Языки России в интернете

Коллекции текстов

Л. Зайдельман И. Крылова,
а также И. Попов, Е. Степанова, Б. Орехов

НИУ Высшая школа экономики

ru-lang@yandex.ru

Летняя школа, 2016

Содержание

Содержание

История проекта

- Проект начался благодаря Центру изучения Интернета 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.

Сколько языков в России?

- Около ста (для точного подсчета 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 нем пишут в интернете почти исключительно за границей;
- Про некоторые не вполне понятно, жив ли еще хоть один носитель (алеутский);
- После консультаций со специалистами мы остановились на цифре 96.

Содержание

Как можно собирать тексты

- Ездить в экспедиции, записывать, оцифровывать, выкладывать в интернет;
- Оцифровывать книги (любые другие печатные издания);
- Находить в интернете.

Примеры полевых корпусов

- Корпуса ИЭА РАН на ненецком, телеутском, шорском и эвенкийском – http://corpora.iea.ras.ru/corpora/
- Корпуса кетского, селькупского, эвенкийского – http://siberian-lang.srcc.msu.ru/ru
- Корпуса с текстами на арчинском, хиналугском, нганасанском, энецком и тофалараском – http://www.philol.msu.ru/~languedoc/rus/index.php
- Корпус бесермянского диалекта удмуртского языка – http://beserman.ru/corpus/search/?interface_language=ru

Содержание

Как собрать тексты из интернета?

- Придумать, как собирать ссылки;
- Собрать ссылки;
- Выкинуть все лишнее;
- Выкачать тексты по ссылкам;
- Выкинуть все лишнее;
- Оставить только тексты на нужном языке (выкинув все лишнее).

Почему просто не взять Википедию?

- Ботозаливки;
- Одной жанровой направленности – энциклопедической.

Топ-10 частотного списка словоформ татарской википедии:

№	Словоформа	Перевод/значение	Встречаемость
1	елга	“река”	132567
2	бассейны	“бассейн”	75706
3	су	“вода”	54689
4	буенча	“по”	48838
5	Русия	“Россия”	48722
6	урнашкан	“расположенный”	38043
7	км	“километр”	36962
8	һәм	“и”	27231
9	кече	“малый”	27203
10	дәүләт	“государство”	26888

Делал ли кто-то такое до нас?

- Проекты Kevin Scannell по сбору текстов из интернета и твиттера (в открытом доступе тольке триграммы и некоторые ссылки)
  - crubadan.org
  - indigenoustweets.com
- Проект Endangered Languages (справочная информация о языках со всякими материалами, картой и примерами текстов) – www.endangeredlanguages.com
- Языки народов России в Интернете. Список ресурсов – www.belti.ru/peoples/eng_index.html

Как собирали тексты мы

- Мы не можем выкачивать весь интернет, отдельно разбираясь с каждой страницей;
- Воспользуемся большой поисковой машиной;
- Находим сайты в поиске по специальным терминам;
- Фильтруем;
- Выкачиваем;
- Определяем язык.

С какими проблемами мы сталкивались?

- 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).
- Грязный интернет (вычищали полуавтоматически);

Содержание

Лексические маркеры

Необходимые условия:

- Слова из этого языка;
- Уникальны для этого языка;
- Частотные.

Лексические маркеры – слова, принадлежащие языку и однозначно его идентифицирующие.

Поиск осуществляется вручную (грамматики, словари, разговорники).

Лексические маркеры

Опциональное условие:

- Не содержат диакритических символов.

Потому что:

- Все пользуются бытовой системой письма (термин А.А. Зализняка);
- Не всегда можно установить однозначные соответствия;
- Например: ӧ → э, о.
  татшӧм → татшэм, татшом (коми-зырянский)

На примере русского языка

- Принадлежат языку: + молоко − мандрувати
- Есть только в одном языке: + возле − через
- Являются частотными словами: + под − горловина
- Не содержат диакритических символов*: + сельдь − селёдка

→ который – хороший маркер для русского языка

* Бытовая письменность (по звучанию или написанию)

Задание

- Разделитесь на команды (по 2-3 человека у одного компьютера).
- Выберите один из языков:
  - аварский,
  - адыгейский,
  - бурятский,
  - ингушский,
  - корякский,
  - лезгинский,
  - татарский,
  - тувинский,
  - удмуртский,
  - хакасский.
- Попробуйте, используя предоставленные разговорники, найти лексические маркеры для этого языка.

Соревнование

А теперь давайте сравним ваши маркеры с нашими! По чьим маркерам находится больше текстов на нужном языке, тот и победил :)

Наши маркеры

Слайды с нашими маркерами были удалены во избежания попадания их в интернет. Полную презентацию можно запросить, связавшись с нами через почту: *ru-lang@yandex.ru*

Содержание

Сбор ссылок

- Мы не можем выкачивать весь интернет, отдельно разбираясь с каждой страницей;
- Воспользуемся большой поисковой машиной;
- Лексические маркеры отправляются как запрос поисковику;

- В результате получаются списки сайтов;
- Повторно отправляем запрос с лексическим маркером и url сайта;

- Получаем список страниц на этом сайте;
- Сортируем полученные сайты.

Для автоматизации запросов используем Яндекс.XML.

Сортировка сайтов

Хорошие: содержат больше 30 страниц на интересующем нас языке.

Бедные: содержит несколько страниц на языке.

Большие: многостраничные сайты, на которых встретилось несколько страниц на данном языке (youtube.com и stihi.ru).

Плохие: сюда попадают сайты с рефератами, словарями и грамматиками, в которых маркеры встречаются вне естественной языковой среды.

Вконтакте: отдельная группа, в которую собираются ссылки на этот ресурс.

В зависимости от того, в какую категорию попал сайт, мы либо выкачиваем его целиком, либо формируем список нужных страниц.

Содержание

Выкачивание текстов

- Для большого интернета используются краулеры. Краулер, которым пользуемся мы, написан с помощью Scrapy, библиотеки для Python;
- Для Вконтакте используется API.

Данные про сайты

- url сайта;
- url страницы сайта;
- Заголовок, если есть;
- Текст между двумя тегами сайта;
- Техническое: дата выкачивания.

Данные про ВКонтакте

О каждом сообществе:

- Название;
- Число участников;
- Список всех постов и комментариев.

О каждом посте/комментарии:

- Текст;
- Дата создания;
- Число комментариев (для постов);
- Id поста-родителя (для комментариев);
- Информация об авторе: id, дата рождения; город проживания, пол.

Содержание

Бытовая система письма и code-mixing по-башкирски

Идентификация языка

(здорового человека)

Задача

Нужно определить, на каком языке из заданного списка написан текст.

- Текст представляется в виде упорядоченного по частоте набора буквенных n-грамм;
- Для каждого языка из списка создается эталонный набор буквенных n-грамм, также упорядоченный;
- Вычисляются расстояния между текстом и эталонами каждого из языков;
- Выигрывает язык, эталон которого оказался самым близким.

Идентификация языка

(человека, который связался с малыми языками)

Проблема

У нас нет текстов, чтобы сформировать эталоны.

Задача

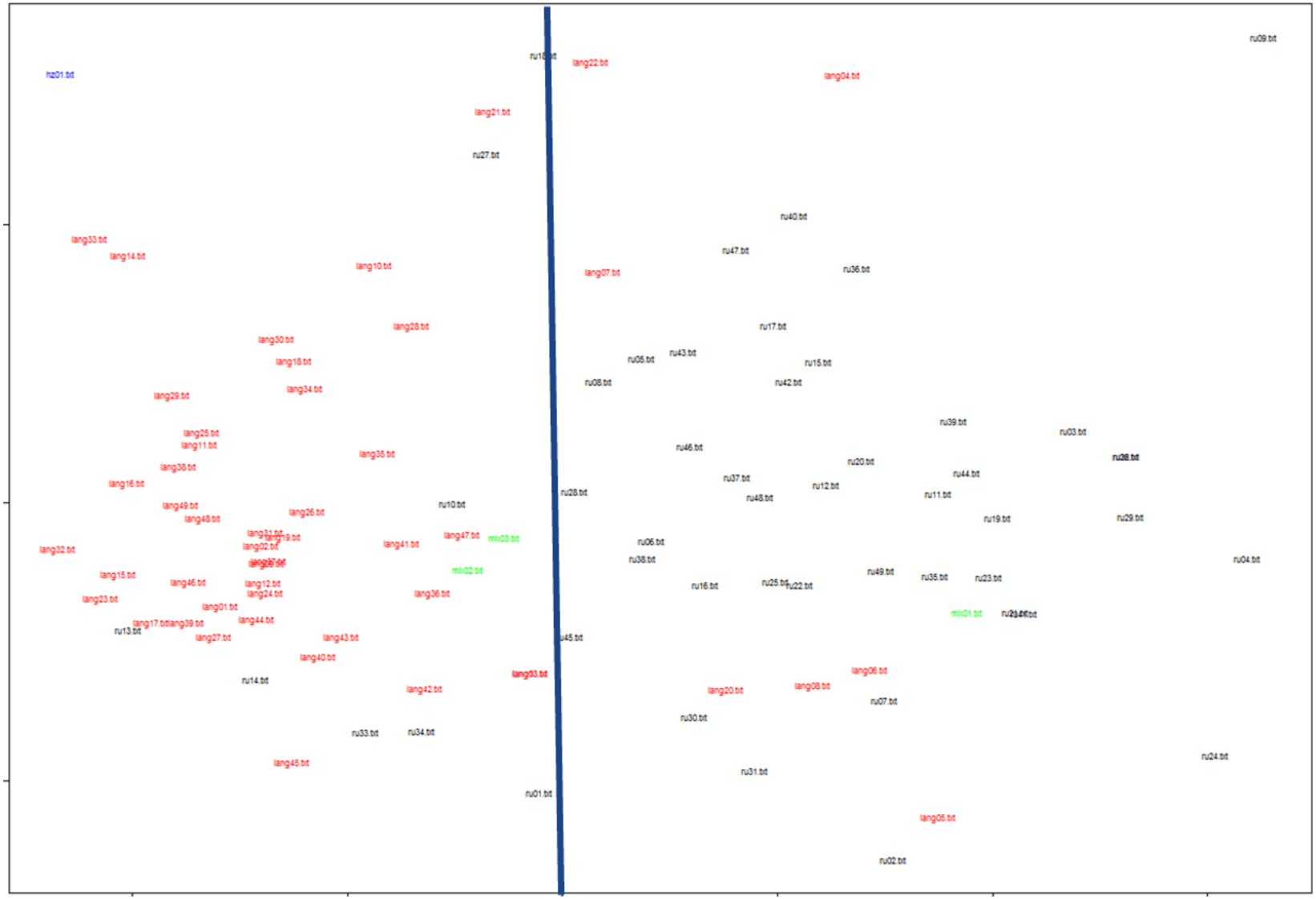
Нужно понять, написан ли текст на ожидаемом языке.

Идентификация языка

(человека, который связался с малыми языками)

- Текст представляется в виде упорядоченного по частоте набора буквенных n-грамм;
- Эталонный набор буквенных n-грамм создается только для русского языка;
- Вычисляется расстояние между текстом и эталоном для русского языка;
- Эмпирическим путем была определена граница отсечения;
- Если текст оказался очень близок к эталону (расстояние меньше границы), текст на русском;
- В противном случае считаем, что текст написан на ожидаемом языке;
- Дополнительно избавляемся от мусорных текстов.

Идентификация языка. Что получилось



Задание

Найти и категоризовать распространенные ошибки инструмента определения языка.
По возможности придумать и предложить решение для исправления ошибок.
Сделать мини-доклад для участников остальных групп.

Формат данных. Интернет


```
{
    "download_date": "2016-01-25 10:28:15.197639",
    "url": "http://www.abazashta.com/club/forum/forum2/
topic2/messages/?PAGEN_1=2",
    "domain": "www.abazashta.com",
    "language": "abq",
    "header": "",
    "text": {
        "85": {
            "language": "abq",
            "text": "Адац-ач1выйа йг1анышцтуа сахща?"
        },
        ...
    }
}
```

Формат данных. Вконтакте


```
"club40933116": {
    "name": "Учим аварский язык вместе",
    "posts": {
        "143": {
        "sort": 1357381128,
        "author": {
            "bdate": "26.6.1992",
            "city": "Махачкала",
            "sex": 1,
            "id": 11182
            },
        "language": "ava",
        "date": "2013-01-05 13:18:48",
        "text": "гьаналъ ханклал -мясные хинкал",
        "comments": {}
        },
```

Содержание

Содержание

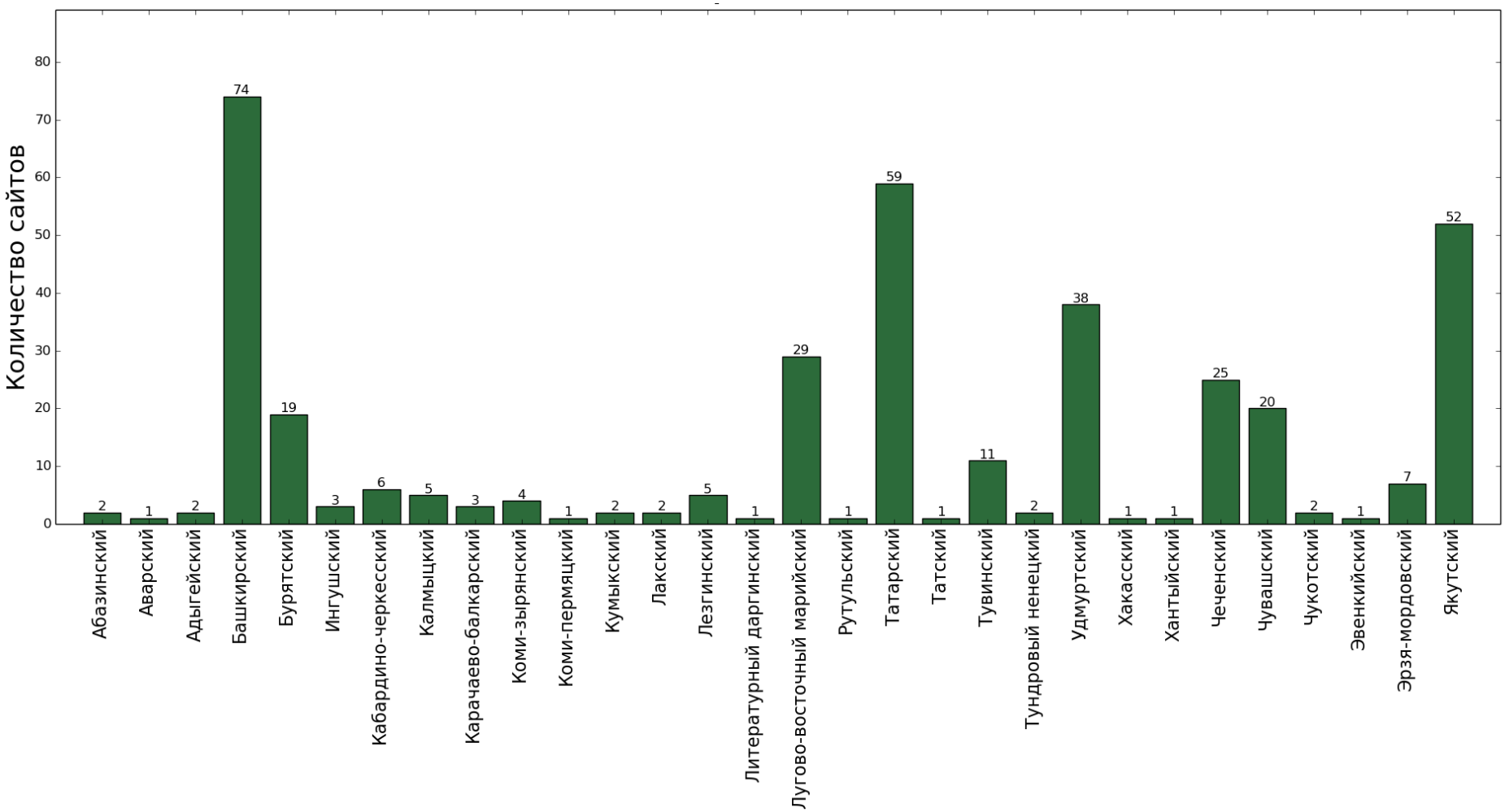
Интернет

- Искали 49 языков;
- Нашли 48 языков, для которых есть хотя бы одна страница;
- 30 языков, для которых есть хотя бы один сайт на малом языке;
- 379 сайтов на малых языках;
- Выкачали чуть меньше.

На поиск ушло 239 дней.

Интернет

Сколько сайтов на каждом языке?



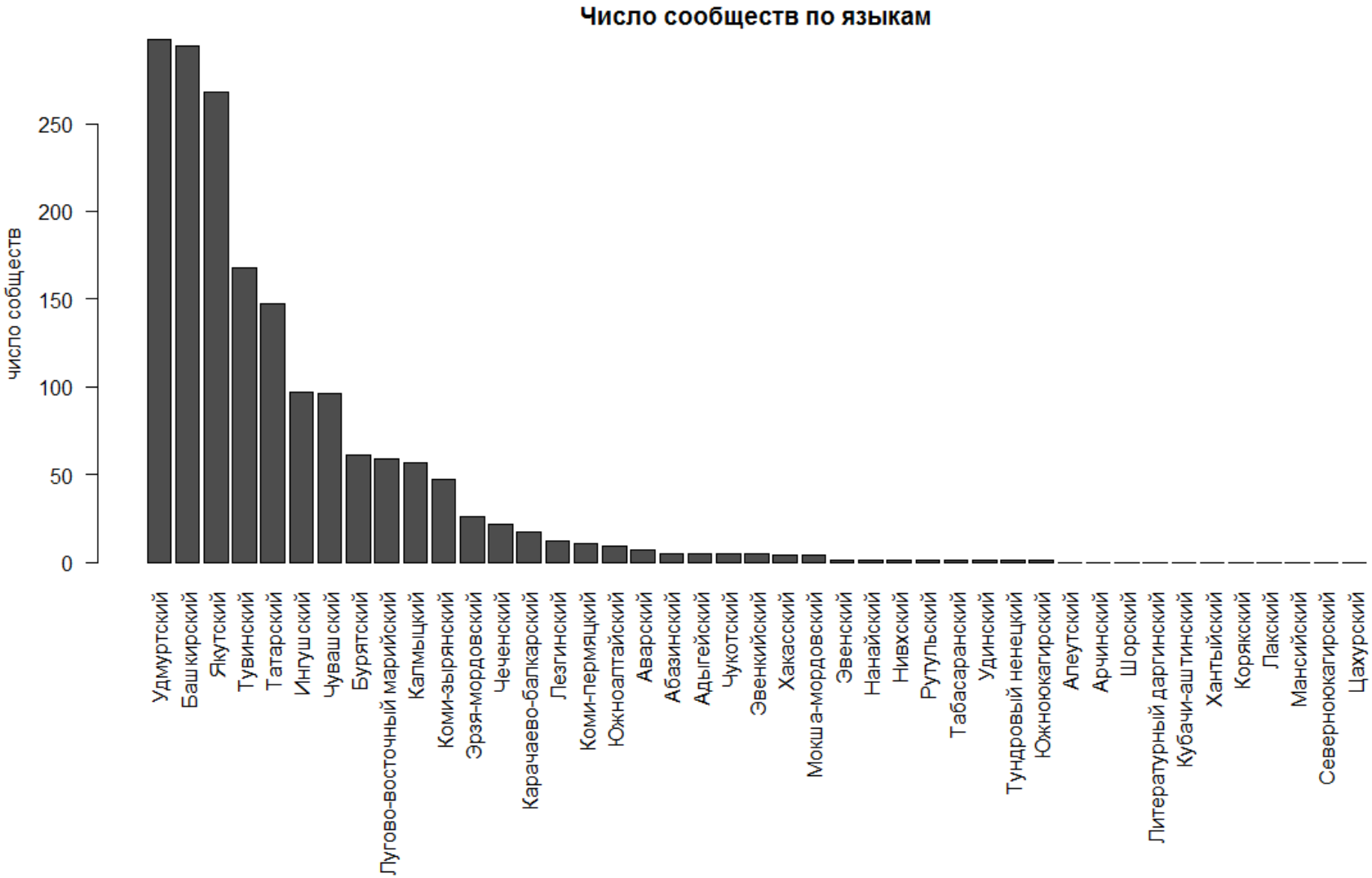
Другие данные на сайте проекта: http://web-corpora.net/minorlangs

Содержание

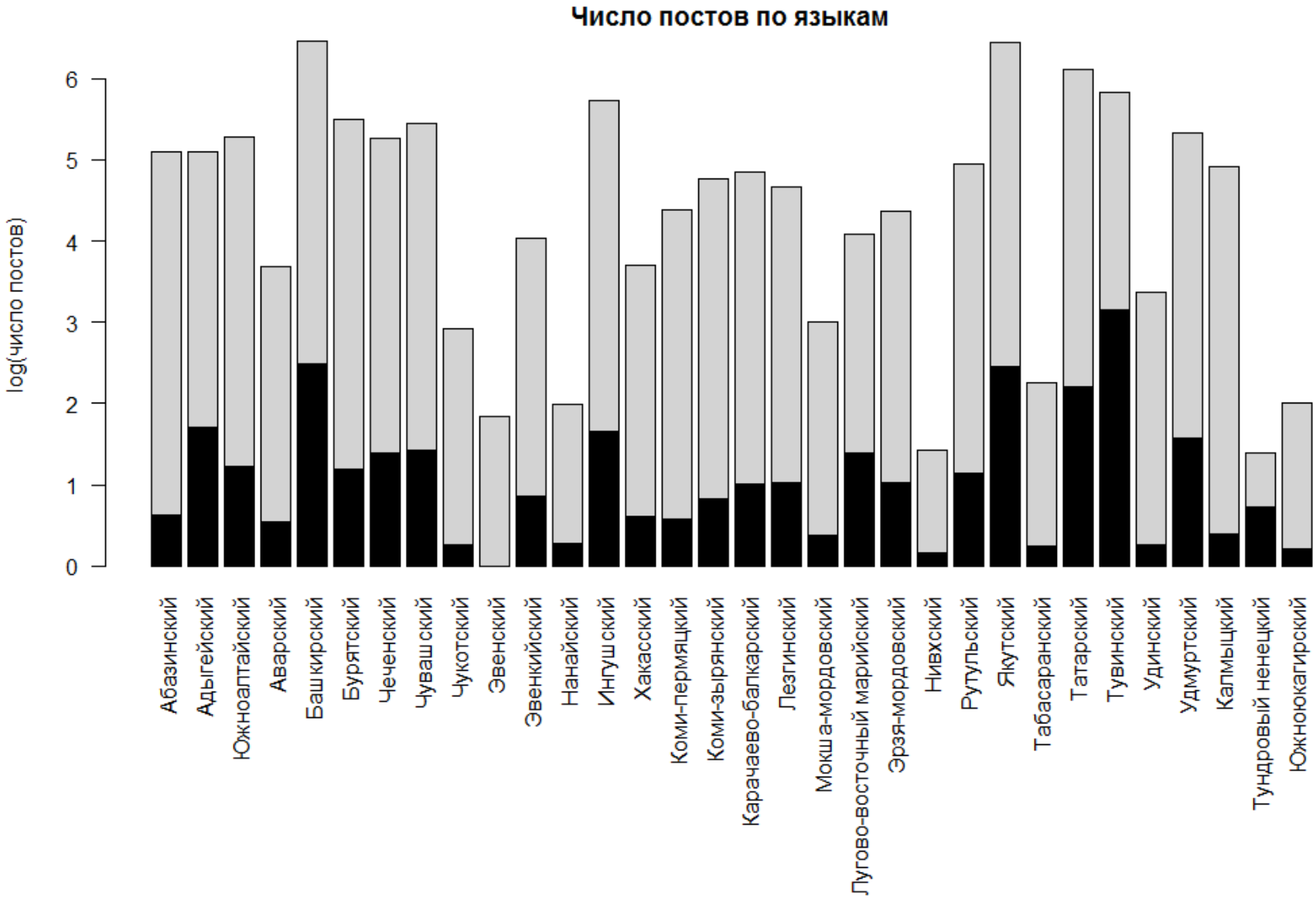
ВКонтакте

- Искали 49 языков;
- Нашли 35 языков;
- 1630 сообществ, в которых есть хотя бы 1 текст;
- Но не все содержат тексты только на миноритарном языке.

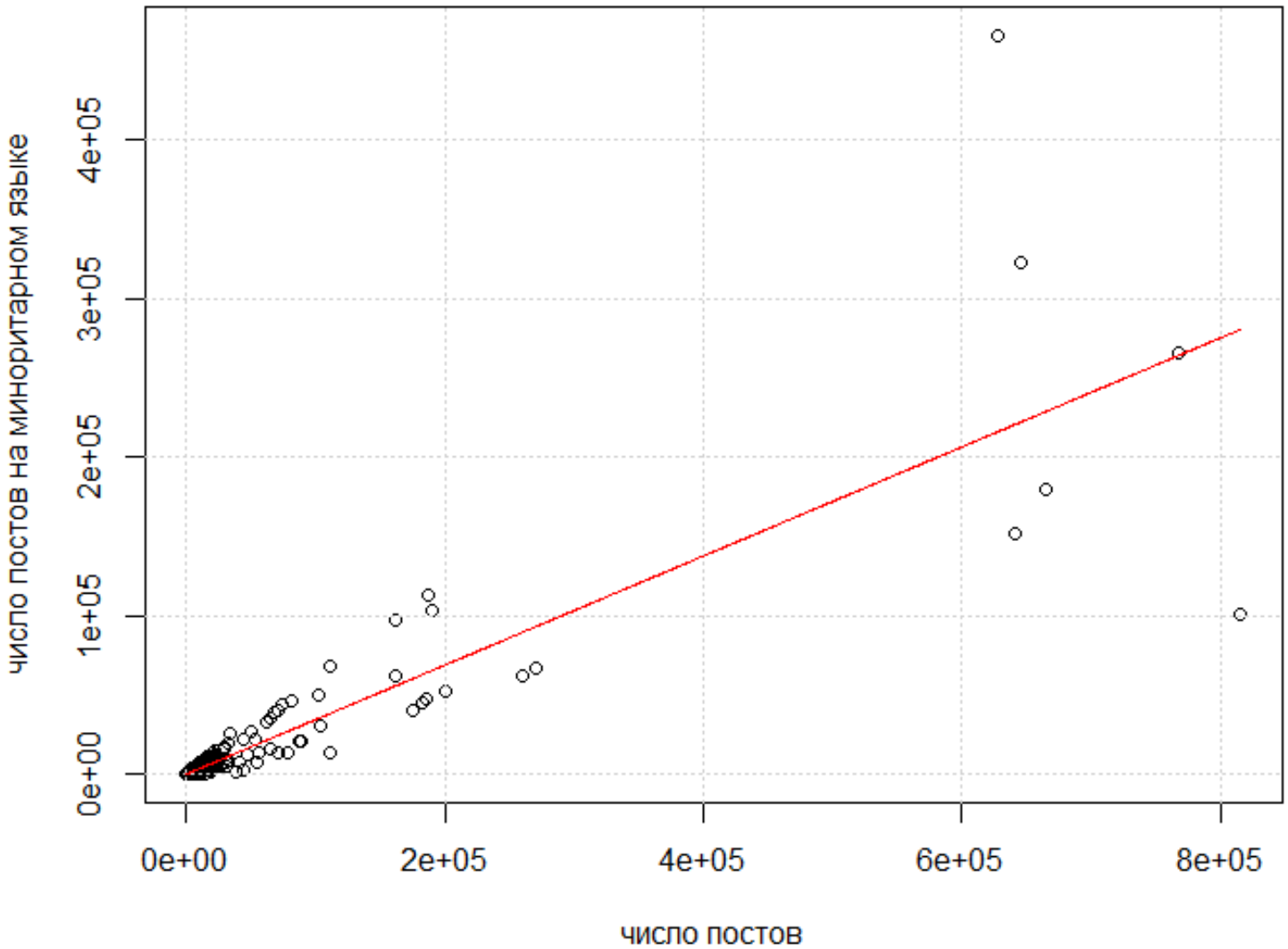
Сообщества и языки



Посты и языки



Зависимость постов на языке от общего числа постов



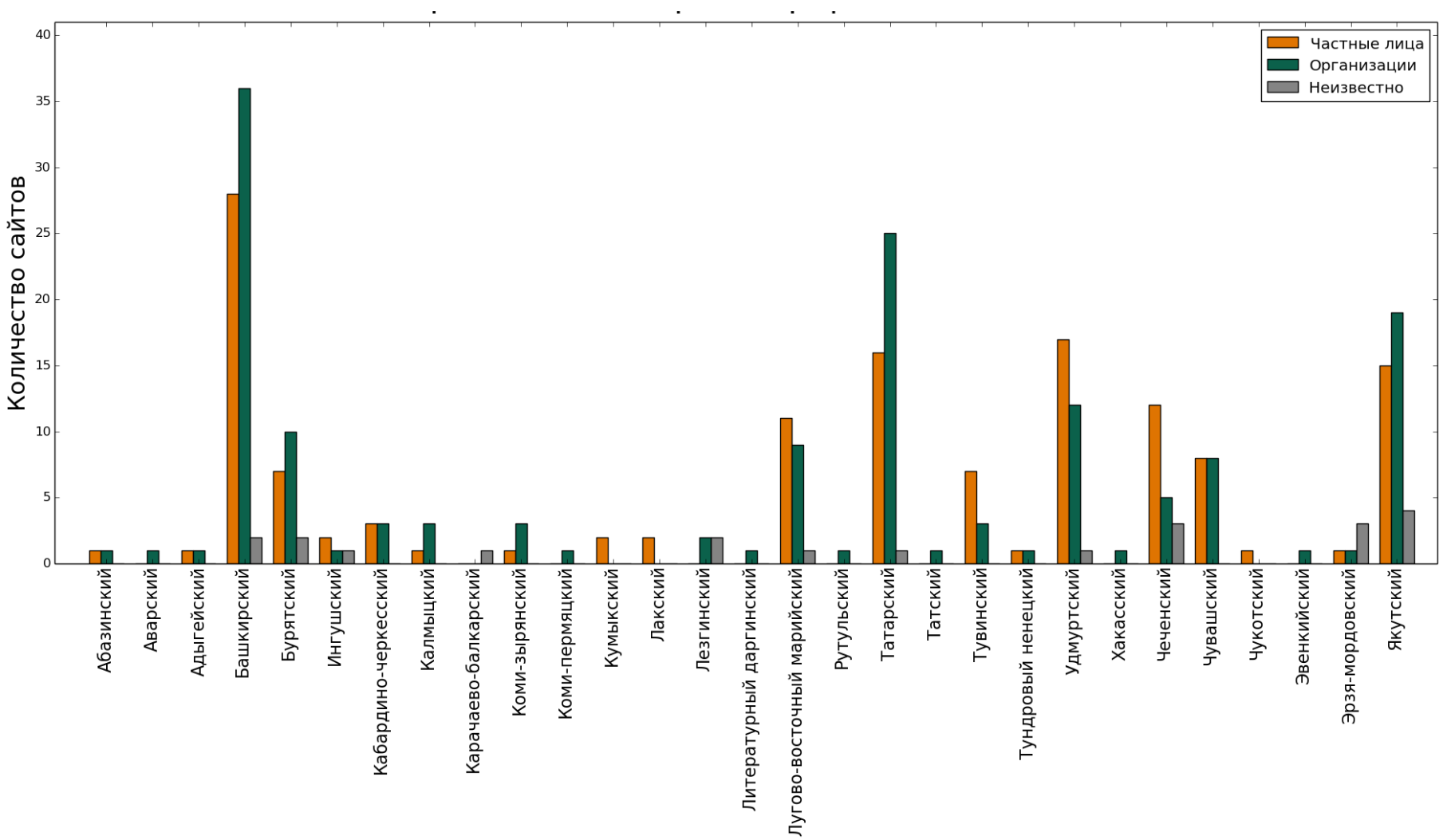
Сайт

http://web-corpora.net/minorlangs

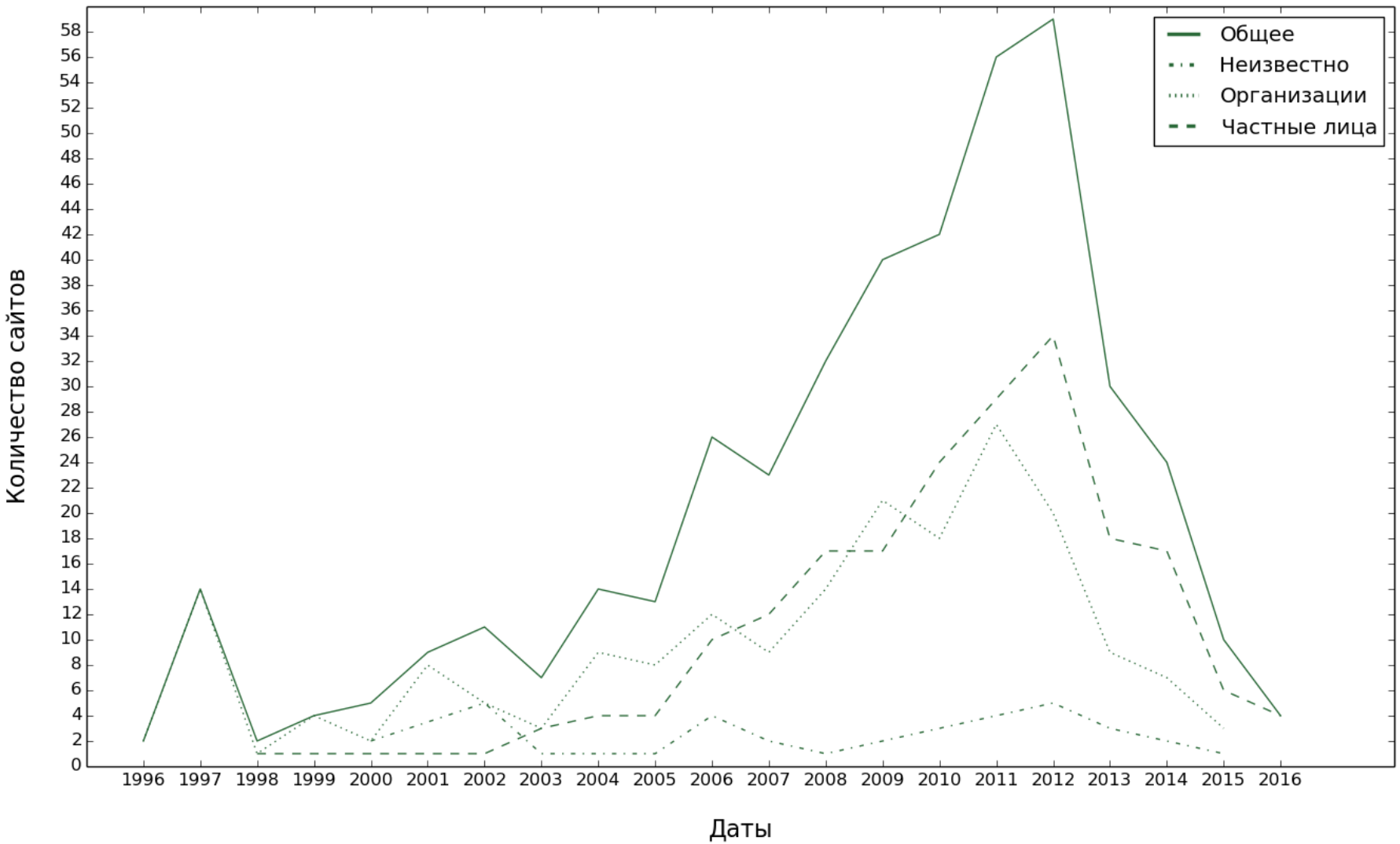
Содержание

Содержание

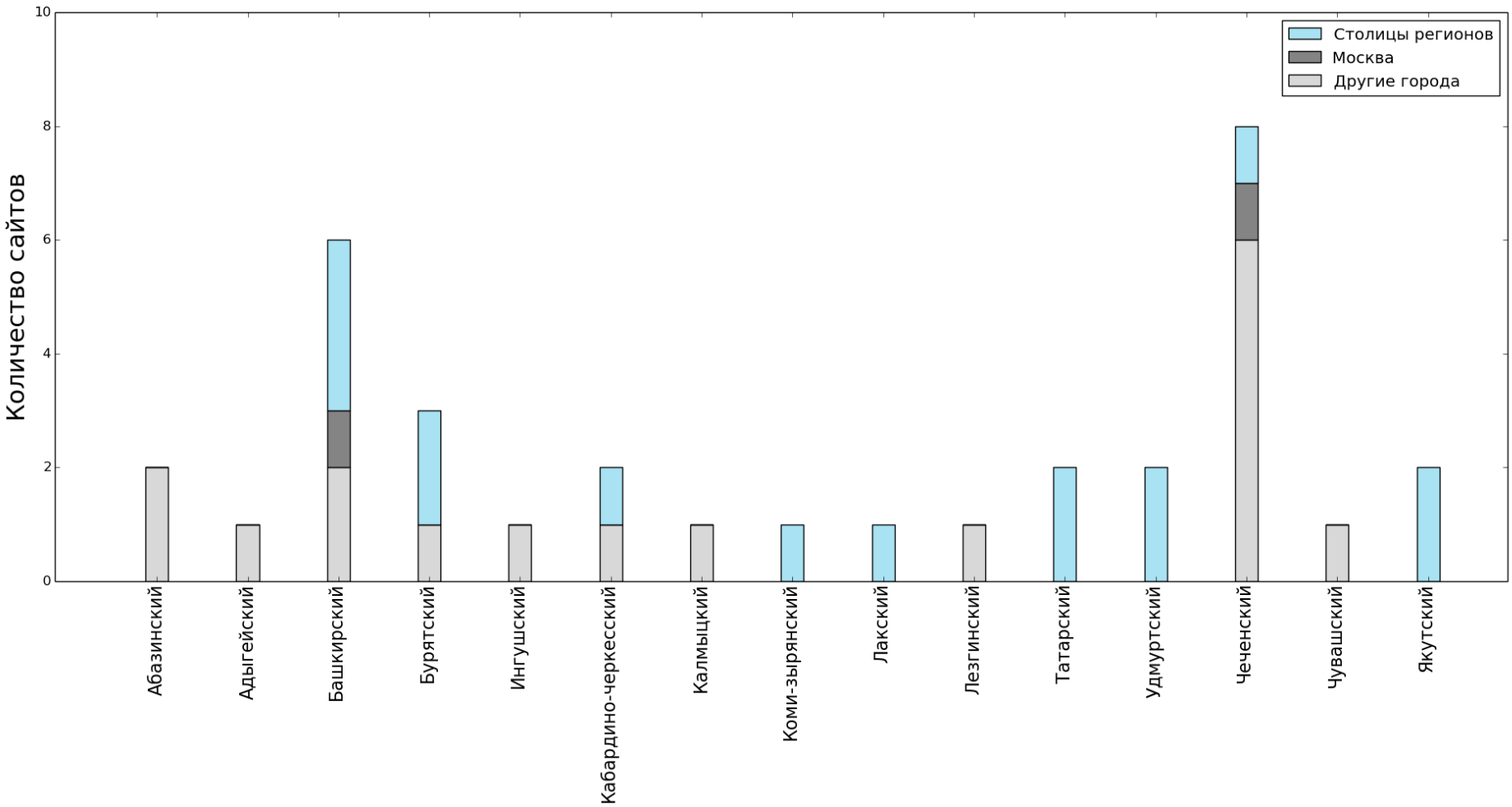
Кто регистрирует домены



Когда регистрируют домены



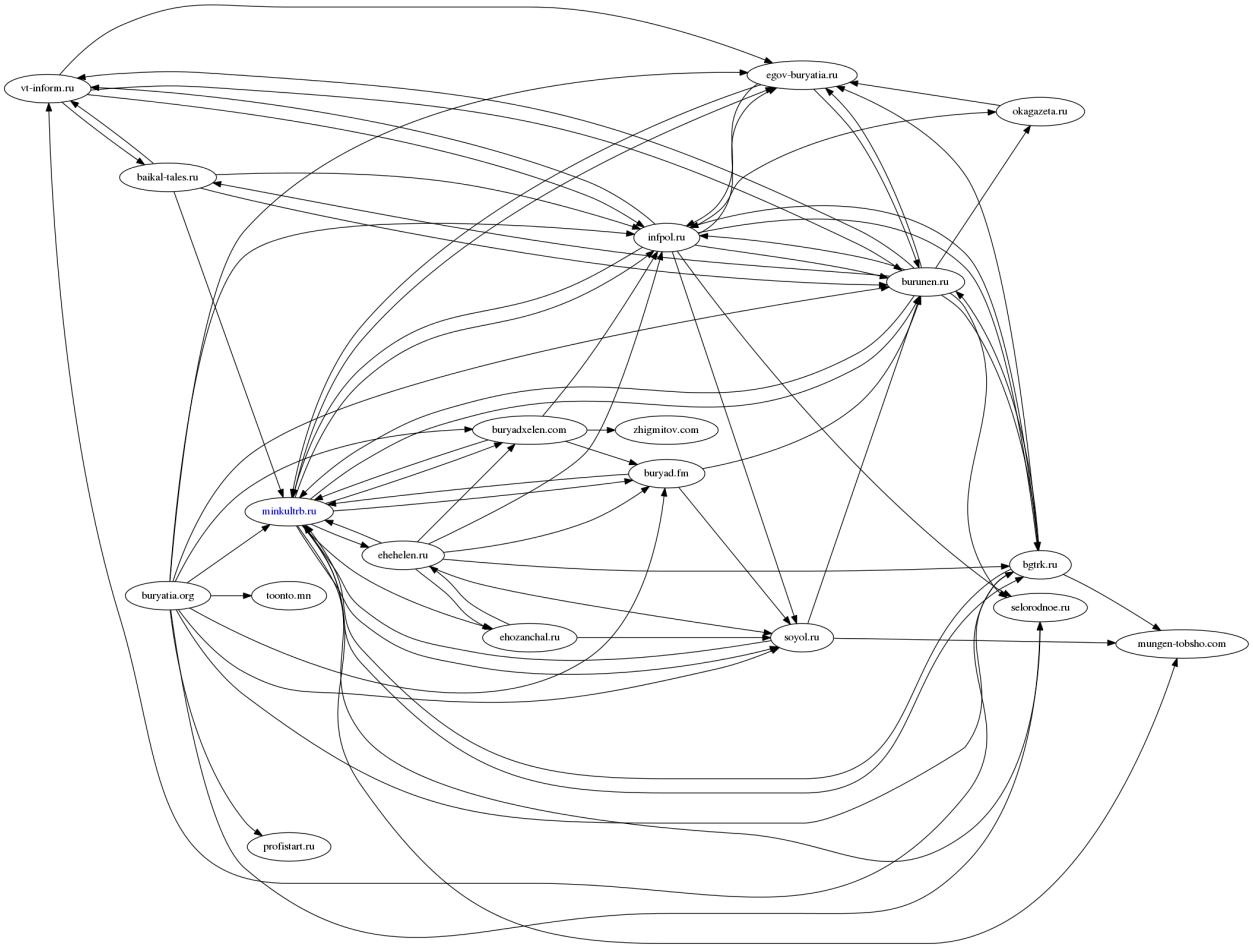
Где регистрируют домены



Веб-граф национальных интернетов

Характеристики:
Ассортативный
Не слабо связный

Веб-граф бурятского национального интернета



Характеристики:
Дисассортативный
Слабо связный

Содержание

Можно построить значимую модель линейной регрессии, объясняющую 65-75% данных.

В модель будут входить параметры:

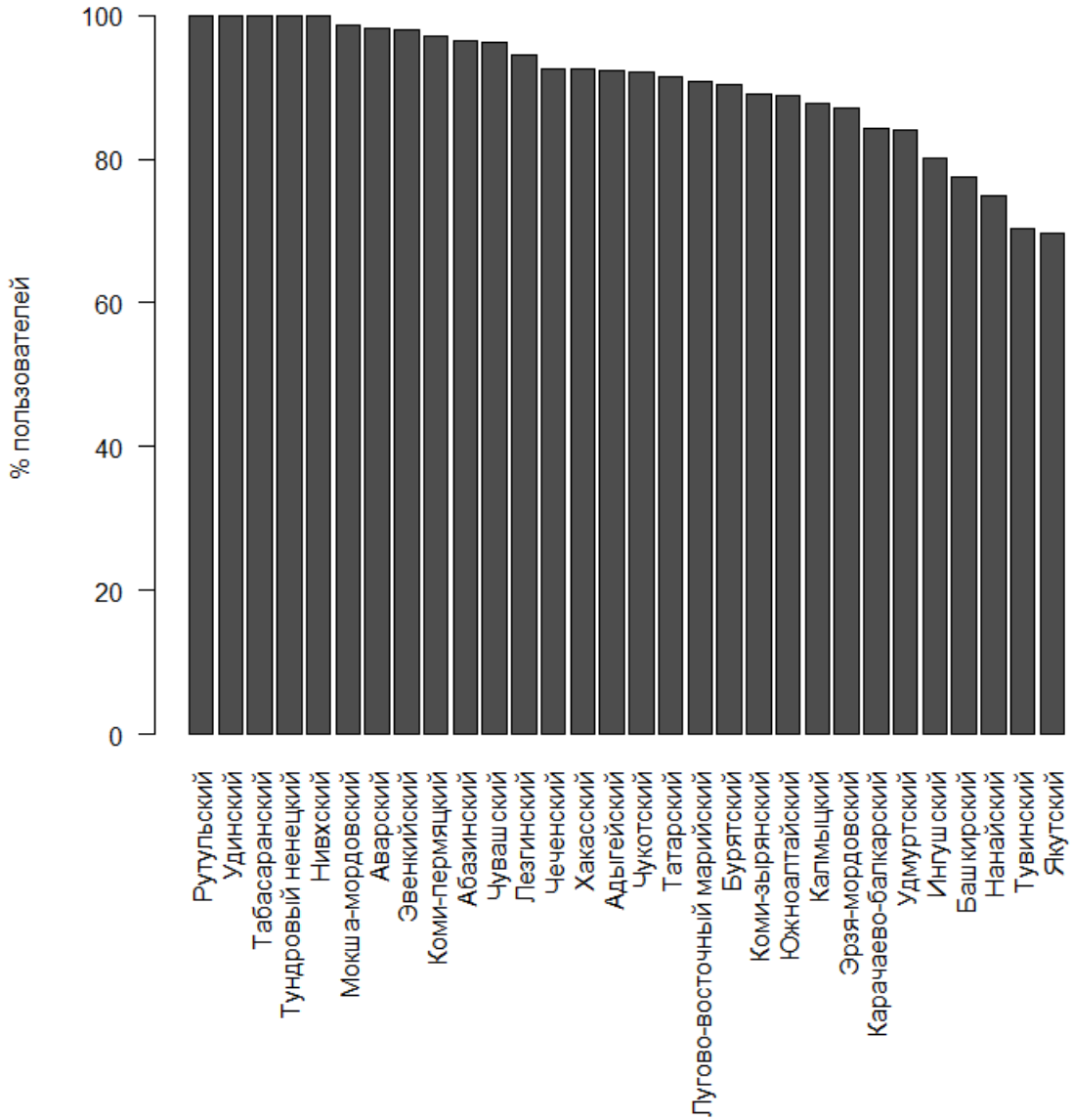
- ln(кол-во статей в вики за 2016г);
- ln(кол-во токенов в ВКонтакте);
- ln(число носителей за 2010г);
- прирост населения за 2014г;
- расходы региона за 2014г;
- ln(число газет за 1996г).

Содержание

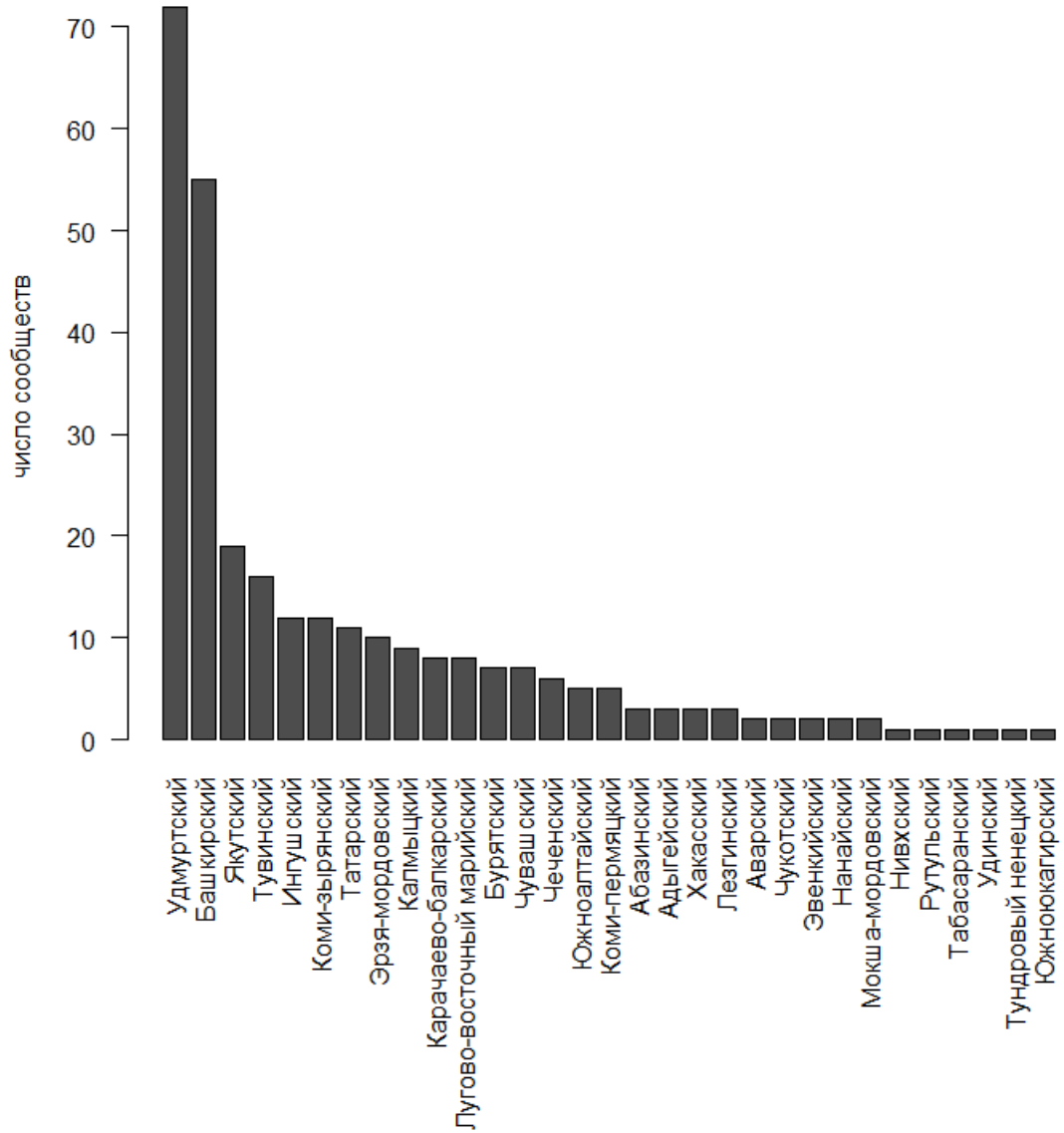
Пользователи

- Пользователь является носителем миноритарного языка, если он написал по крайней мере один пост на этом языке;
- Количество пользователей-носителей не зависит линейно от количества носителей, известного из переписи населения;
- Среди носителей 23 языков регионом-лидером является титульный регион;
- Средний возраст носителей всех языков находится в промежутке 19,6-31,3 лет; самым «юным» языком является чукотский, а самым «взрослым» – коми-зырянский;
- Абсолютное большинство всех пользователей участвует в жизни только одного сообщества, «говорящего» на миноритарном языке: наименьший показатель у пользователей якутского языка – 69,6%;
- 57,4% пользователей-носителей написали только один пост на миноритарном языке; средний показатель равен четырем.

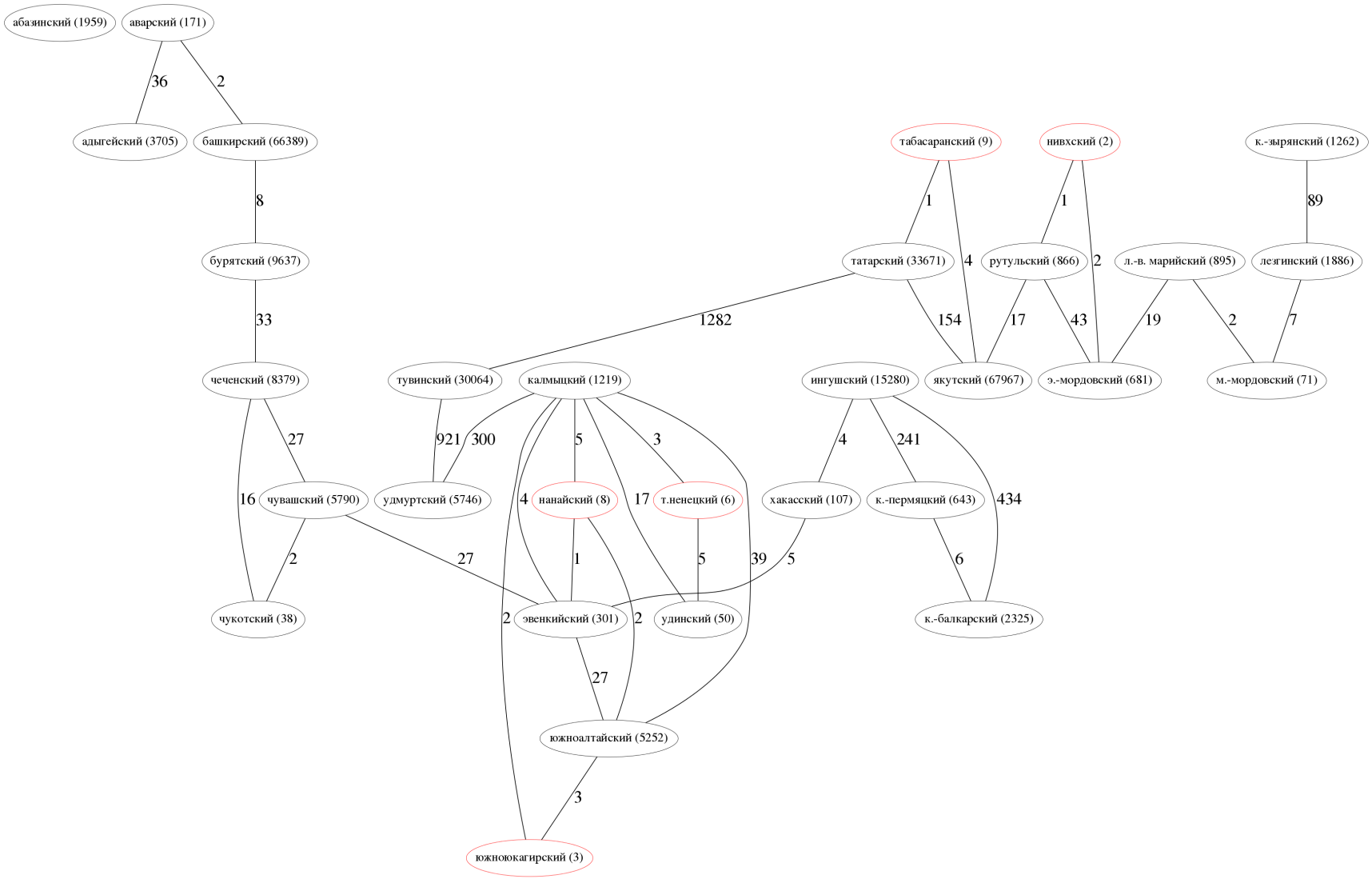
Доля написавших только в одно сообщество



Максимальное число сообществ у одного пользователя



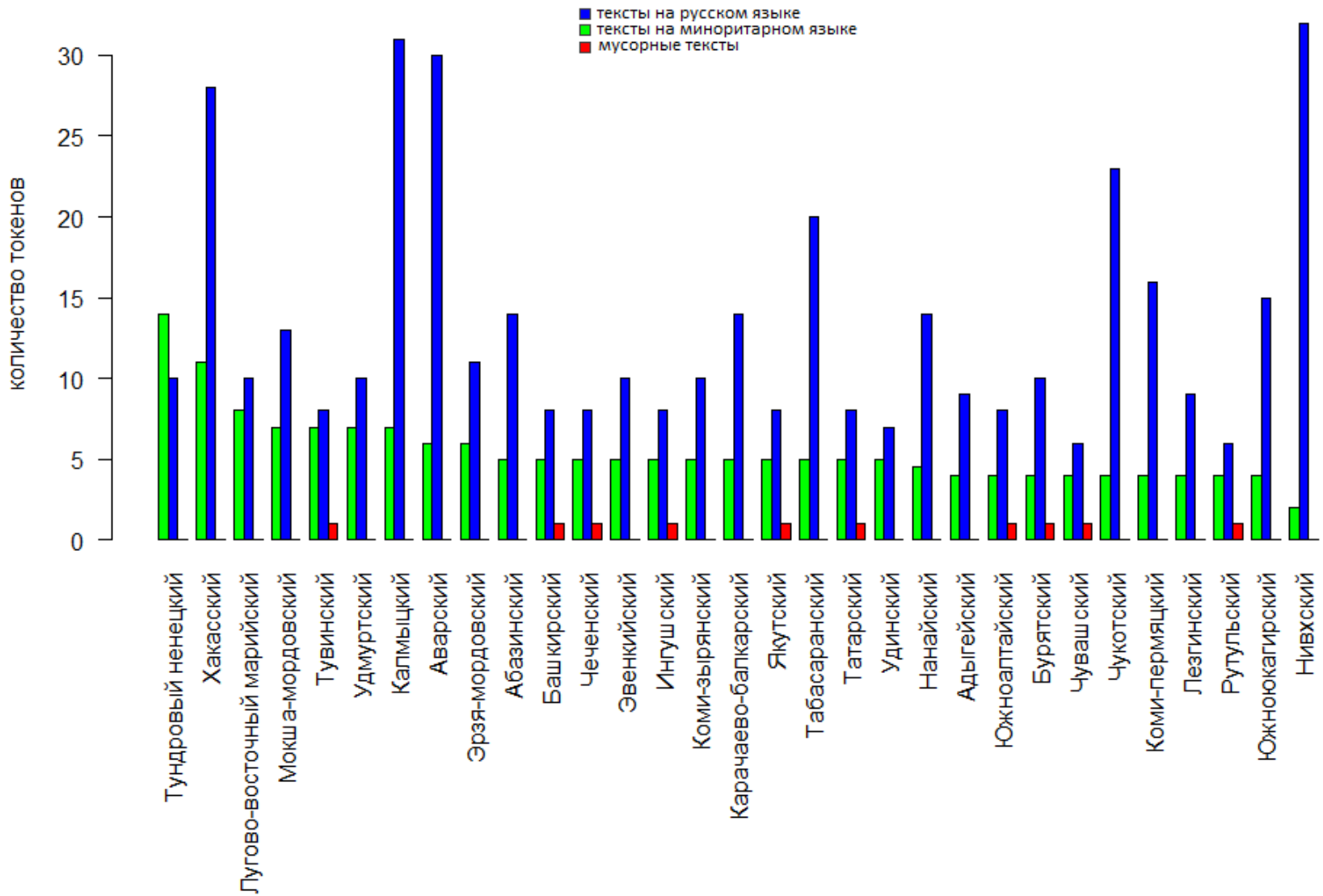
Языковые способности пользователей



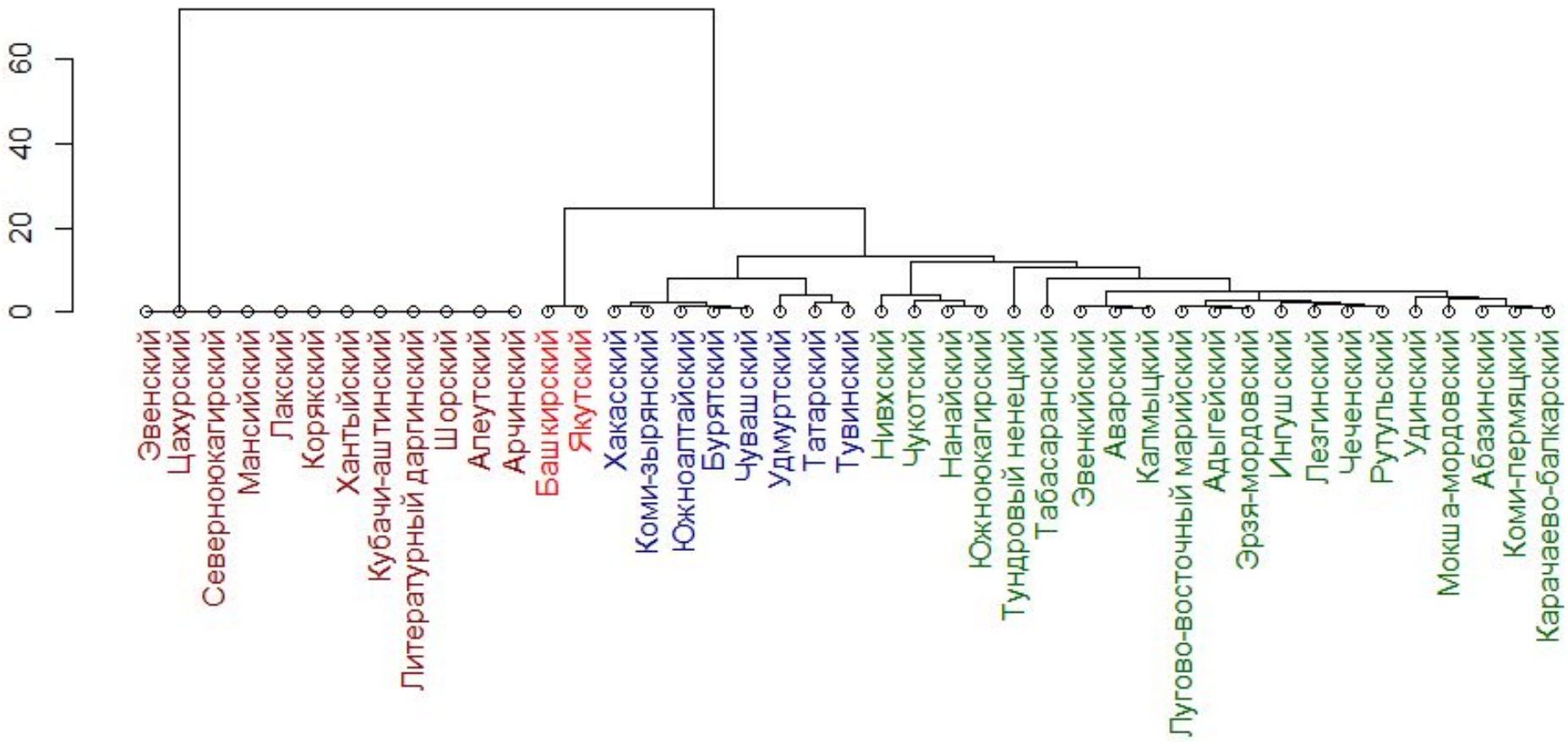
Сообщества

- Количество сообществ на миноритарном языке не зависит линейно от количества пользователей, «говорящих» на этом языке;
- Тексты на миноритарном языке короче текстов на русском: в среднем текст на миноритарном языке состоит из 5 слов, а на русском – из 13;
- Треть всех текстов сообщества написана на миноритарном языке;

Длина текстов



Иерархическая кластеризация языков



Портрет типичного носителя

Типичный пользователь-носитель миноритарного языка:

- Возраст: 19-31;
- Место проживания: титульный регион;
- Являясь носителем миноритарного языка, в основном «говорит» по-русски;
- Вероятнее всего, принимает участие в жизни только одного сообщества.

Типичное сообщество, «говорящее» на миноритарном языке:

- Треть всех постов написана на миноритарном языке;
- Длина текста на миноритарном языке в среднем около пяти слов.

Содержание

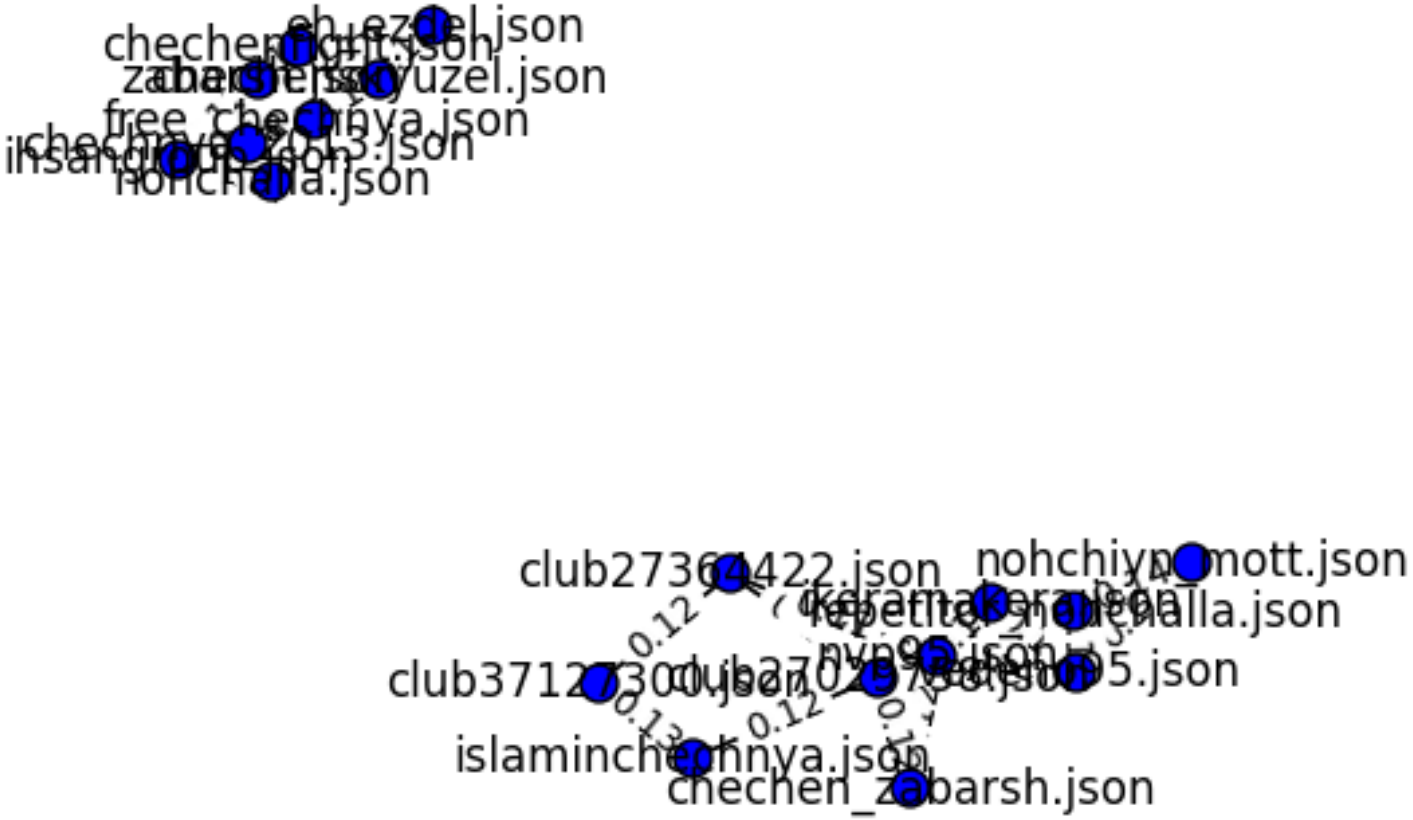
Можем ли мы узнать тематику записей?

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

Пример данных


```
    "213": {
        "date": "2015-07-18 01:49:02",
        "text": "-Астог1фируллох1, 1аллеелай!\n-Х1ухилла?\n
-Охьахаал т1аккха дуьцар ду хьун!\n-Хии, схьадица х1ухилла.",
        "author": {
            "bdate": "8.8.1999",
            "id": 2841,
            "city": "Уфа",
            "sex": 2 }
    },
    "115": {
        "date": "2015-03-11 21:09:03",
        "text": "-Я от тебя ухожу, ты меня не любишь!\n
-Дорогая что ты говоришь?!",
        "author": "chechen_zabarsh"
    }
```

Результат: чеченский



Как получить такие графы?

Предобработка данных

- Оставить только тексты из всех данных;
- Определить язык (русский, мусор, малый) и оставить только тексты на языке;
- Токенизировать;
- Убрать пунктуацию и смайлики;
- Убрать наиболее частотные слова из каждого сообщества.

Как получить такие графы?

Метрика близости

- Каждое сообщество – вектор длины N, где N – сумма всех токенов во всех сообществах одного языка;
- Каждая компонента вектора соответствует одному токену и отражает вес (число употреблений) этого токена в конкретном сообществе;
- Для сравнения полученных векторов мы использовали косинусную меру сходства.

Как получить такие графы?

Расчет метрик и построение графа

Вершины: сообщества.
Ребра: соединяют сообщества друг с другом.
Вес: косинусная мера схожести.

Дополнительные условия:

- рисовали только ребра с весом > 0.1;
- рисовали только вершины, у которых есть исходящие и/или входящие ребра.

Графовые метрики

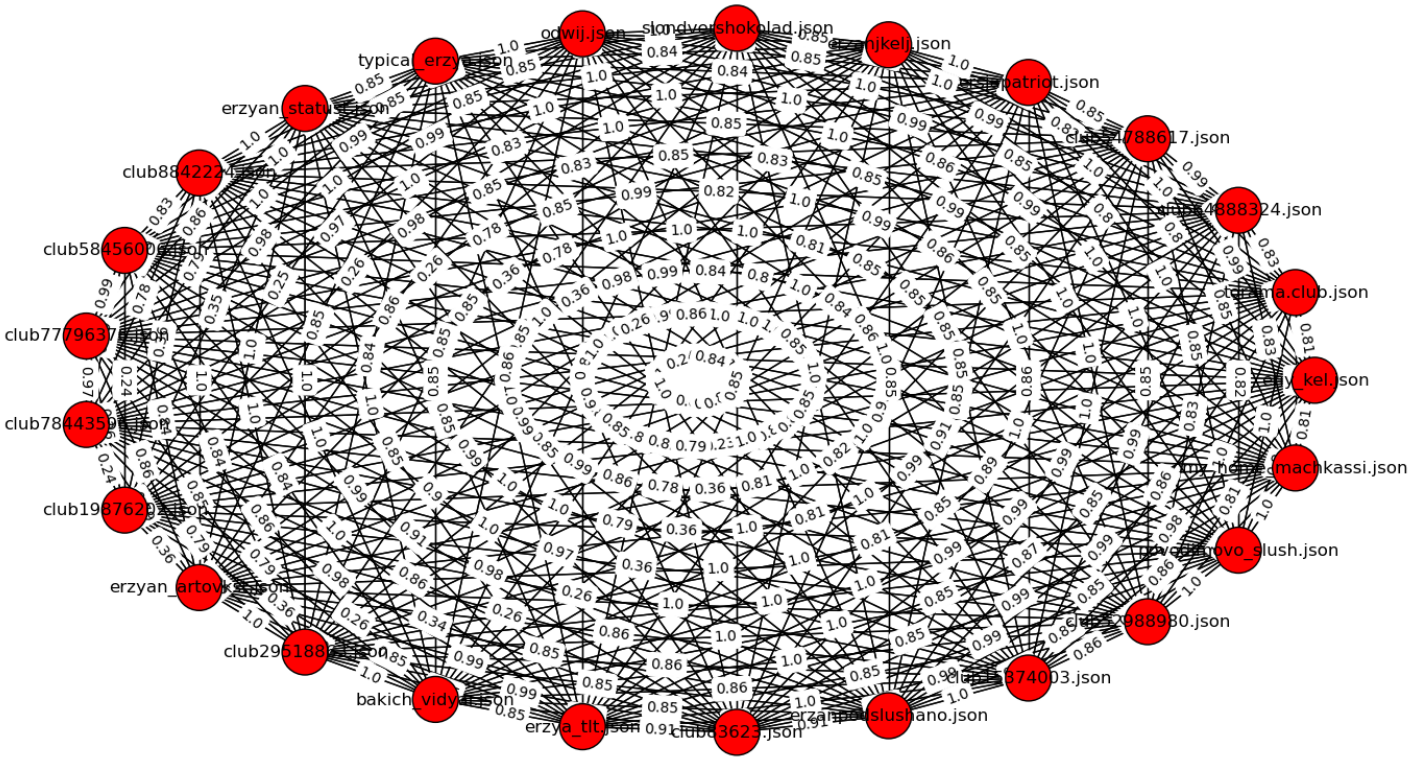
Степень центральности вершины – число ребер, входящих в эту вершину.

Среди вершин-сообществ мы находим вершину с наибольшей степенью центральности (Maxdc). Мы рассчитываем, что сообщество, соответствующее этой вершине, является наиболее типичным среди сообществ этого языка.

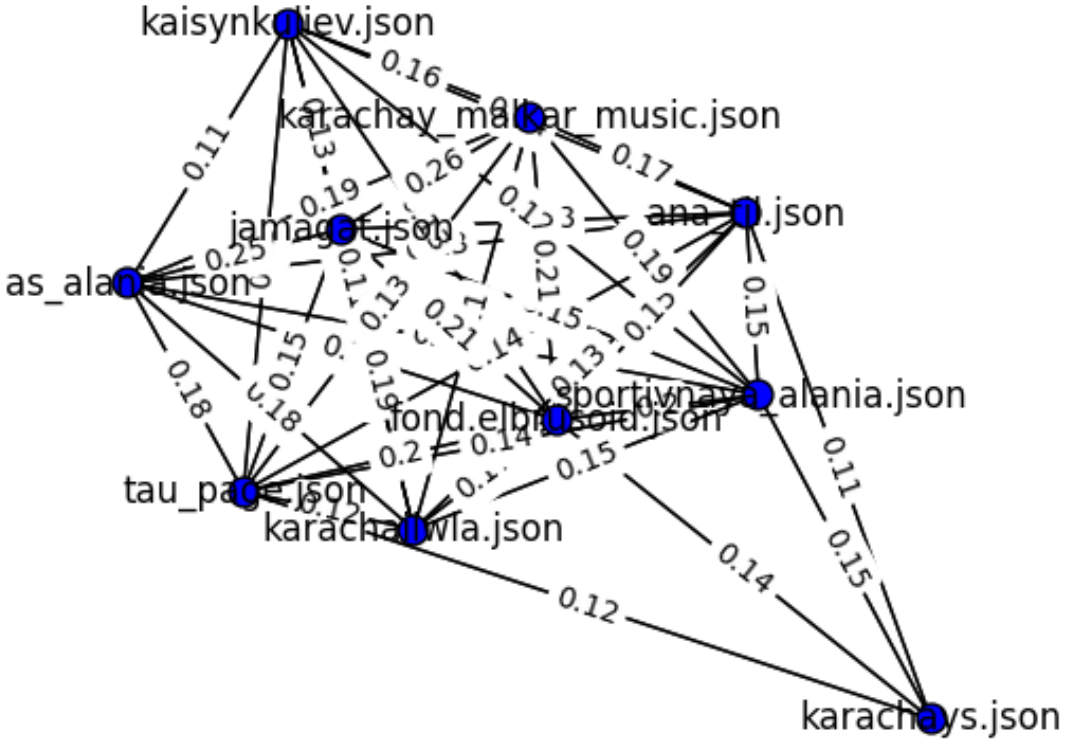
Ассортативность – метрика, которая показывает похожесть вершин.

Коэффициент ассортативности (AC) принимает значения от -1 до 1. Мы ожидаем, что сообщества, «говорящие» на определенные темы, будут связаны ребрами с сообществами, которые «говорят» на эти же темы.

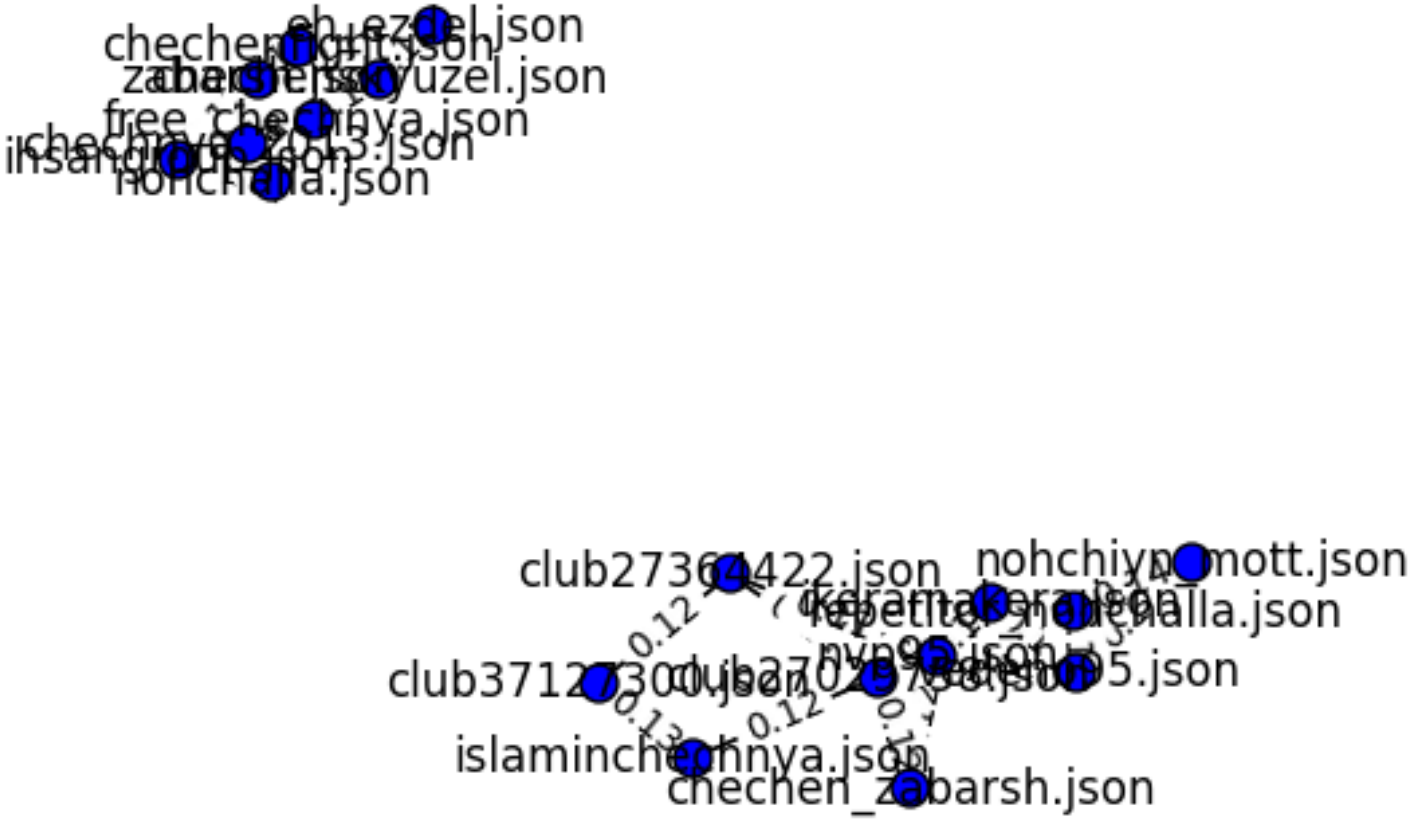
Результат: что будет, если не выставить порог?



Карачаево-балкарский



Чеченский


nohchiyn_mott.json
islamin_chechnya.json
chechen_zabarsh.json

Чеченский юмор

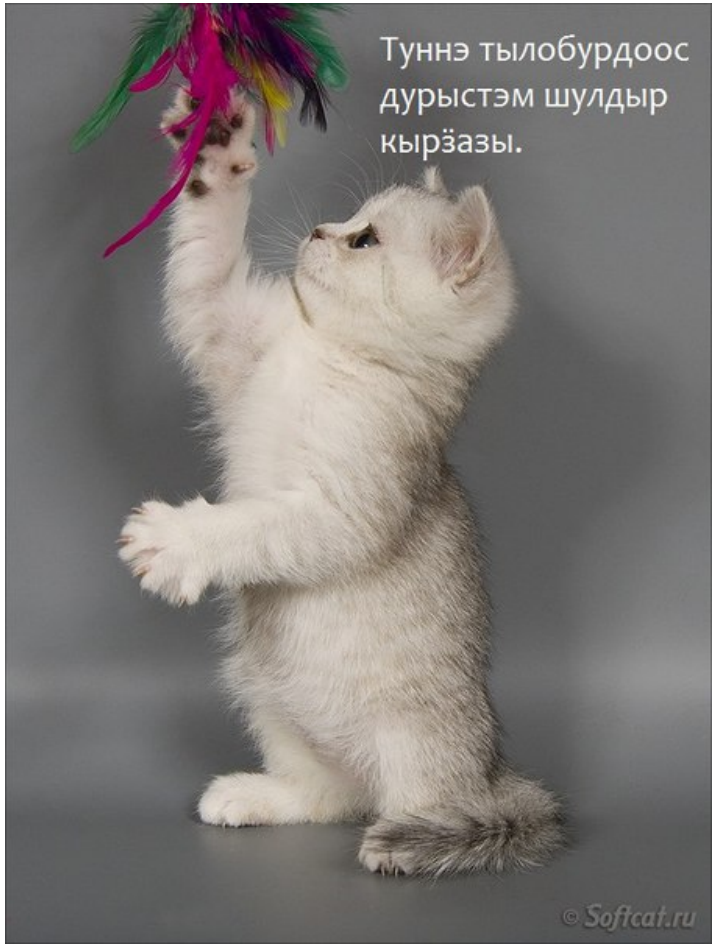
Задание

Гипотеза

Составляющие несвязанного графа имеют явную тематическую направленность.

Задача

Проверить гипотезу.


Туннэ тылобурдоос
дурыстэм шулдыр
кырӟазы.
© Softcat.ru