Миноритарные языки РФ в Интернете: количественное описание и анализ данных

Выпускная квалификационная работа
студента 2-го курса магистратуры группы МТР141

Крылова И. В.
Научный руководитель: Орехов Б. В., к.ф.н, доцент

Национальный исследовательский университет
«Высшая школа экономики» (Москва)

10 июня 2016 г.

Цели и задачи

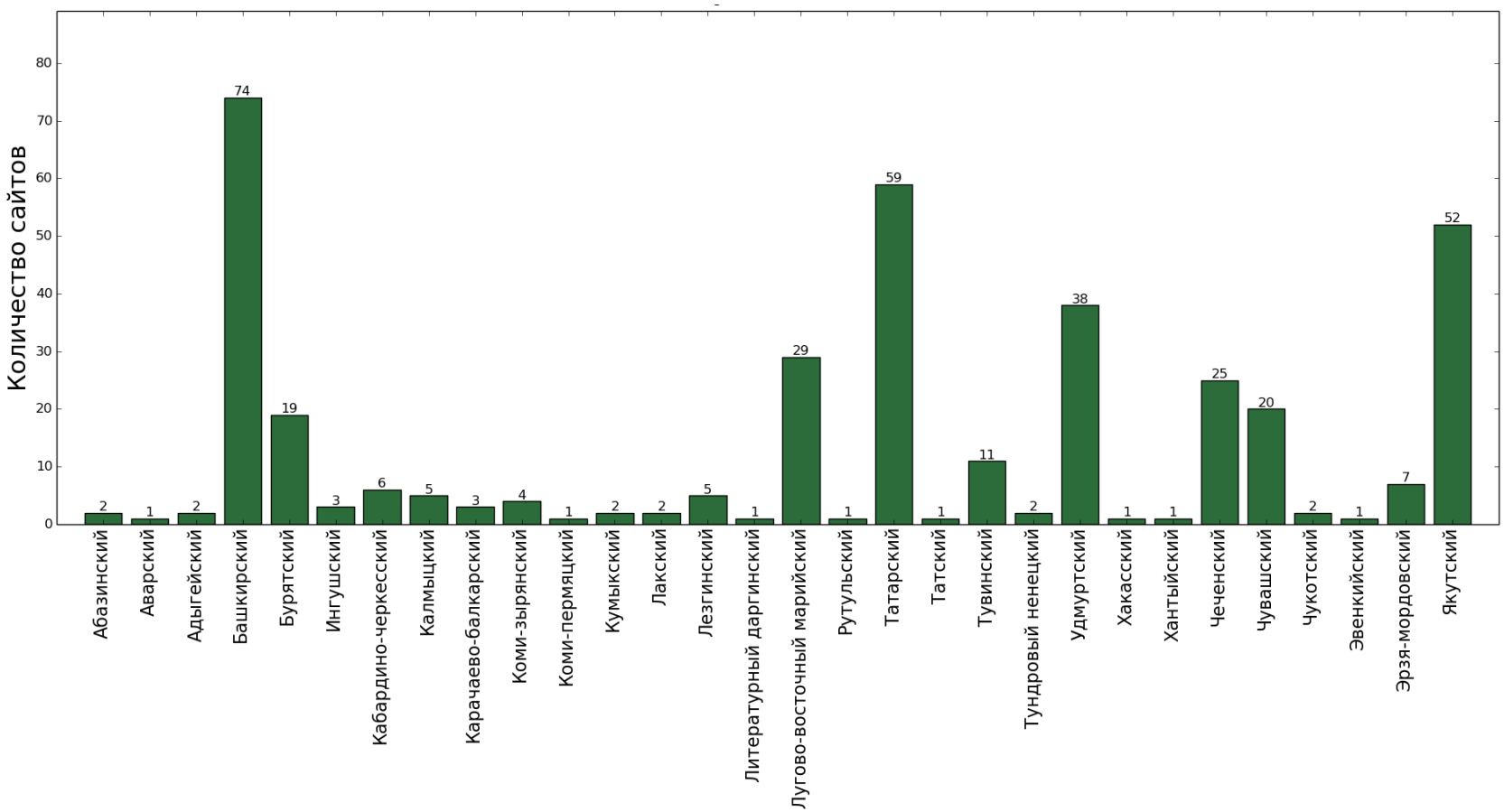
- создание количественного описания интернета миноритарных языков
- определение характеристик национальных интернетов
- поиск параметров, на основе которых можно предсказать, будет ли представлен малый язык в интернете и в каком количестве

Данные

- Интернет-коллекция
  - количество веб-страниц
  - количество токенов
  - медианное значение количества токенов на веб-страницу
  - количество веб-сайтов на малом языке
  - общее количество веб-сайтов со страницами на малом языке
- ВКонтакте-коллекция
- данные из Википедии
- число носителей языка
- данные по титульному региону распространения языка

Количественное описание

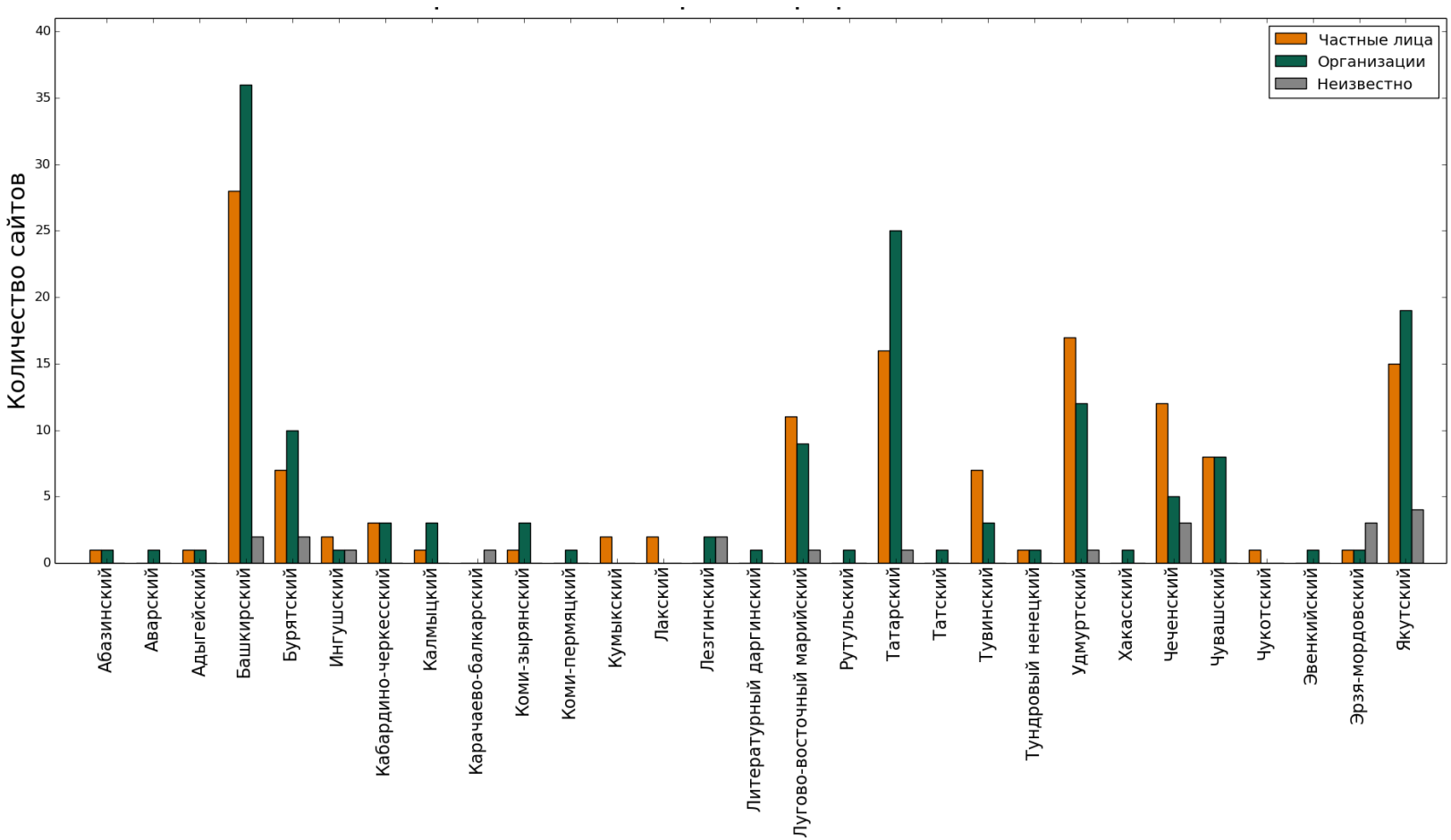
Сайты на малых языках



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

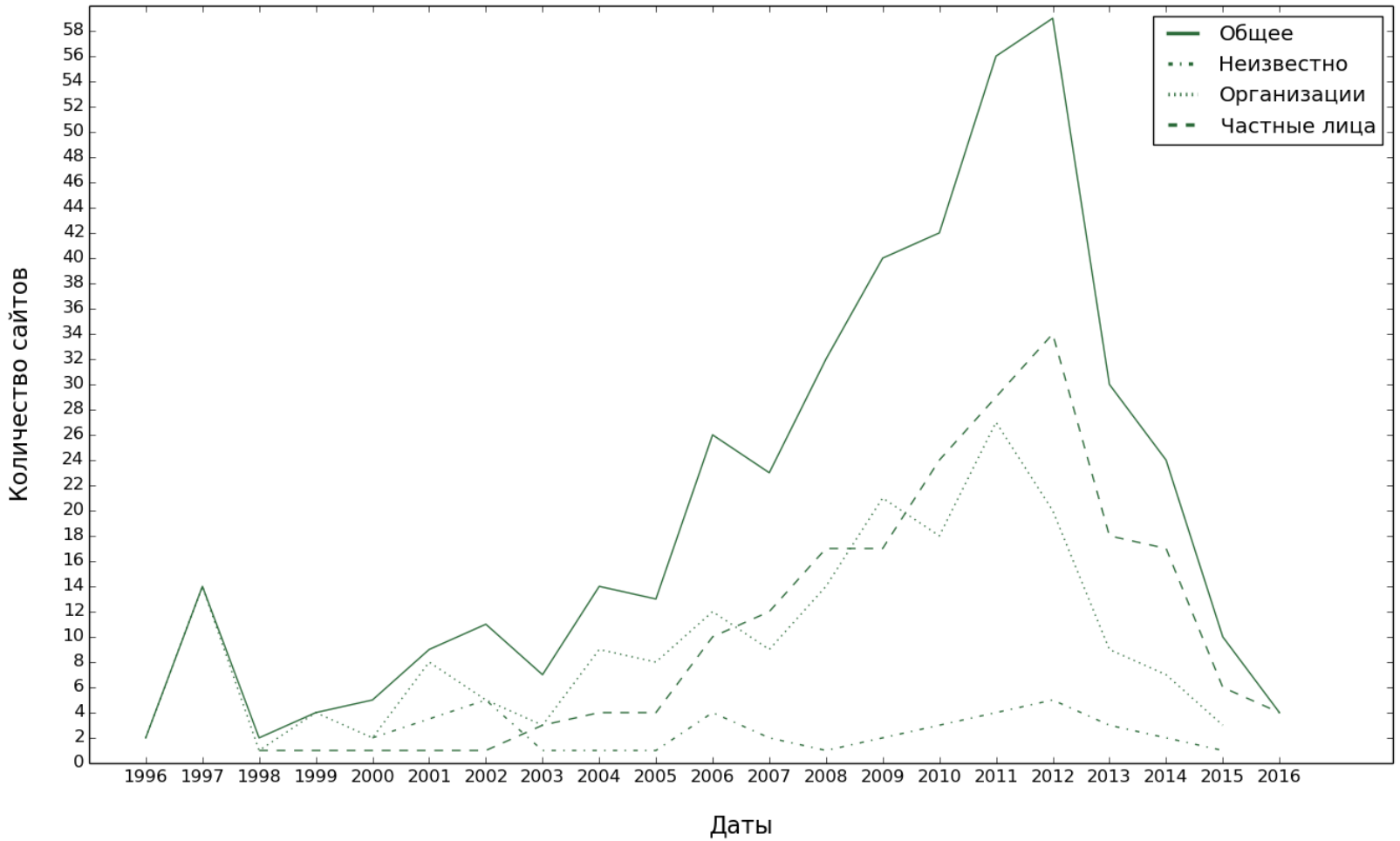
Количественное описание

Регистрационные данные: кто



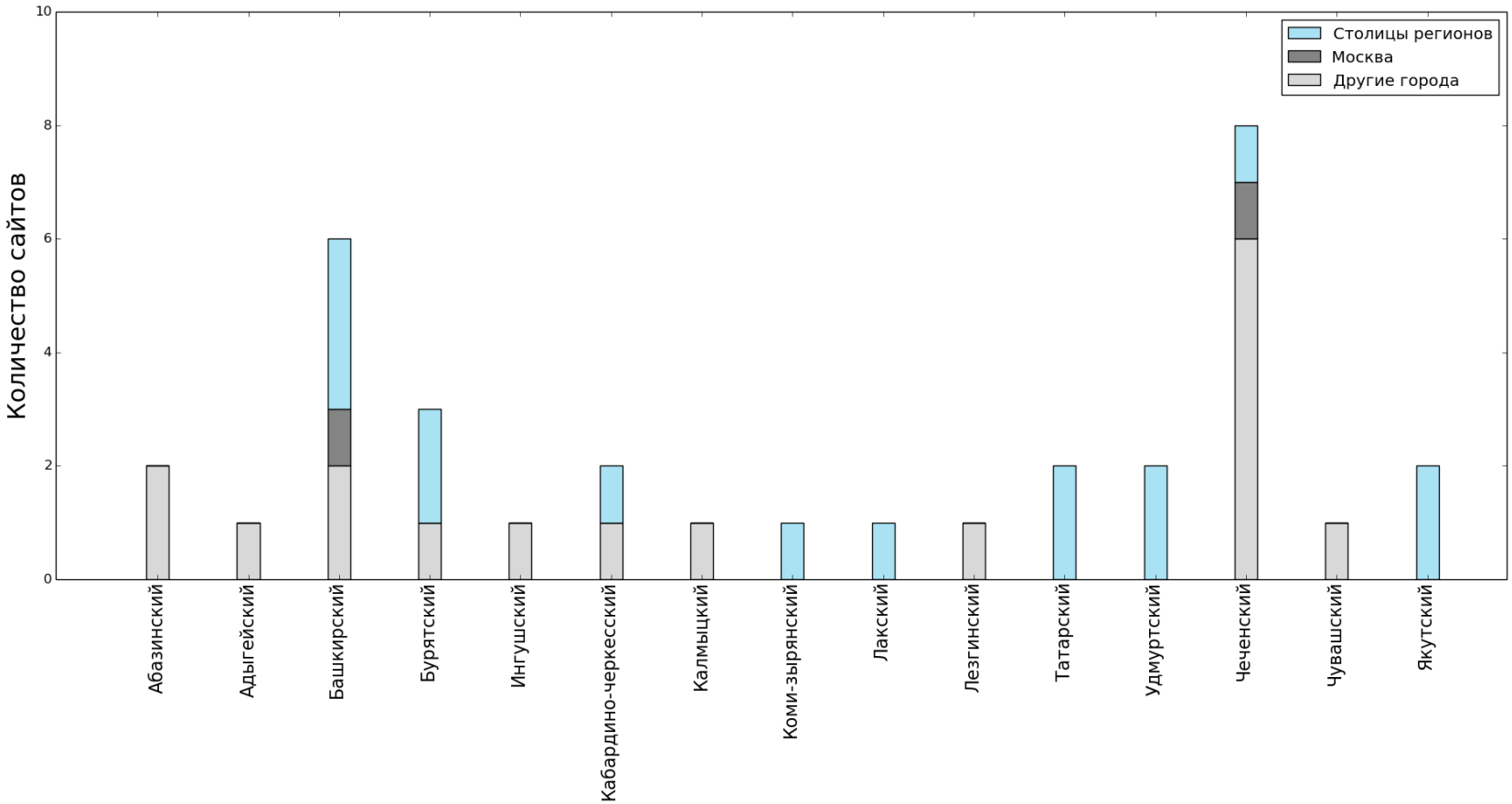
Количественное описание

Регистрационные данные: когда



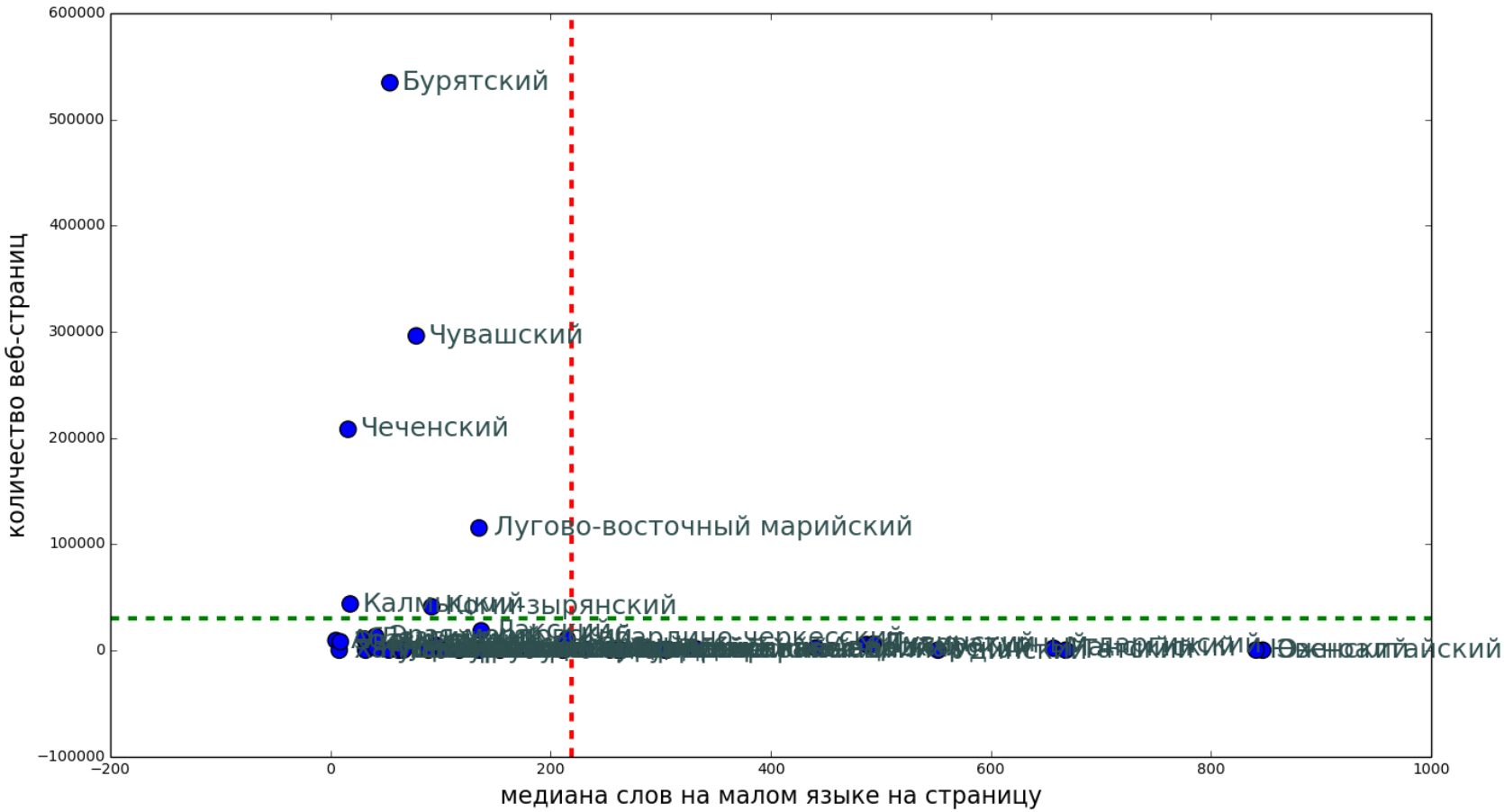
Количественное описание

Регистрационные данные: где



Количественное описание

Медианное значение токенов на страницу относительно количества веб-страниц



Методы анализа данных

- Анализ корреляций
- Регрессионный анализ
- Кластерный анализ
- Построение графа

Результаты

Анализ корреляций

Корреляция Спирмена, сильные связи (по порогу 0.7):

- между всеми онлайн-данными
  - статьи в Википедии за 2014, 2015 и 2016 года
  - сообщества и токены из ВКонтакте
  - сайты, веб-страницы, токены из Интернета

  –> одинаковый уровень интернет-представленности

- число носителей с интернет-параметрами
  - число носителей за 2002 год (меньше за 2010 год)
  - сайты, веб-страницы, токены из Интернета

  –> язык в цифровом мире развивается по своим законам

Результаты

Регрессионный анализ

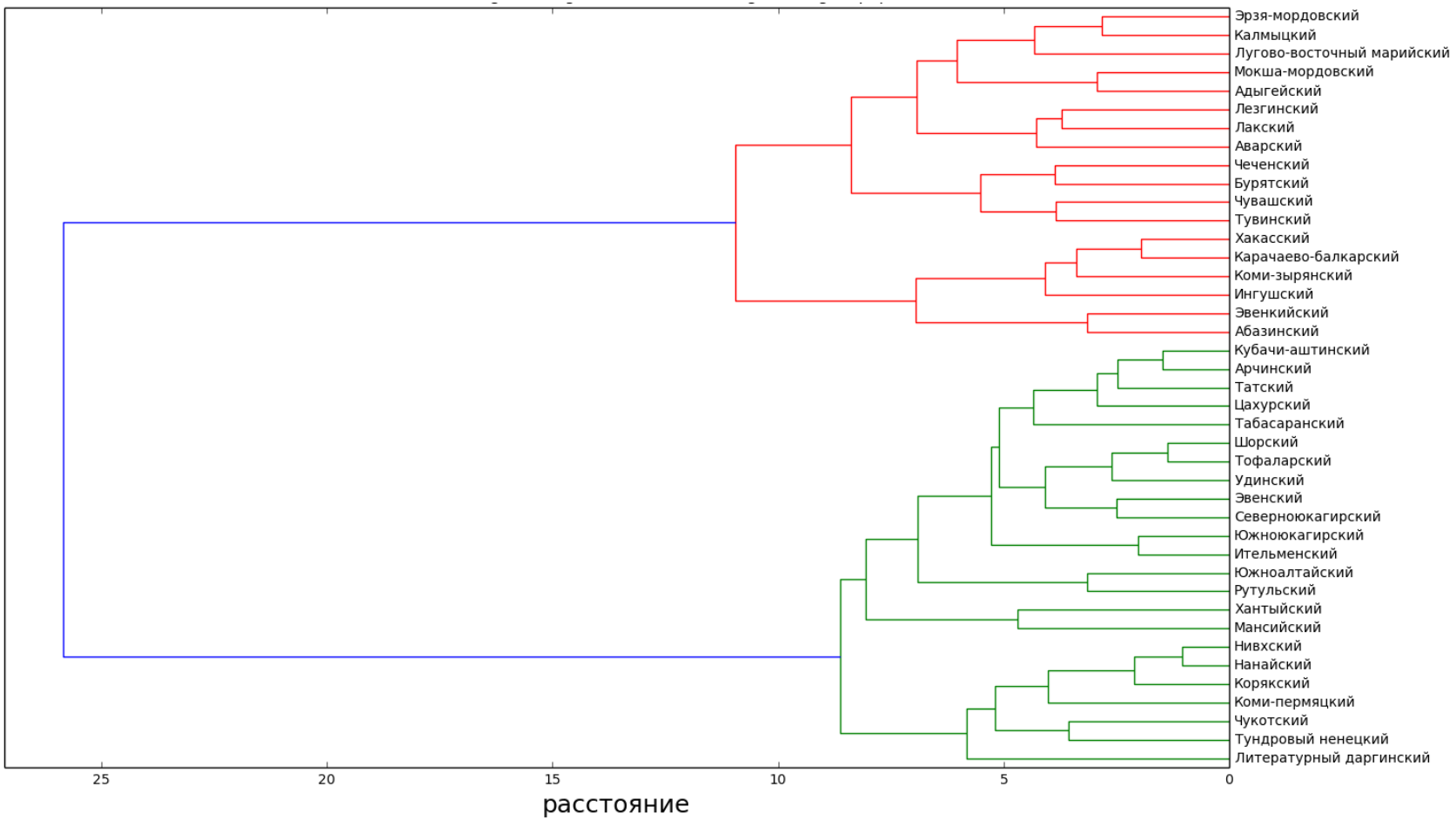
Лучшая модель: adjusted R^2 65-75%, P(F-stat) < 0.001.

Таблица: Значения коэффициентов для модели с зависимой ln(количество сайтов на малом языке)

	coef	std err	t	P>|t|
ln(кол-во статей в вики за 2016г)	0.4293	0.139	3.089	0.004
ln(кол-во токенов в ВКонтакте)	0.2263	0.142	1.595	0.120
ln(число носителей за 2010г)	0.0269	0.158	0.170	0.866
прирост населения на 2014г	0.0989	0.094	1.057	0.298
расходы региона за 2014г	−0.0424	0.102	−0.414	0.681
ln(число газет за 1996г)	0.3255	0.104	3.142	0.003

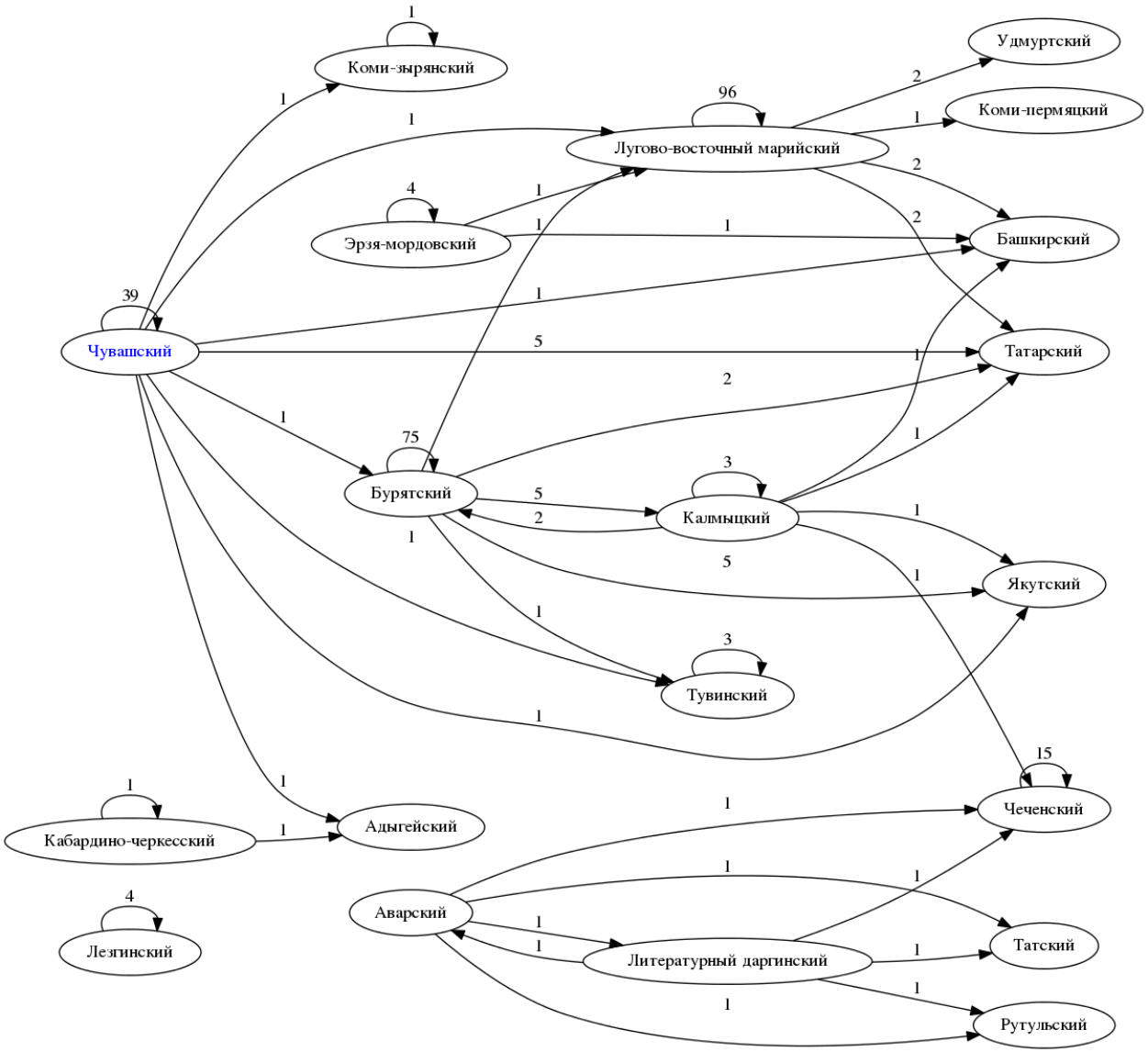
Результаты

Кластерный анализ



Результаты

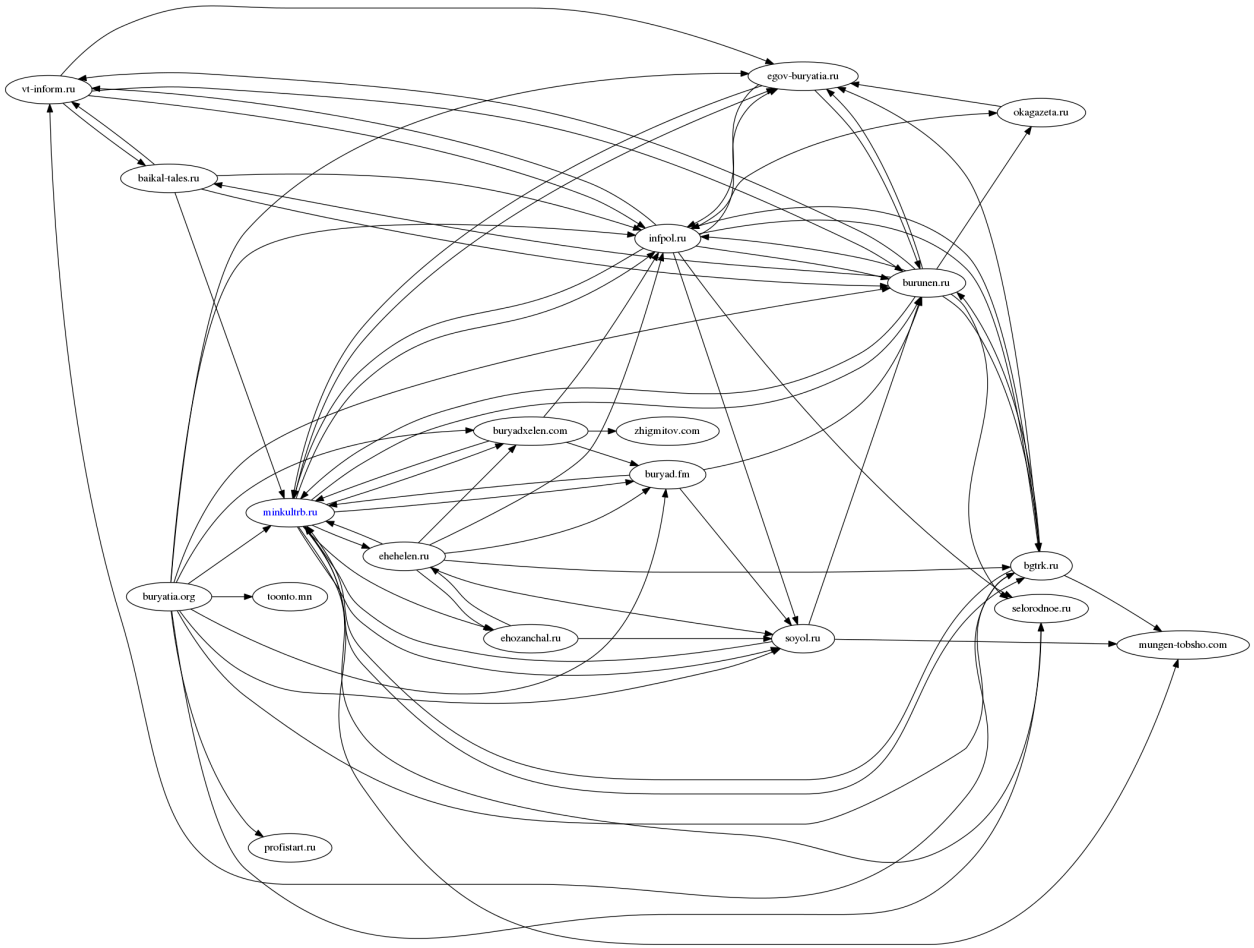
Веб-граф национальных интернетов



Характеристики:
Ассортативный
Не слабо связный

Результаты

Веб-граф бурятского национального интернета



Характеристики:
Дисассортативный
Слабо связный

Спасибо за внимание!