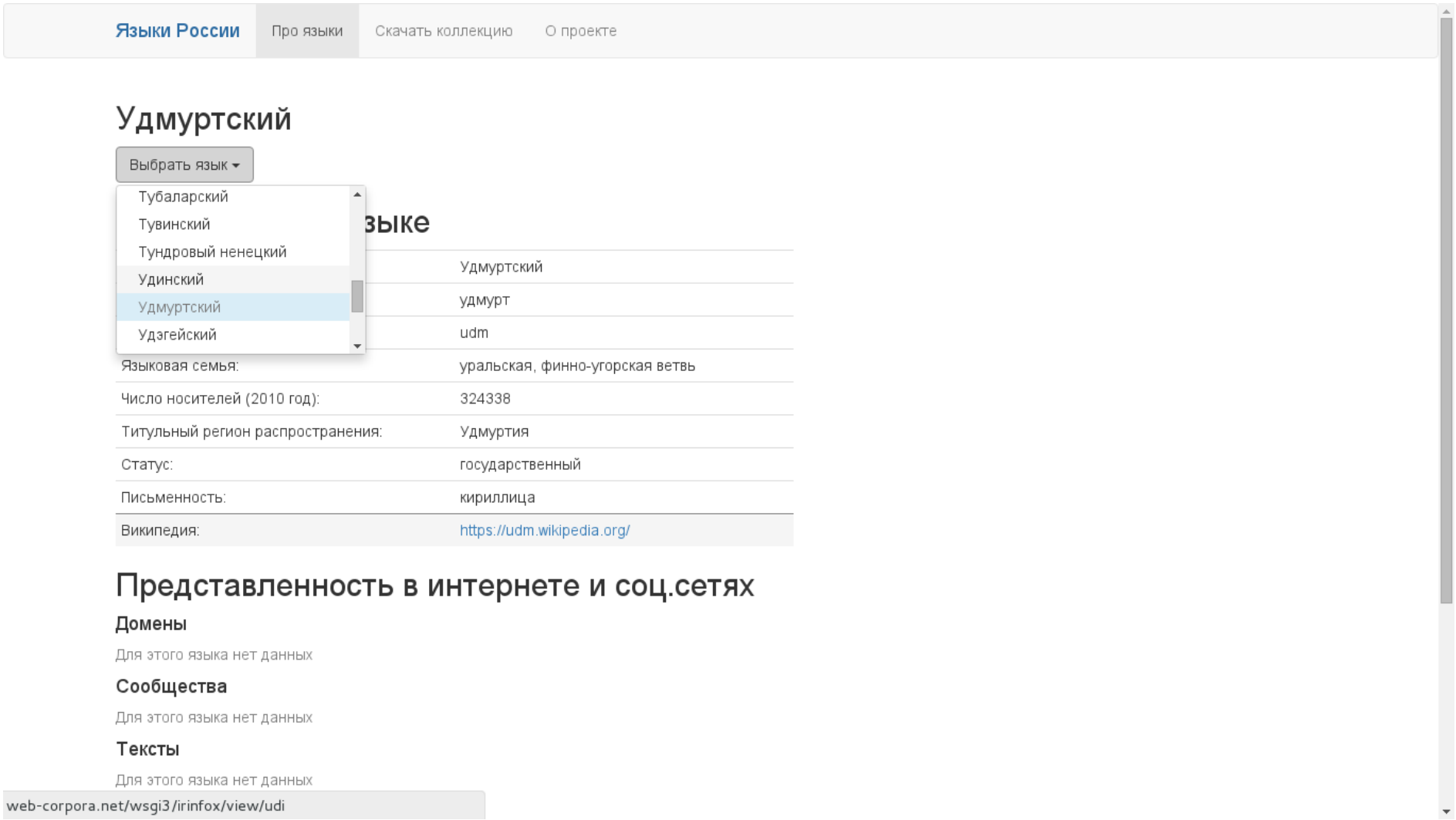
Каталог языков



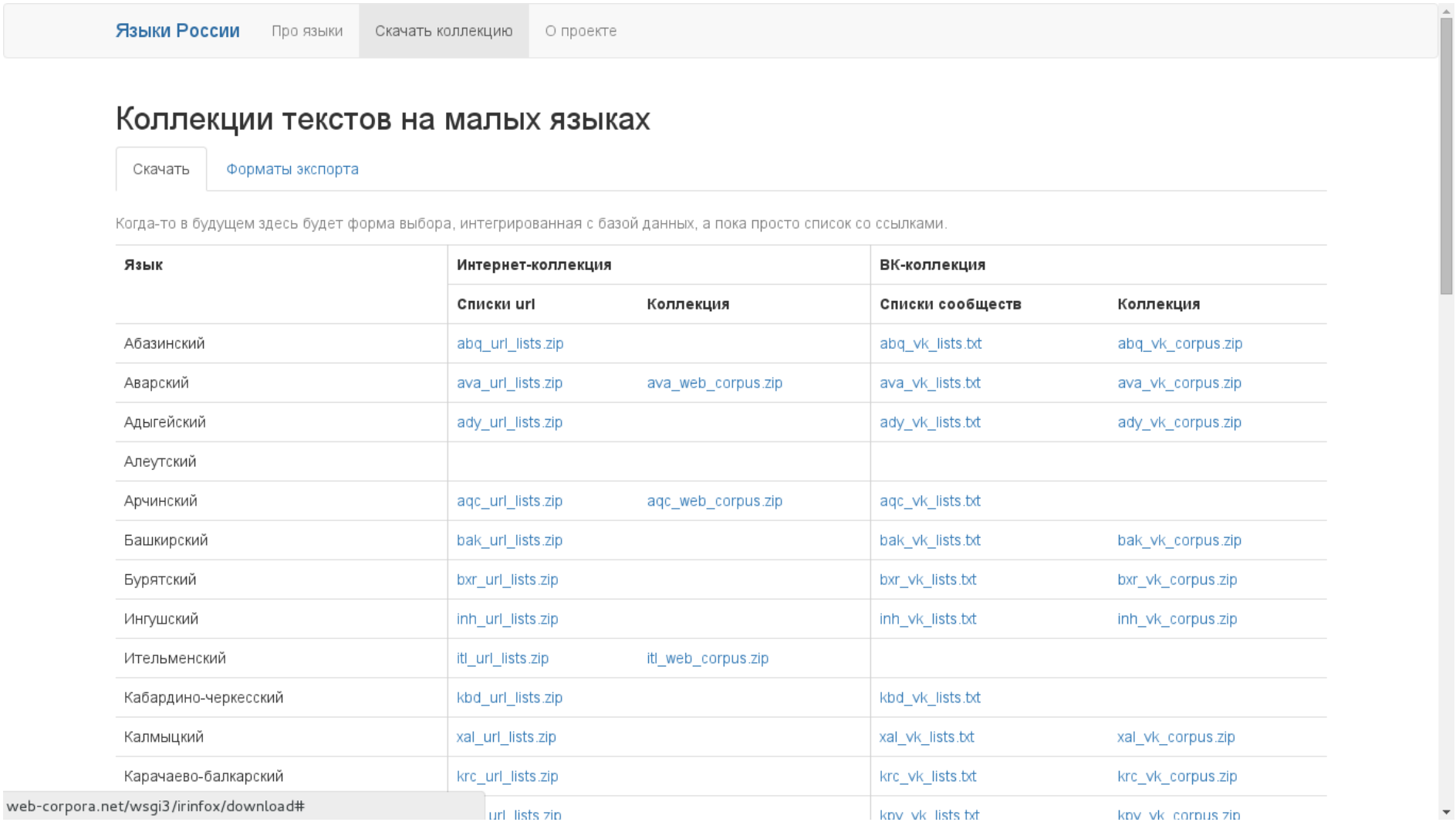
Каталог языков



Каталог языков



Возможность скачать коллекцию



Языки России | Про языки | Скачать коллекцию | О проекте

Коллекции текстов на малых языках

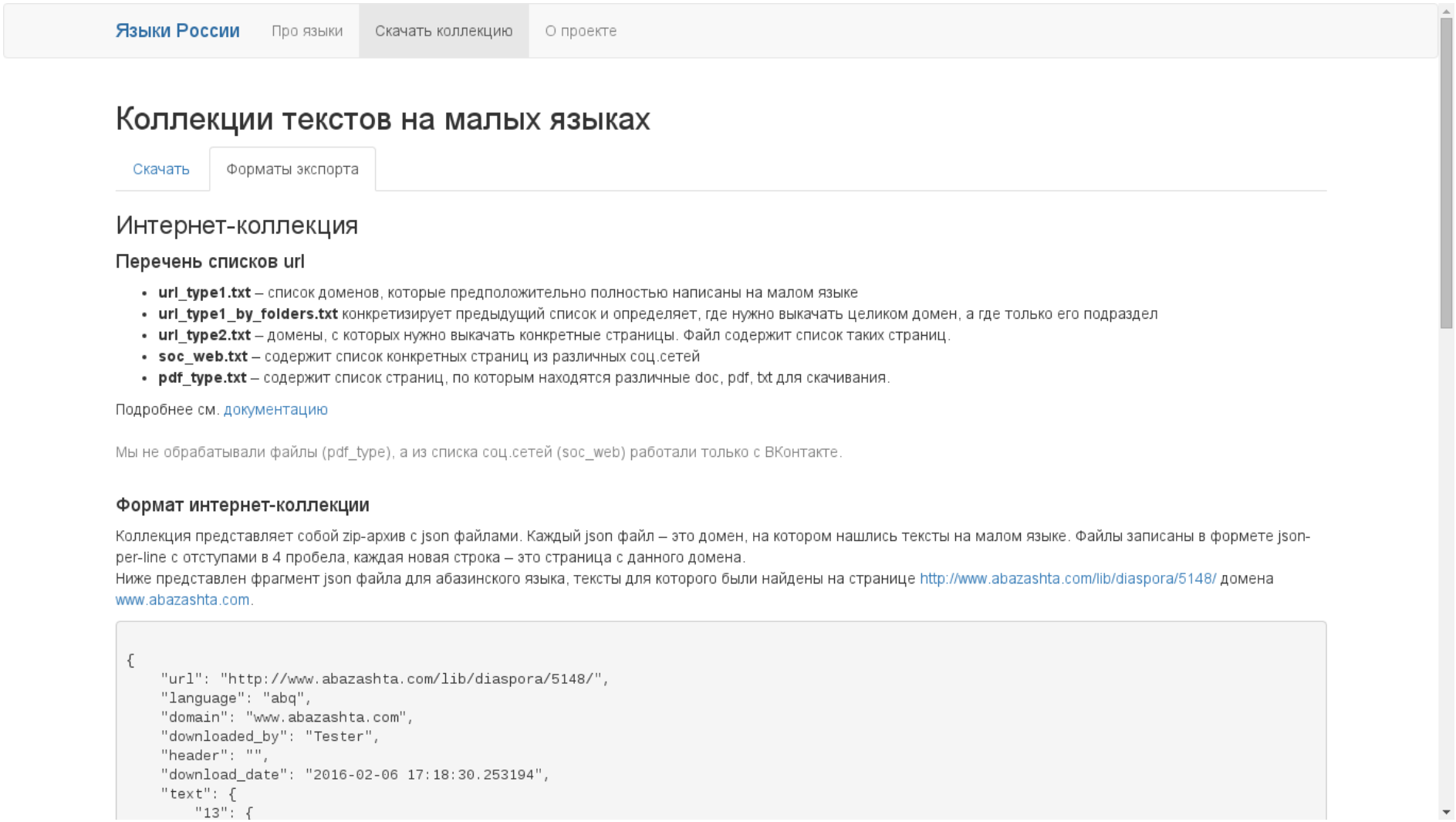
Скачать | Форматы экспорта

Когда-то в будущем здесь будет форма выбора, интегрированная с базой данных, а пока просто список со ссылками.

Язык	Интернет-коллекция		ВК-коллекция	
	Списки url	Коллекция	Списки сообществ	Коллекция
Абазинский	abq_url_lists.zip		abq_vk_lists.txt	abq_vk_corpus.zip
Аварский	ava_url_lists.zip	ava_web_corpus.zip	ava_vk_lists.txt	ava_vk_corpus.zip
Адыгейский	ady_url_lists.zip		ady_vk_lists.txt	ady_vk_corpus.zip
Алеутский				
Арчинский	aqc_url_lists.zip	aqc_web_corpus.zip	aqc_vk_lists.txt	
Башкирский	bak_url_lists.zip		bak_vk_lists.txt	bak_vk_corpus.zip
Бурятский	bxr_url_lists.zip		bxr_vk_lists.txt	bxr_vk_corpus.zip
Ингушский	inh_url_lists.zip		inh_vk_lists.txt	inh_vk_corpus.zip
Ительменский	itl_url_lists.zip	itl_web_corpus.zip		
Кабардино-черкесский	kbd_url_lists.zip		kbd_vk_lists.txt	
Калмыцкий	xal_url_lists.zip		xal_vk_lists.txt	xal_vk_corpus.zip
Карачаево-балкарский	krc_url_lists.zip		krc_vk_lists.txt	krc_vk_corpus.zip

web-corpora.net/wsgi3/irinfox/download#

Описание формата данных



Языки России | Про языки | Скачать коллекцию | О проекте

Коллекции текстов на малых языках

Скачать | Форматы экспорта

Интернет-коллекция

Перечень списков url

- **url_type1.txt** – список доменов, которые предположительно полностью написаны на малом языке
- **url_type1_by_folders.txt** конкретизирует предыдущий список и определяет, где нужно выкачать целиком домен, а где только его подраздел
- **url_type2.txt** – домены, с которых нужно выкачать конкретные страницы. Файл содержит список таких страниц.
- **soc_web.txt** – содержит список конкретных страниц из различных соц.сетей
- **pdf_type.txt** – содержит список страниц, по которым находятся различные doc, pdf, txt для скачивания.

Подробнее см. документацию

Мы не обрабатывали файлы (pdf_type), а из списка соц.сетей (soc_web) работали только с ВКонтакте.

Формат интернет-коллекции

Коллекция представляет собой zip-архив с json файлами. Каждый json файл – это домен, на котором нашлись тексты на малом языке. Файлы записаны в формете json-per-line с отступами в 4 пробела, каждая новая строка – это страница с данного домена.
Ниже представлен фрагмент json файла для абазинского языка, тексты для которого были найдены на странице http://www.abazashta.com/lib/diaspora/5148/ домена www.abazashta.com.

```
{
    "url": "http://www.abazashta.com/lib/diaspora/5148/",
    "language": "abq",
    "domain": "www.abazashta.com",
    "downloaded_by": "Tester",
    "header": "",
    "download_date": "2016-02-06 17:18:30.253194",
    "text": {
        "13": {
```

Содержание

Цифры

- Из 96 языков мы искали тексты на 49 самых распространённых;
- что-то нашлось для примерно 30;
- 383 доменов;
- 1735 сообществ VK.com;
- 239.4 дней ушло на поиск.

Сколько доменов у каждого языка?

Таблица: Языки и домены

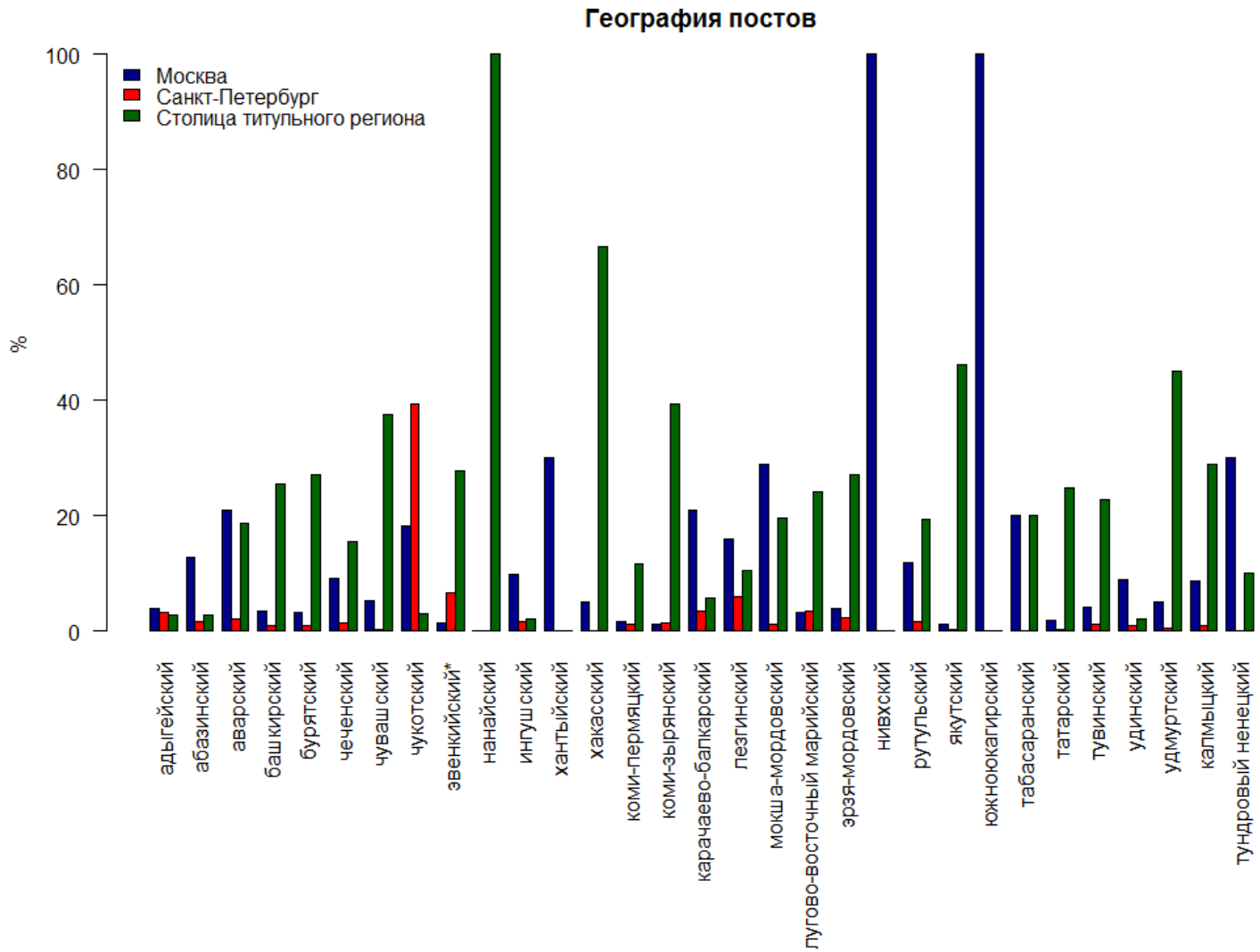
Язык	Носителей	Доменов
Башкирский	1 150 000	74
Якутский	450 140	53
Татарский	4 280 000	59
Чувашский	1 152 404	20
Адыгейский	117 489	10
Калмыцкий	80 546	5
Мокшанский	2 025	0
Ительменский	7	0

Зависимость от числа носителей?

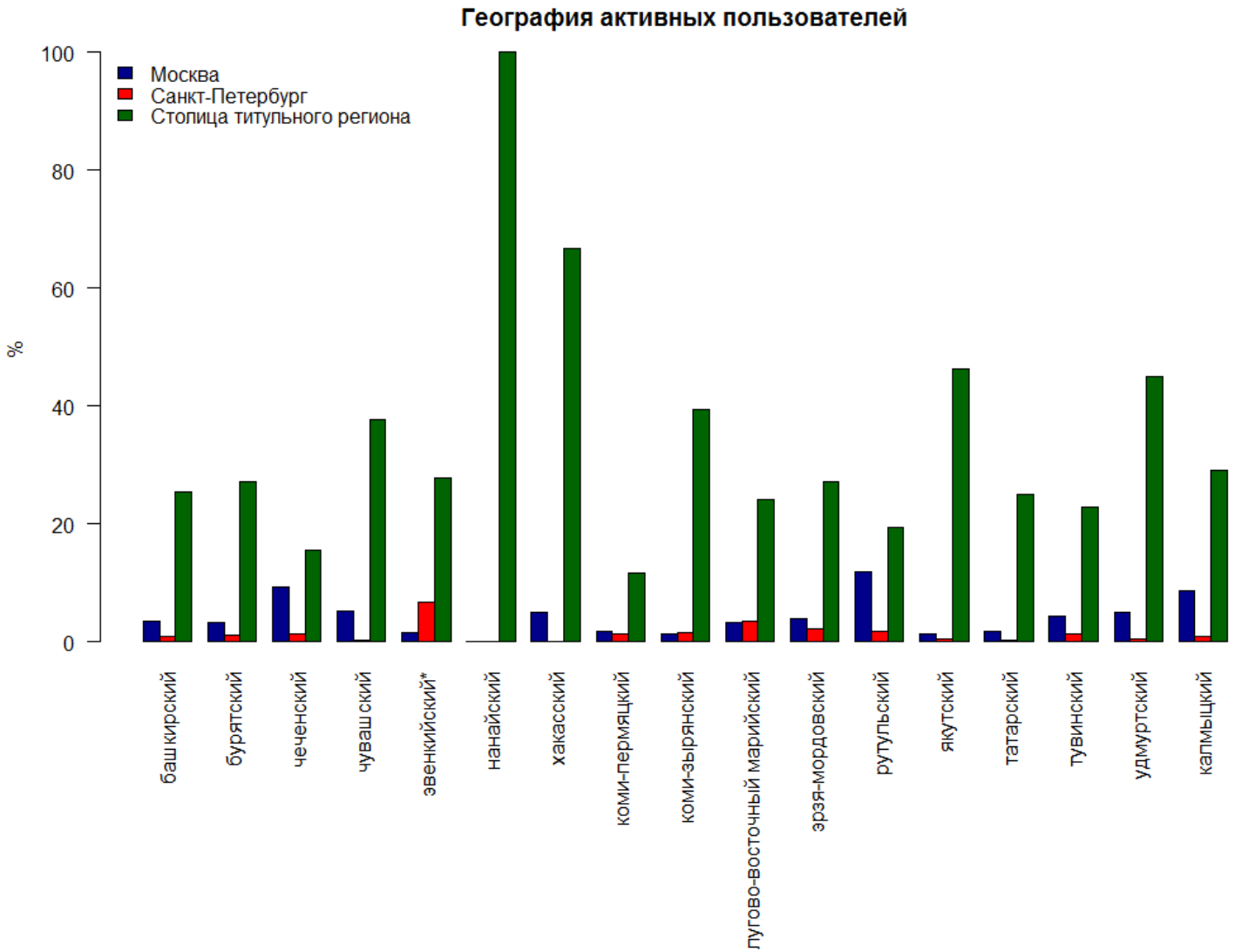
Коэффициент корреляции между числом доменов и числом носителей: 0.739.

География

Так как мы сохраняем информацию профиля, можно посчитать, кто говорит на языке



Сколько пишут из региона



Сколько пишут из столиц