Малые языки России в интернете

Л.Зайдельман

а также: И.Крылова, Б.Орехов, И.Попов, Е.Степанова

19 марта 2016

Содержание

Содержание

О малых языках

- В России больше ста языков, не считая русского и языков, являющихся титульными в других государствах.
- Часть языков не имеет письменности.
- Для некоторых языков открыт вопрос «язык vs. диалект».
- Для части языков, увы, непонятно, жив ли еще хоть один носитель.

Содержание

Малые языки в интернете

- Есть ли малые языки России в интернете? А где? А сколько?
- Все носители малых языков России двуязычны, поэтому кажется, что можно не создавать ресурсов на малых языках.
- Но есть же Википедия! (На самом деле нет)

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).

Лексические маркеры

Необходимые условия:

- Слова принадлежат языку;
- Уникальны для этого языка;
- Частотные.

Лексические маркеры

Слова, принадлежащие языку и однозначно его идентифицирующие.

Лексические маркеры

Опциональное условие:

- Не содержат диакритических символов.

Потому что:

- Все пользуются бытовой системой письма;
- Не всегда можно установить однозначные соответствия;
- Например: ö $\rightarrow$ э, о.

Сбор ссылок

- Мы не можем выкачивать весь интернет, отдельно разбираясь с каждой страницей;
- Воспользуемся большой поисковой машиной (например, Яндексом);
- Лексические маркеры отправляются как запрос поисковику;
- В результате получаются списки доменов.
- Сортируем домены.

Сортировка доменов

- Хорошие: содержат больше 30 страниц на интересующем нас языке;
- Бедные: содержит несколько страниц на языке;
- Большие: многостраничные сайты, на котором встретилось несколько страниц на данном языке (youtube.com и stihi.ru);
- Черный список: 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 домен, мы либо выкачиваем его целиком, либо формируем список нужных страниц.

Выкачивание текстов

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

Формат данных. Интернет


```
"download_date": "2016-01-25 10:28:15.197639",
"url": "http://www.abazashta.com/club/forum/forum2/topic2/messa
"domain": "www.abazashta.com",
"language": "abq",
"header": "",
"text": {
    "85": {
        "language": "abq",
        "text": "Адац-ач1выйа йг1анышщтуа 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 ханklал -мясные хинкал",
        "comments": {}
    },
```

Идентификация языка

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.

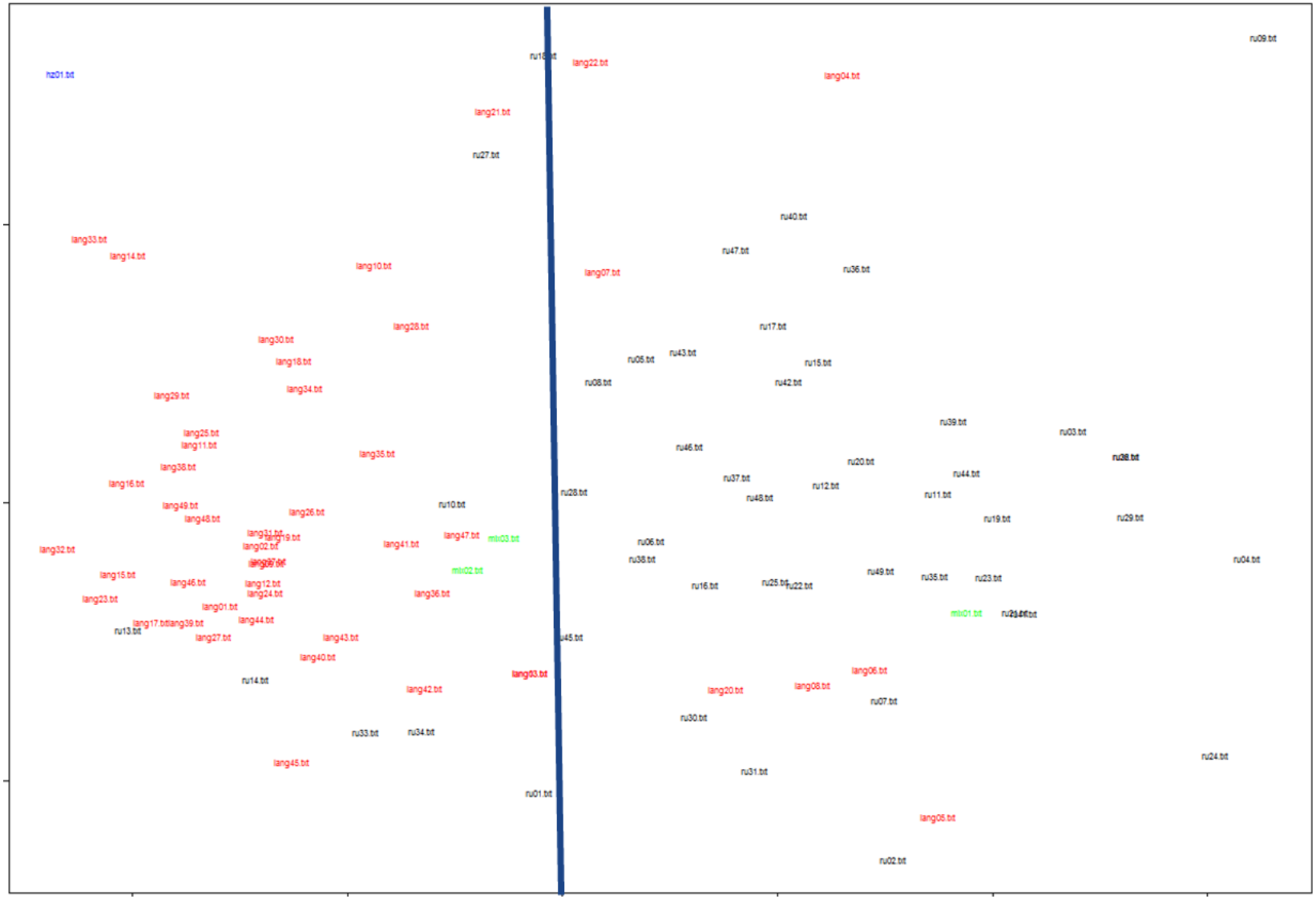
Задача

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



Содержание

Что получилось

- 96 языков;
- для 49 самых распространенных проведен поиск;
- что-то нашлось для примерно 30;
- 368 доменов;
- 1735 сообществ Вконтакте.

Содержание

Что интересного можно узнать

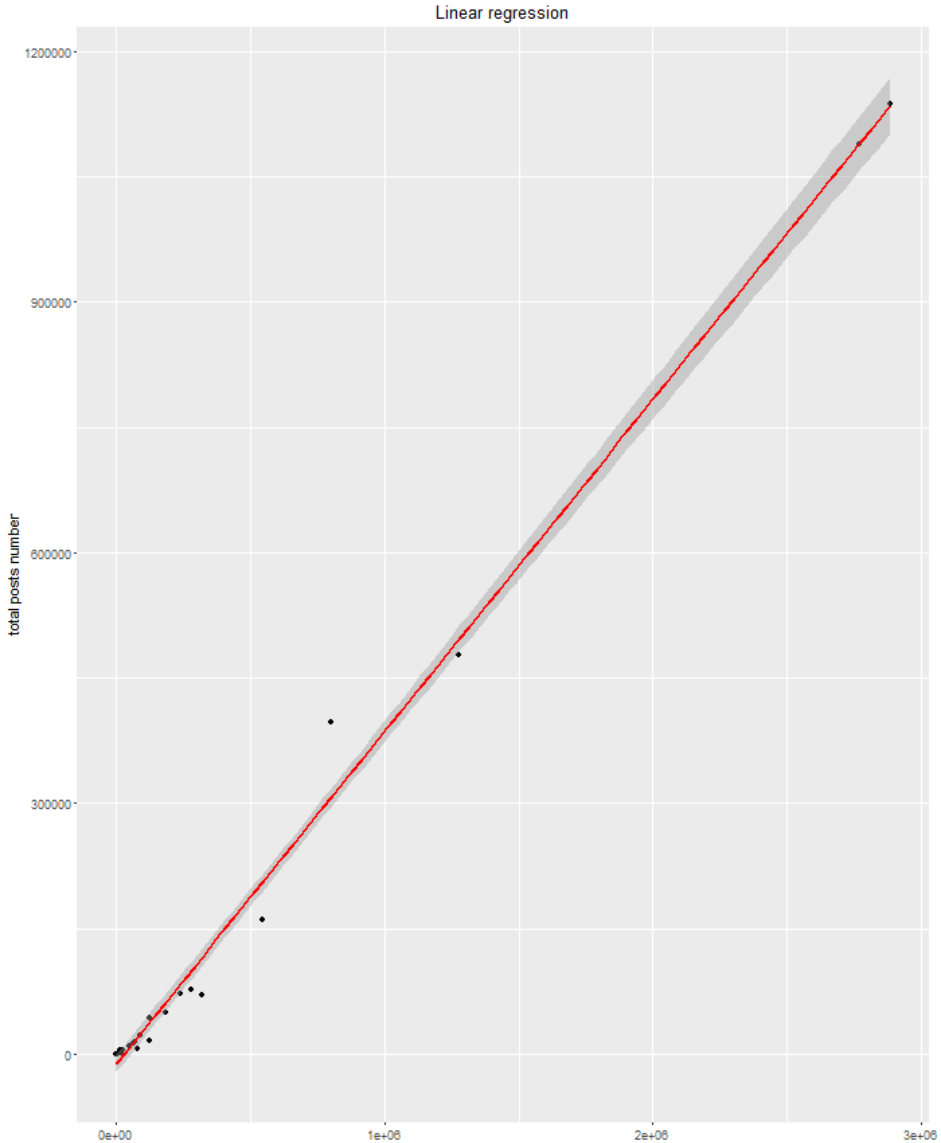
Чем больше язык, тем больше его в интернете?

Оказалось, что ни число доменов, ни число сообществ Вконтакте, ни число текстов, полученных из этих сообществ, не коррелируют с числом носителей.

Чем больше сообщество, тем больше в нем говорят на малом языке?

И да и нет. С одной стороны, в крупных сообществах текстов на малом языке действительно больше, чем в более мелких. С другой стороны, во всех сообществах соотношение текстов остается на одном уровне: постов на малом языке примерно в 2,5 раза меньше, чем всего постов в сообществе.

Зависимость числа постов на языке от общего числа постов



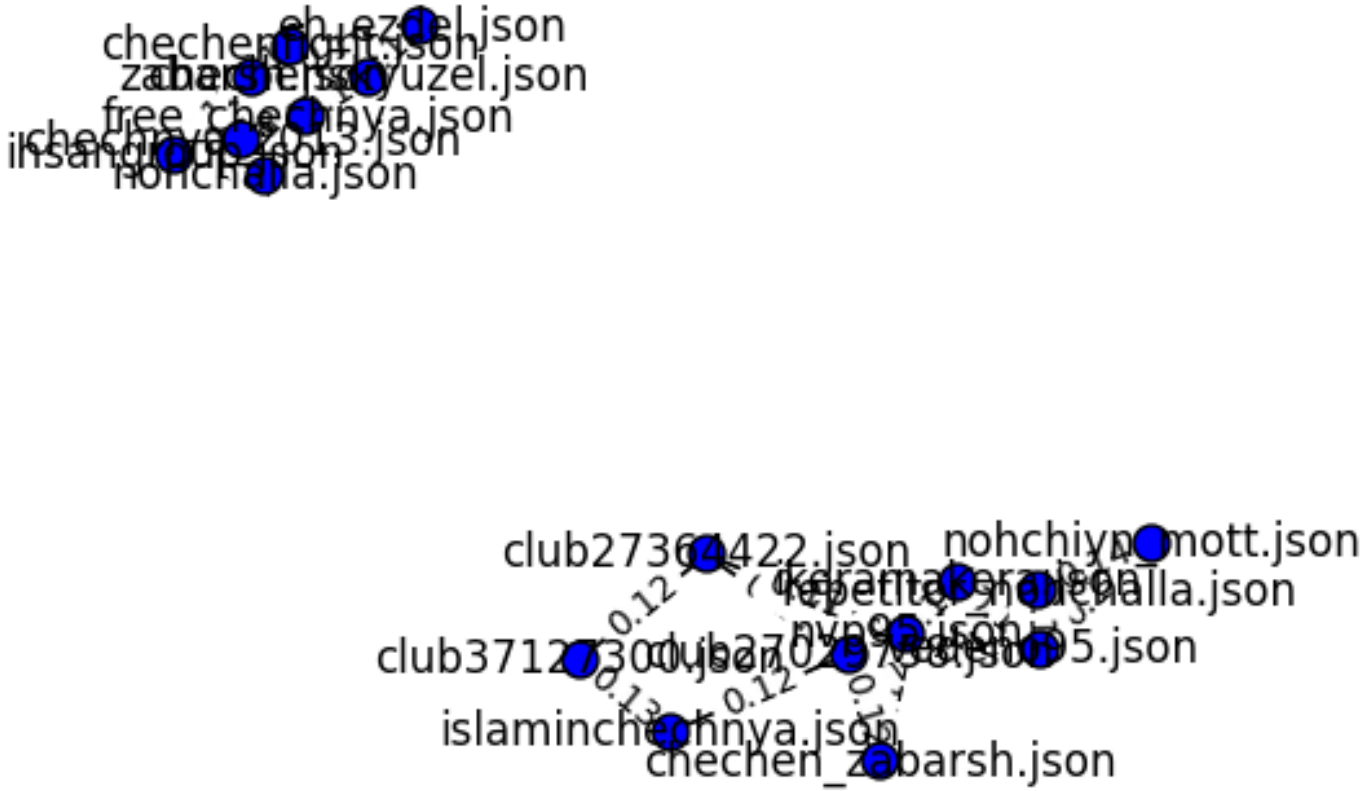
Что интересного можно узнать

Можем ли мы узнать, о чем пишут на малых языках?

В идеальном случае нам нужны были бы 30 экспертов, чтобы прочесть тексты на всех языках.

Хорошо, что у компьютерной лингвистики есть свои способы определения тематики.

Чеченский



Чеченский юмор

Карачаево-балкарский