Языки России в Интернете

Л.Зайдельман, И.Крылова, Б.Орехов

НИУ Высшая школа экономики

nevmenandr@gmail.com

10 марта 2016

Содержание

Содержание

История проекта

- Проект начался благодаря Центру изучения Интрнета и общества РЭШ;
- На ранних этапах над ним работали Борис Орехов и Кирилл Решетников;
- Затем работа переехала в НИУ ВШЭ;
- Кирилл больше не занимается этой проблематикой;
- Зато к работе подключились магистранты Школы лингвистики НИУ ВШЭ;
  - Л. Зайдельман;
  - И. Крылова;
  - И. Попов;
  - Е. Степанова.

Содержание

Языки не в фокусе коммерческих разработок

- Большими языками занимаются, потому что это коммерчески выгодно;
- Про большие языки в Интернете известно многое:
  - сколько носителей являются пользователями?
  - сколько (примерно) сайтов есть на этом языке?
  - сколько написано на этом языке в Сети?
- Малые языки коммерчески неинтересны;
- Все их носители двуязычны и дешевле разработать ресурс на большом языке;
- Компании создают сервисы или инструменты NLP на малых языках только из имиджевых соображений, то есть по остаточному принципу;
- Поэтому малые языки являются уделом академического сообщества.

Сколько языков в России?

- Около ста (для точного подсчёта нужно решить проблему язык vs. диалект);
- При этом 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);
- После консультаций со специалистами мы остановились на цифре 96.

Что уже сделано?

- Качественные исследования (например, [Сахарных 2008]), их становится труднее проводить по мере увеличения сегмента Сети;
- Количественные исследования пока единичны ([Орехов, Галлямов 2012], [Орехов, Решетников 2016]);
- Замечено, что новые средства коммуникации помогают носителям малых языков поддерживать коммуникацию.

Сколько языков России живут в Интернете?

- Чтобы выяснить это, нужно их все найти;
- Правда ли, что о распространенности языка можно судить по Википедии?

Как мы искали?

- Мы не можем выкачивать весь Интернет и разбираться с каждой страницей, на каком языке она написана;
- Воспользуемся большими поисковиками;
- Поисковые термины (слова-маркеры) отправляются как запрос поисковику;
- Результаты фильтруются и получаются списки доменов и сообществ в соцсетях;
- Сайты и сообщества выкачиваются и парсятся.

Проблемы

- Нет заранее собранных хороших поисковых терминов (собраны вручную);
- Code-mixing (сделана программа определения языка);
- Бытовая система письма (разработана система правил перевода);
- Автоматические обращения к поисковикам лимитированы (ничего не поделаешь).

Бытовая система письма и code-mixing по-башкирски

Что получилось?

- 96 языков;
- для 49 самых распространённых проведён поиск;
- что-то нашлось для примерно 30;
- 368 доменов;
- 218 дней ушло на поиск.

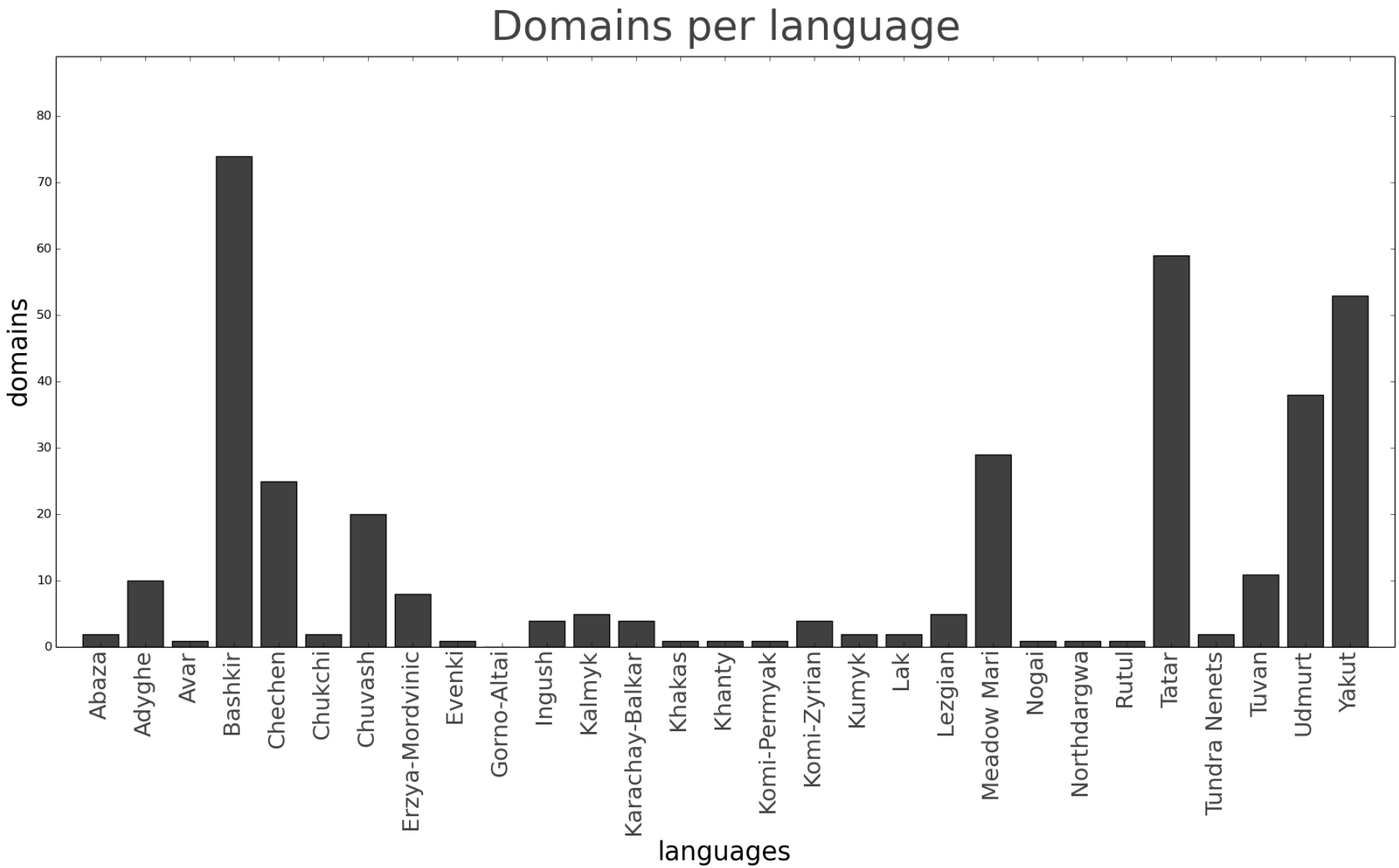
Содержание

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

Сколько доменов у каждого языка?



А число носителей?

Коэффициент корреляции между числом доменов и числом носителей: 0.128.

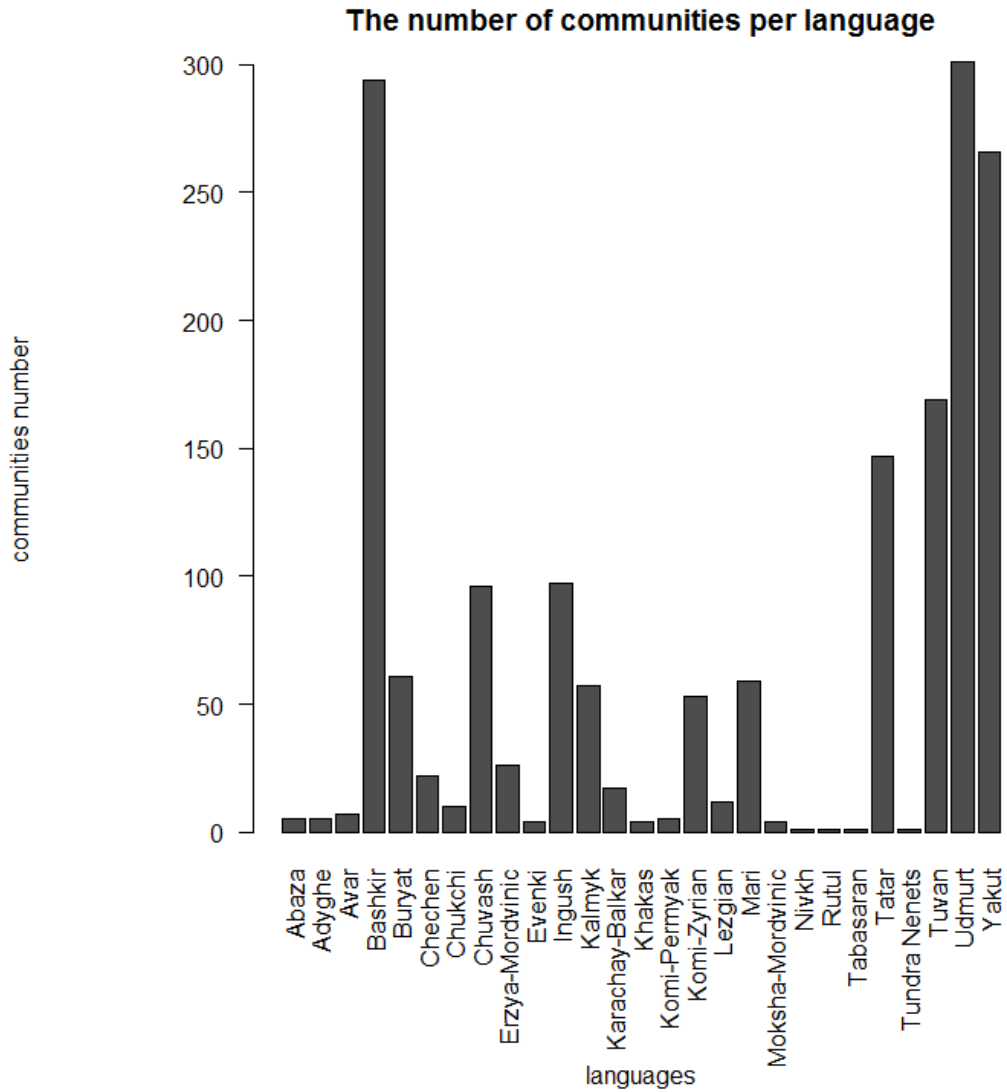
Содержание

VK.com

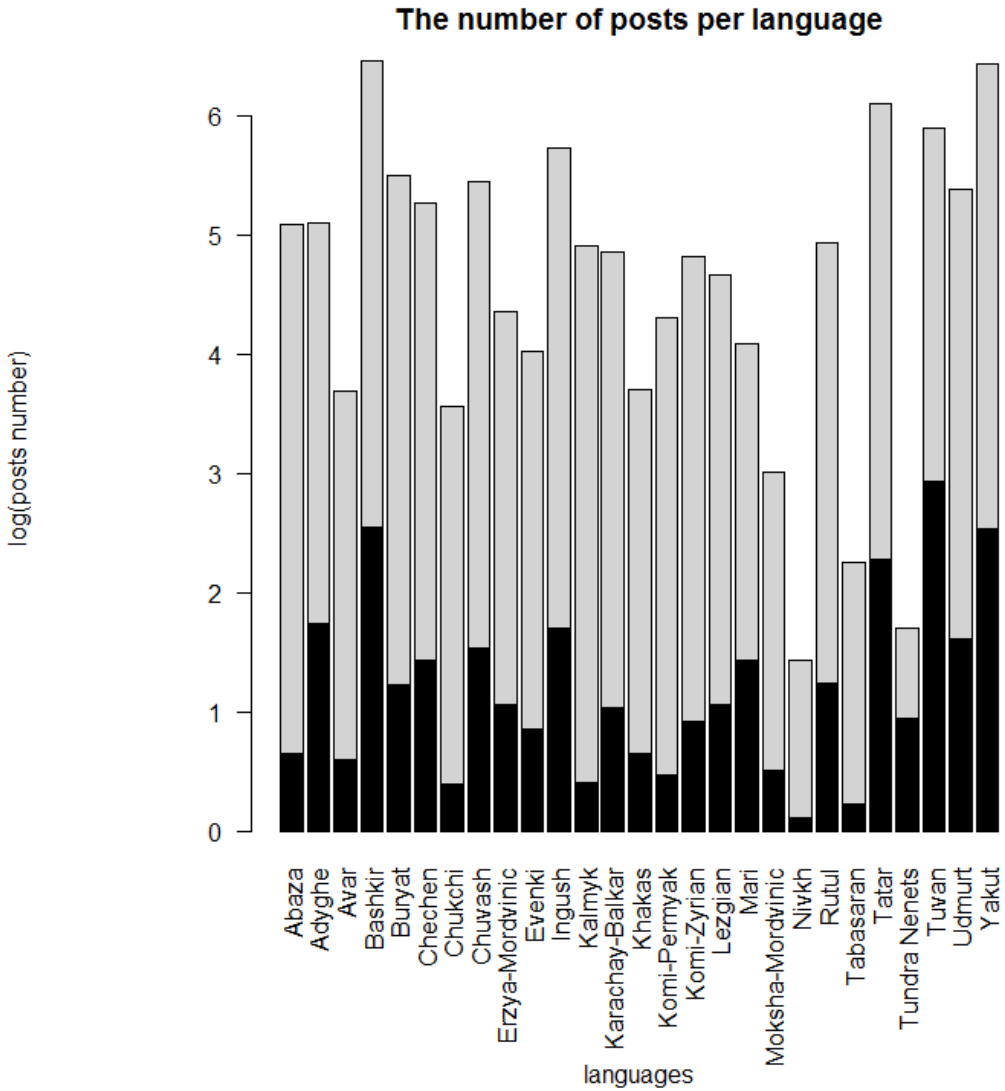
Эксперты свидетельствуют, что в FB наши языки не так распространены

- 27 языков.
- 1735 сообществ.
- Но не все содержат тексты только на миноритарном языке.

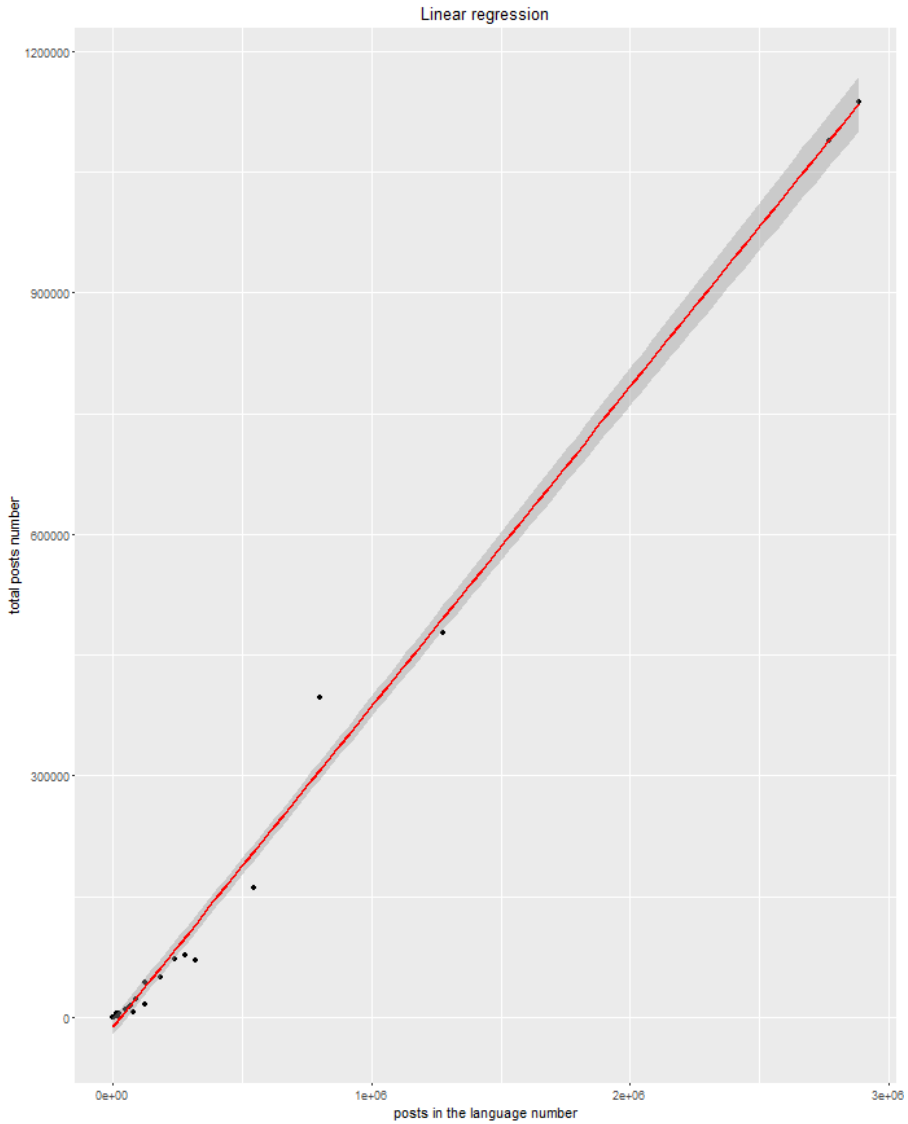
Сообщества и языки



Записи и языки



Зависимость постов на языке от числа постов

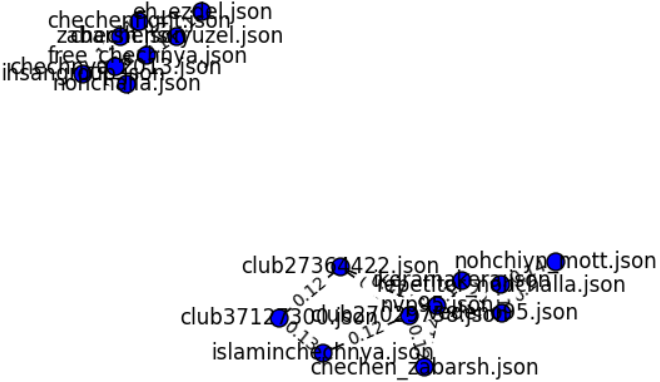


Содержание

Можем ли мы узнать тематику записей?

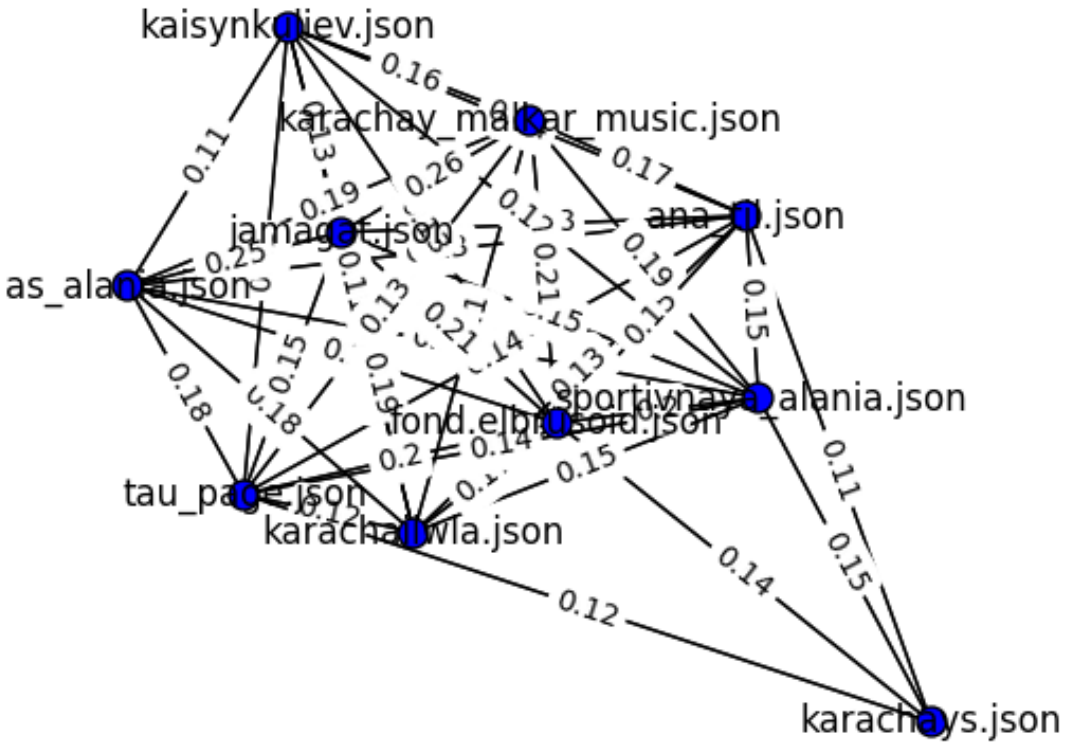
- В идеальном случае нам нужны были бы 30 экспертов, чтобы всё это прочесть.
- Но у компьютерной лингвистики есть свои способы определения тематики.

Чеченский


club27364422.json
free_chechnya.json
chechen_zabarsh.json

Чеченский юмор

Карачаево-балкарский



Карельский


omamua.json
0.11
club4969726.json
0.14
0.15
club4619459.json
0.15
0.11
club3127978.json
0.14
me_karjalaiset.json

Итоги

- Миноритарные языки в Интернете есть.
- Языку нужно как минимум 40 тыс. носителей, чтобы пробиться в Сеть.
- Число носителей в наших данных коррелирует только с числом статей в Википедии.

Итоги

Таблица : Корреляции

	Вики (2015)	Сообществ	Постов	Доменов	Носителей
Вики (2015)	1	0.218	0.284	0.317	0.746
Сообществ	0.218	1	0.790	0.284	0.324
Постов	0.284	0.790	1	0.447	0.368
Доменов	0.317	0.284	0.447	1	0.128
Носителей	0.746	0.324	0.368	0.128	1

Спасибо!