ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Компьютерная лингвистика»

Крылова Ирина Владимировна

Миноритарные языки РФ в Интернете: количественное описание и анализ данных

Representation of minor languages of Russia on the Internet: quantitative description and data analysis

Выпускная квалификационная работа студента 2 курса магистратуры группы МТР141

Академический руководитель
образовательной программы
канд. филологических наук, доц.
А. А. Бонч-Осмоловская

_____

« » _____ 2016 г.

Научный руководитель

канд. филологических наук, доц.
Б. В. Орехов

_____

Москва 2016

Оглавление

1. Введение **1**

2. Малые языки и их интернет **3**

3. Описание данных **10**

3.1. Сайты на малых языках 10

3.1.1. Количество сайтов на малых языках 10

3.1.2. География, персоналии и даты 12

3.2. Медианное значение токенов на страницу текста на малом языке . 29

4. Анализ данных **32**

4.1. Подготовка данных 35

4.2. Анализ корреляций 37

4.3. Регрессионный анализ данных 39

4.4. Кластерный анализ данных 45

4.5. Веб-графы интернетов малых языков 47

5. Заключение **51**

Список использованных источников и литературы **56**

Приложения **60**

A. Список больших сайтов, с которых скачивались только конкретные страницы **60**

B. Дополнительные графики распределения регистраций доменов **63**

C. Таблицы регистраций доменов по городам **65**

D. Таблицы распределения сайтов по городам на основе IP-адресов **66**

E. Код и документация **67**

F. Веб-графы национальных интернетов **67**

1. Введение

Исследование посвящено изучению представленности миноритарных языков России в Интернете. Миноритарными мы называем языки этнических меньшиств Российской Федерации, которые при этом не являются государственными в другой стране. В этом исследовании термины малый и миноритарный язык используются как синонимы, возможно, также использование сокращения м. я. в таблицах.

По нашим данным на территории Российской Федерации зафиксировано порядка 100 малых языков. Часть (15%) из этих языков бесписьменна. Для оставшихся же представляется возможным собрать текстовые интернет-коллекции, которые позволяют проводить исследования и разрабатывать инструменты автоматической обработки текста. Такие текстовые коллекции для примерно 40 языков были собраны в рамках магистерского проекта и сделали возможным текущее исследование.

Изучение представленности миноритарного языка в Интернете интересно как с чисто академической точки зрения, так и с прикладной. Сопоставление распространённости языка в сети с другими релевантными для языка параметрами позволяет понять, как функционируют миноритарные сообщества в современных условиях. С прикладной же точки зрения количественное описание интернета малых языков позволяет крупным компаниям оценить свои шансы на получение прибыли от продвижения своих сервисов и рекламы на малых языках. Объёмы интернета на чеченском, ингушском, чукотском и других языках позволяют представить объём аудитории и, соответственно, общие коммерческие перспективы освоения этой ниши.

До сих пор в основном проводились качественные исследования интернетов отдельных малых языков с акцентом на то, какие тематические сегменты занимают национальные интернеты и как важно присутствие малого языка в цифровом пространстве для сохранения и ревитализации языков под угрозой исчезновения. За редким исключением количественных исследований в области частотного анализа интернет-лексики малых языков и нескольких масштабных зарубежных проектов по сбору текстовых коллекций малых языков всего мира

и анализу цифровой жизни всех когда-либо существовавших языков (включая уже мёртвые и находящиеся сейчас под угрозой исчезновения).

Целью исследования является создание количественного описания интернета миноритарных языков (далее также национального интернета), определение характеристик национальных интернетов и поиск параметров, на основе которых можно предсказать, в каком количестве и будет ли представлен малый язык в интернете. К исследованию привлечены разные внешние данные о языке: число носителей языка, данные по региону распространения языка (экономическое благополучие, интернетизированность и другие), данные о представленности языка в Википедии (число статей, активных участников и администраторов языковых разделов), данные по малому языку из социальной сети Вконтакте. А также данные по собранным интернет-коллекциям: количество сайтов на малом языке (как показатель самостоятельности языка в интернете), количество веб-страниц, токенов, медианное значение количества токенов на веб-страницу (как показатель устойчивости/закреплённости языка в Интернете).

Такое количественное описание для многих языков будет создано впервые. Используя его и разные методы анализа данных такие как анализ корреляций, многофакторная линейная регрессия и кластерный анализ мы попробуем определить, на основе каких параметров возможно предсказать представленность языка в интернете. Для определения характеристик национальных интернетов прибегнем к построению веб-графа.

В первой главе работы описаны существующие исследования интернета малых языков, а также упомянуты основные социолингвистические теоретические работы по этой теме, во второй главе представлено количественное описание некоторых важных параметров представленности малого языка в интернете, а также проведён анализ регистрационных данных для сайтов на малом языке. В третьей главе описан анализ имеющихся онлайн и оффлайн данных по малым языкам.

2. Малые языки и их интернет

Изучению миноритарных языков посвящено немалое количество работ. Исследователей привлекают как лингвистические и грамматические особенности таких языков, так и социальные или, вернее, социолингвистические аспекты существования и поддержания малых языков в условиях доминирования главного государственного языка. Пожалуй, одно из самых полных и разносторонних описаний миноритарных языков на территории Российской Федерации представлено в книге Б. Комри (Comrie 1981). Б. Комри рассматривает языки с позиций генетической, типологической, социологической классификаций, учитывает влияние русского языка и историческое прошлое языков. Книга представляет собой обширное теоретическое описание языков прежнего СССР и включает описание лингвистических (фонологических, морфологических, синтаксических) особенностей языков по языковым семьям.

Стоит упомянуть и два более свежих издания, посвящённых малым языкам России. В 2002 году под редакцией В. П. Нерознака была издана книга «Языки народов России» (*Языки народов России. Красная книга* 2002). В книге собраны истории исследований практически всех языков народов России с лингвистической и социолингвистической точек зрения, а также приведены сведения об их носителях и даны рекомендации по сохранению и развитию языка. В 2012 году совместными усилиями финских и русских исследователей была написана книга «Как и зачем сохранять языки?» (Замятин, Пасанен, Саарикиви 2012). Акцент в первую очередь делается на финно-угорских языках, и в отличие от первых двух масштабных теоретических изданий эта книга скорее адресована носителям малых языков, чем исследователям.

В XXI веке мы живём в парадигме «всё, что нельзя найти в интернете (с помощью поисковых машин), не существует». Исследователи придумывают специальные методики сбора сбалансированных обширных текстовых коллекций из интернета (Baroni [и др.] 2009), изучают особенности цифровой жизни языка, опираясь на имеющиеся тексты и сопоставляя эти данные с оффлайн-наблюдениями теоретических исследований. Всё это в полной мере должно относится и к миноритарным языкам, однако на практике немногочисленные корпуса ма-

лых языков пополняются в основном за счёт вручную перепечатанных текстов из экспедиций.

Исследователи и СМИ отмечают особый вклад современных технологий в благополучие малых языков. В статье (Das 2013) утверждается, что интернет даёт новый виток для развития малых языков, предоставляя возможность носителям языка говорить друг с другом несмотря на разделяющие их расстояния. Для исследователей сферы обработки языка, таким образом, представляется возможность использовать новый источник текстов малых языков.

Среди исследований представленности и особенностей поведения малых языков в интернете мы различаем отечественные и зарубежные, качественные и количественные. Большая часть известных нам отечественных работ про малые языки имеет качественных характер. Как правило такие работы посвящены какому-то одному языку и включают ссылки на конкретные сообщества и сайты с текстами на этом языке, но не полное количественное описание представленности языка в интернете.

Известные нам работы Д. М. Сахарных и К. Пишлёгера про удмуртский интернет носят скорее качественный характер. В кратких интервью Д. М. Сахарных (Сахарных 2006; Сахарных 2008) рассказывает о немногочисленности сайтов на удмуртском и о том, что мешает развитию удмуртского интернета. В качестве основных проблем Д. М. Сахарных отмечает отношение удмуртоговорящего населения к интернету как к библиотеке (т.е. к тому, где можно получить информацию, а не создавать свой контент), технические и бюджетные сложности (мало компьютеров, низкая интернетизация населения, отсутствие удмуртских шрифтов и раскладок), а также незаинтересованность лингвистов-удмуртов в популяризации языка и как следствие отсутствие легкодоступных материалов по изучению удмуртского языка. В статье 2008-го года (Сахарных 2008) Д. М. Сахарных говорит о позитивных сдвигах в пространстве удмуртского интернета и благотворном влиянии появляющихся социальных сетей на формирование отношения этноса к интернету как к пространству, в котором можно жить, творить и размещать свои тексты. «В ноябре 2007 года в социальной сети Вконтакте.ру была открыта коммуникационная группа «Удмуртлык», с помо-

щью которой удалось в очень сжатые сроки (в течение нескольких месяцев) привлечь в интернет, ввести в круг тематического общения и общения на национальном языке такое количество пользователей (студентов, молодых учёных и др.), которое до сих пор не могло быть аккумулировано в рамках ни одного из существовавших до этого удмуртских национальных интернет-проектов» (Сахарных 2008).

На декабрь 2013-го года количество сообществ на удмуртском ВКонтакте достигало уже сотни (Пишлёгер 2014). В своих статьях (Пишлёгер 2013a; Пишлёгер 2013b; Пишлёгер 2014) К. Пишлёгер исследует поведение удмуртов в социальных сетях и описывает влияние русского на современный удмуртский язык, которое выражается в таких явлениях как переключение кодов («суропожо»), смешение языков, стилей и диалектов. В (Пишлёгер 2014) утверждается, что такое смешение характерно не только для переписки в социальных сетях, но и для повседневного устного общения. Это подтверждается словами респондентов, которые утверждают, что на чистом удмуртском говорят только иностранцы, в то время как сами удмурты часто говорят на смеси русского с удмуртским. По мнению К. Пишлёгера хорошо даже такое развитие и существование языка, а академическая пуристическая критика может носить только разрушительный характер.

В качестве примеров количественных исследований малых языков можно привести статьи (Орехов, Галлямов 2013), (Орехов, Решетников 2014) и (Орехов, Махмутов 2014). Б. В. Орехов и А. А. Галлямов разбивают башкирские сайты по тематике на медийные (ориентированные на печатные СМИ и перенесённые в веб) и все остальные (изначально веб-ориентированные), а также проводят сравнительный лексикочастотный анализ полученных текстов с 30 сайтов на башкирском. Исследователи отмечают ожидаемое своеобразие башкирской интернет-лексики в сравнении с частотными словарями на основе печатных научных и художественных текстов, а также официально-деловой акцент, обусловленный преобладанием публикуемых в этой тематике документов. В работе Б. В. Орехова и А. А. Галлямова подчёркивается, что современный Башнет ещё «не может быть рабочим инструментом для того, в чём давно пре-

успел Рунет: в нём нельзя искать телефоны, товары, оставлять комментарии, в нём мало сайтов с возможностью регистрации (а значит, привлекающих постоянную аудиторию пользователей)» (Орехов, Галлямов 2013).

В статье (Орехов, Махмутов 2014) сформулированы некоторые научные подходы к изучению этнических сегментов интернета. В частности там приводится классификация сайтов на малых языках по соотношению количества текстов на миноритарном и русском языках, авторы различают номинальную (минимальную), второстепенную, паритетную (равный объём с другим языком) и доминирующую степень присутствия малого языка на сайте. Исследователи рассматривают присутствие сайтов на национальных государственных языках РФ в официальном, познавательном и социальном секторе, а также приводят данные по наличию национального языка на официальных ресурсах республик РФ. Для рассмотренных (~ 30) языков в большинстве своём заполнен познавательный сектор и в меньшей степени два других, «на большинстве официальных ресурсов республик РФ миноритарные языки отсутствуют вообще, на некоторых интерфейс на национальном языке находится на стадии разработки или имеет номинальный характер». Б. В. Орехов и З. А. Махмутов утверждают, что для того чтобы национальный интернет стал «инструментом лингвокультурной экологии» необходимо сделать его полезным в повседневной жизни пользователей, а именно на малом языке должны быть созданы ресурсы, включающие услуги электронной почты, новости, содержащие знания, предоставляющие площадки для общения и самовыражения, а также возможности приобретения товаров через интернет-магазины, игры и развлечения. На момент написания статьи наиболее развитыми национальными интернетами были признаны башкирский и татарский, что коррелирует и с численностью этих этносов.

В этом же сборнике отдельно был рассмотрен татарский интернет (Махмутов 2014). Исследователь отмечает, что татарский интернет представляет собой развитую сегментарную систему, среди сайтов на татарском есть и официальные сайты, и музыкальные, и сайты знакомств. Однако, доля татарского на этих сайтах значительно уступает русскому, что говорит о «нарушении лингвокуль-

турной вертикальной трансмиссии среди татар и свидетельствуют о все более возрастающей коммуникативной функции русского языка внутри этнической группы».

Это отнюдь не исчерпывающий список работ о малых языках и их интернете, мы привели здесь лишь часть из них про наиболее освещённые в научной литературе малые языки РФ и позволяющие представить общую картину положения миноритарных языков во всемирной паутине.

Для проведения количественных исследований по интернету малых языков прежде всего необходимо собрать текстовые коллекции и корпуса на этих языках. Наиболее простым решением кажется выбор Википедии в качестве такого источника текстов. Однако, во-первых, тексты Википедии прежде всего одной и той же энциклопедической направленности и не дают полного представления о статусе присутствия малого языка в интернете, хотя вполне вероятно, что при отсутствии Википедии на малом языке, тексты на этом языке будут отсутствовать и на других ресурсах сети. Во-вторых, в (Орехов, Решетников 2014) приведены достаточно убедительные доводы в пользу непригодности Википедии как адекватной коллекции текстов на малых языках. В статье рассматриваются наиболее представленные в интернет-энциклопедии башкирский и татарский языки, а также мордовские, марийские и два языка коми. Как показал анализ частотных списков словоформ, составленных на основе википедийных текстов, во всех языках преобладают автоматически сгенерированные статьи («ботозаливки») небольшого объёма какой-нибудь одной тематической направленности.

Большая часть корпусов малых языков собрана в полуавтоматическом (т.е. ссылки для таких корпусов были получены вручную, после чего были скачаны и обработаны тексты) или ручном режиме (перепечатанные тексты из полевых экспедиций). Например, сотрудниками и аспирантами Института этнологии и антропологии РАН собраны маленькие корпуса на ненецком, телеутском, шорском и эвенкийском (*Корпусы ИЭА РАН* б.г.). Тексты, включённые в корпуса, собраны в лингвистических экспедициях, перепечатаны лингвистом и размечены морфологическим аннотатором. Стоит упомянуть и несколько аналогичных проектов с текстами на кетском, селькупском, эвенкийском (*Малые*

языки Сибири: наше культурное наследие б.г.) и текстами на арчинском, хиналугском, нганасанском, энецком и тофалараском (*Центр мультимедийных языковых архивов LangueDOC* б.г.). В качестве примеров полуавтоматических корпусов, часть текстов для которых собрана в интернете, можно привести корпуса бурятского (*Бурятский корпус* 2012), удмуртского (*Корпус удмуртского языка* 2015), лезгинского (*Корпус лезгинского литературного языка* 2011), татарского (*Татарский национальный корпус «Туган тел»* б.г.) и калмыцкого (*Калмыцкий корпус* б.г.).

Самым масштабным проектом по автоматизированному сбору текстовых интернет-коллекций является An Crúbadán. Проект был запущен в 2007 году К. Сканеллом (Scannell 2007) и на данный момент предоставляет информацию по более чем двум тысячам миноритарным языкам и другим языкам с нехваткой лингвистических инструментов обработки. Несмотря на то, что собранные текстовые коллекции недоступны для скачивания, веб-сайт проекта предоставляет много полезной информации, включая словесные униграммы и биграммы, буквенные триграммы, списки веб-страниц, на которых можно найти тексты на конкретном языке, количество найденных документов и слов, а также ссылки на другую справочную информацию по малому языку и ссылки на существующие спеллчекеры малых языков. К. Сканелл ведёт и другой проект, ориентированный скорее на веб 2.0 — (*Indigenous Tweets* б.г.) и (*Indigenous Blogs* б.г.). Проект призван помочь заинтересованным в поддержании родного языка пользователям найти друг друга. В нём собраны ссылки на различные твиттер-аккаунты и блоги на малых языках с указанием количества пользователей и твиттов для твиттера и количества блогов, постов и слов для блогов.

Ввиду отсутствия доступных текстовых коллекций по всем интересующим нас миноритарным языкам РФ, в рамках проекта «Языки России» была разработана и опробована методика сбора цифровых коллекций текстов в целом напоминающая «seed words» из (Baroni [и др.] 2009). Методика подробно описана в (Zaydelman [и др.] в печати), а результирующие коллекции текстов по каждому из обработанных языков доступны на сайте проекта (*Языки России* б.г.).

Среди других проектов и исследований, непосредственно относящихся к ма-

лым языкам и текстам на них, можно упомянуть статью Т. Архангельского, О. Беляева и А. Выдрина (Arkhangeskiy, Belyaev, Vydrin 2012) про автоматический морфологический аннотатор для миноритарных языков и бакалаврскую работу студента Высшей школы экономики про инструмент определения языка (Самойленко 2015).

Многие миноритарные языки входят в категорию языков под угрозой исчезновения (англ. Endangered Languages). Подобные классификации составляются на основе различных офлайн-данных доступных о языке, таких как количество носителей языка, статус языка (официальный/государственный), сферы использования языка, возрастной состав носителей, различных нелингвистических факторов (например, экономических) и некоторых других (*Endangered Languages* б.г.). В 2013-м году А. Корнаи опубликовал исследование (Kornai 2013), в котором адаптировал традиционную схему EGIDS (англ. Expanded Graded Intergenerational Disruption Scale) к «цифровой жизни» языка. Он классифицировал 8 426 языка на процветающие (thriving), жизнеспособные (vital), ушедшие в наследие/неразвивающиеся (heritage) и неподвижные/мёртвые (still)), используя данные с всевозможных доступных интернет-ресурсов, и пришёл к выводу, что лишь 4% всех языков имеют цифровое будущее.

Очевидно, что в научном мире существует интерес в исследовании малых языков, сборе для них корпусов и разрабатывании компьютерно-лингвистических инструментов для таких языков. Исследователи и СМИ возглагают надежды на то, что интернет будет способствовать сохранению и ревитализации миноритарных языков. Некоторые из них особо отмечают благотворное влияние социальных сетей как инструмента, провоцирующего активное поведение носителей малых языков, и утверждают, что несмотря на то, что этносы общаются на смеси русского и национальных языков, такое общение всё равно способствует сохранению языка. Однако, существующие количественные исследования интернета малых языков свидетельствуют о том, что на большинстве сайтов национального интернета тексты на миноритарных языках присутствуют лишь номинально, что лексика интернета малых языков скорее официально-деловая, чем разговорная, и что для фомирования полноценного национально-

го интернета полезного для пользователя необходимо заполнение коммерческих и развлекательных ниш с интерфейсами на малых языках. Отметим, что при этом недавнее обширное исследование А. Корнаи заставляет усомниться в том, что большинство мировых языков смогут пересечь этот барьер и обрести полноценное цифровое существование.

В своём исследовании на основе собранных текстовых коллекций из интернета мы попытаемся оценить присутствие миноритарных языков в интернете в терминах веб-старниц и токенов. Дать представление об особенностях национальных интернетов малых языков и попытаемся определить, связана ли цифровое существование языка с какими-либо офлайн-параметрами.

3. Описание данных

3.1. Сайты на малых языках

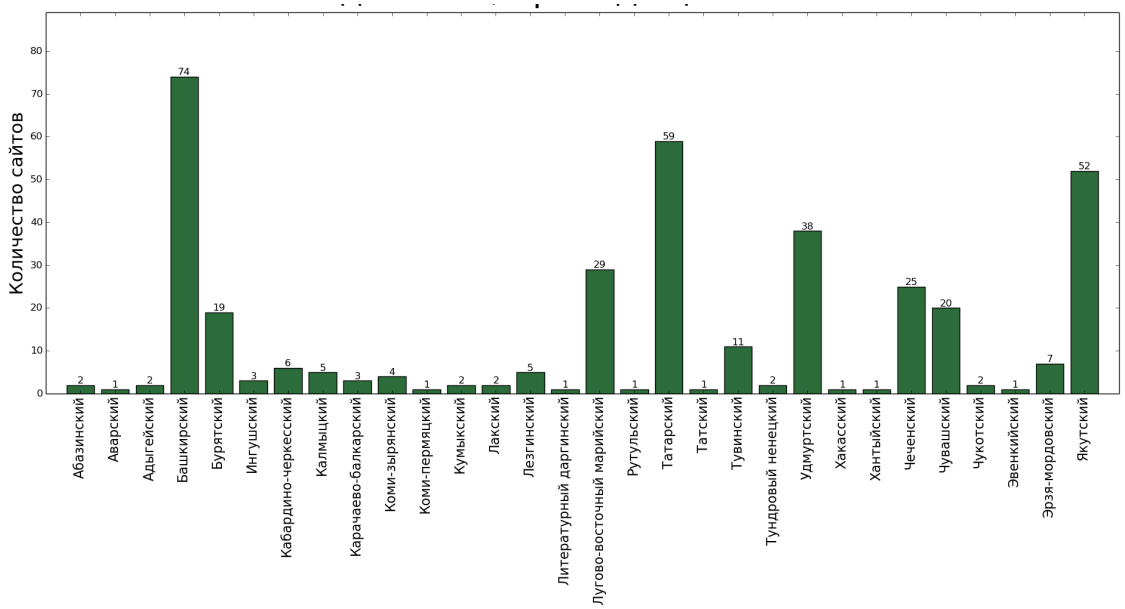
Из всех данных коллекций текстов на малом языке особый интерес для нас представляют сайты, большинство страниц которых написаны на малом языке (далее – сайты на малых языках). При формировании списка ссылок для скачивания текстов такими сайтами считались сайты с не менее 30 страницами на малом языке. Исключение составляют общеизвестные большие сайты такие как stihi.ru и youtube.com, которые не считались сайтами на малом языке вне зависимости от количества найденных страниц. Список всех таких сайтов, дополненный менее известными большими сайтами со страницами на разных языках, приведён в Приложении А.

3.1.1. Количество сайтов на малых языках

Из 49 языков, для которых выполнялся поиск информации, только для 30 удалось найти хотя бы один сайт, преимущественно написанный на малом языке. На рис. 1 представлены все такие языки. Для читаемости мы исключили из графика языки, для которых не было найдено ни одного сайта на малом языке, а именно: алеутский, арчинский, ительменский, корякский, кубачи-аштинский, мансийский, мокша-мордовский, нанайский, нивхский, ногайский, се-

верноюкагирский, табасаранский, тофаларский, удинский, хантыйский, цахурский, шорский, эвенский, южноалтайский и южноюкагирский языки.

Рис. 1: Распределение по количеству сайтов на малых языках



По рисунку 1 видно, что наиболее представленными в интернете являются башкирский, татарский и якутский языки. За ними следуют удмуртский, лугово-восточный марийский, чеченский, чувашский и бурятский. Полученные данные в общем-то согласуются и с чутьём крупных коммерческих компаний – например, у «Яндекса» создана бета-версия «Яндекс.Переводчика» для башкирского, татарского и удмуртского языков (*Яндекс.Переводчик* б.г.). Можно предположить, что в скором времени появится и онлайн-переводчик с якутского. Впрочем, кажется, якутский ускользает не только из фокуса коммерческих компаний, но и от внимания исследователей национальных интернетов. Для татарского и башкирского интернета придуманы названия по аналогии с «Рунетом» – «Башнет» и «Татнет», для якутского интернета такого самоназвания нет. Якутский уступает башкирскому и татарскому также по числу носителей (всего 450 тыс. человек по данным переписи населения за 2010-ый год против 1 млн башкиров и 4 млн татар) и по количеству статей в интернет-энциклопедии Википедии: по данным на апрель 2016-го года количество статей в якутской википедии в 3 и в 6 раз меньше количества статей, чем в башкирской и татарской

википедиях, соответственно. Не нашли мы и отдельных исследований, посвящённых качественному или количественному описанию якутского интернета.

Отметим, что из рисунка выше были исключены сайты, отмеченные по нашей методике как «омонимичные». Без дополнительной консультации с экспертами мы затрудняемся отнести такие сайты к одному из языков, поэтому «омонимичные» сайты участвуют только в следующем подразделе, где приводится общая статистика по национальным интернетам без привязки к какому-то конкретному языку. К сожалению, наших мощностей и временных ресурсов не хватило на получение текстов по четырём самым крупным по числу сайтов языкам (башкирском, татарском, якутском, удмуртском), поэтому для них в нашем распоряжении сейчас только найденные (а не выкачанные) в сети данные, которые рассматриваются только в этом и следующем подразделах. В дальнейшем анализе ни «омонимичные сайты», ни наиболее представленные в интернете языки не участвуют.

3.1.2. География, персоналии и даты

Ценной является не только информация о наличии и количестве сайтов на малых языках, но и то, кто, когда и где создаёт площадки для размещения информации на них. С помощью python-библиотеки, консольной утилиты и вспомогательных сайтов, реализующих программный интерфейс к whois-записям можно узнать регистрационные данные владельцов доменных имён.

Сразу оговоримся, что понятия «сайт» и «домен» приравнивать не совсем справедливо, т. к. домен – это элемент адреса документа, обозначающий символьное имя узла, с которого данный документ получен; а сайт – это сама совокупность документов, для адресов которых совпадает домен. С помощью поисковой машины «Яндекса» мы, конечно, искали сайты (т. е. сами тексты и документы), но сайт в интернете обязательно располагается по какому-то адресу. Понятно, что выборка сайтов «Яндекса» на страницах поисковой выдачи может быть в чём-то неполной (не все сайты индексируются, какие-то могут быть запрещены для поисковых роботов), а в чём-то избыточной (из базы индексации заблокированный или удалённый сайт исчезает не мгновенно),

но исследование даркнета крайне трудоёмкая задача и не входит в рамки нашей работы, а информация о сайтах с текстами на малом языке, пусть и удалённых, не кажется нам избыточной.

Мы также отдаём себе отчёт в том, что понятие «регистрация домена» не то-же самое, что и «регистрация сайта». И что некоторые сайты могли «располагаться» на другом домене, а потом содержательно (т. е. сами тексты) быть перенесены на новый. Но второе понятие не имеет строгого определения и задача поиска даты создания сайта кажется крайне труднореализуемой, поэтому в данном подразделе мы будем оперировать лишь первым понятием и иметь в виду, что имеющиеся данные не предоставляют возможности делать строгие утверждения о возрасте национальных интернетов.

Объём и полнота регистрационной информации, которую можно получить зависит от уровня домена и конкретного whois-сервера. Для большинства сайтов whois-клиент предоставляет информацию о доменах второго уровня. Но если в случае с сайтами kalebtatar.ru, www.aznakaevo-rt.ru мы можем получить информацию о датах регистрации конкретно этих доменов и их владельцов, то в случае с сайтами aulassoki.ucoz.ru, moradimgazeta.wordpress.com и tatarca-text.narod.ru информация о доменах второго уровня (ucoz.ru, narod.ru, worpress.com) не представляет для нас особой ценности, т. к. относится к организациям, регистровавшим хостинги сайтов и блогов, а не к активистам, создавшим эти конкретные сайты.

В связи с этим полученные с помощью whois-клиента данные были перепроверены и очищены от:

- дублей (был оставлен один сайт из группы дублирующих)
  - сайтов и их аналогов с префиксом www, old, archive (chm.mari-centr.ru и www.mari-centr.ru)
  - сайтов и их форумов (freekalmykia.org и forum.freekalmykia.org)
  - сайтов и их подразделов, оформленных как домен более низкого уровня (culcentr.ykt.ru, sakhasire.ykt.ru, news.ykt.ru)

- сайтохостингов (narod.ru, ucoz.ru, wordpress.com, ucoz.net, ucoz.com, 3dn.ru, my1.ru, moy.su, h14.ru, h15.ru)
- блогохостингов (livejournal.com, blogspot.com.tr, blogspot.ru)

В наших данных также есть спорная группа сайтов, в которую входят сайты на доменах 3-го уровня, но регистрационную информацию для которых можно получить только о домене 2-го уровня. Эти сайты отличаются от третьей группы дублей тем, что оформлены как отдельные полноценные сайты, а не подразделы одного сайта. Например, это сайты в домене tuva.ru — gov.tuva.ru, ovur.tuva.ru, tigi.tuva.ru, и в домене s-vfu.ru — www.s-vfu.ru, alampa.s-vfu.ru, iolonkho.s-vfu.ru. Информация о датах в данном случае получается немного задублированной (2 – 3 одинаковые даты вместо одной), однако данные по тем, кто зарегистрировал эти домены важно сохранить (точнее то, что одна и та же группа людей создала больше одного сайта), поэтому было принято решение оставлять все сайты из этой группы.

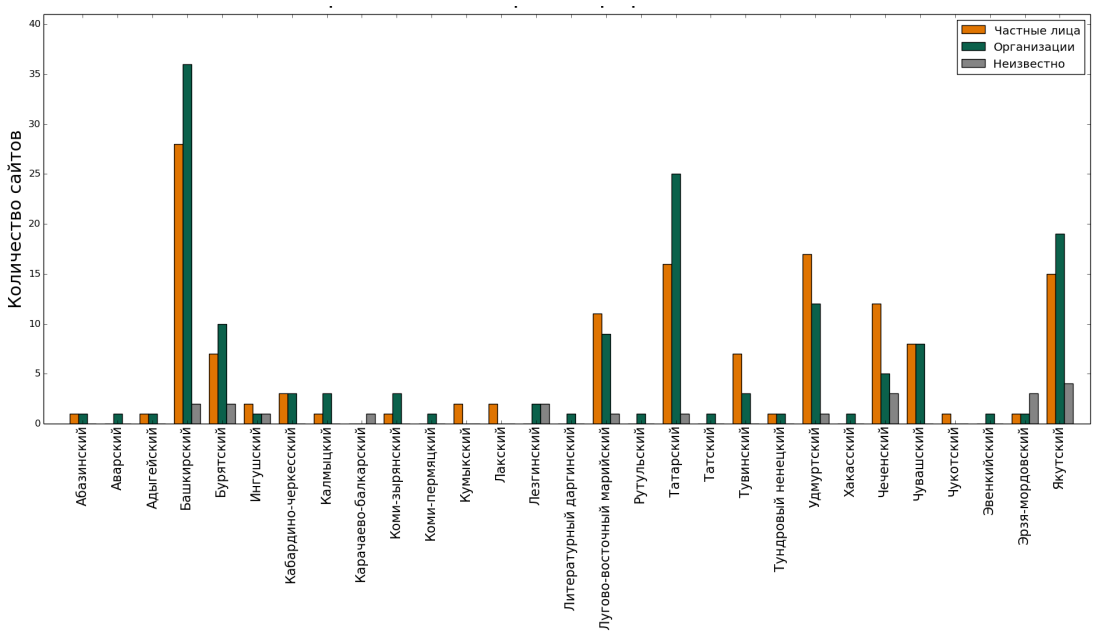
Изначально в нашем распоряжении было 528 сайтов (доменных имени) на 29 малых языках. После того как мы убрали дубли, сайтохостинги и блогохостинги, осталось 432 сайта. Для 124 из них мы затрудняемся однозначно сказать, на каком именно малом языке эти сайты написаны. Такие сайты при создании коллекций текстов мы называли омонимичными и не включали в коллекцию текстов конкретного языка. Однако, считаем правомерным использовать их в этом обзоре данных по малым языкам для создания общей картины продвижения миноритарных языков в интернете.

На рис. 2 представлена информация о владельцах сайтов[1] для каждого конкретного малого языка, для которого у нас есть данные, т. е. омонимичные сайты на графике не представлены. Мы разбили регистрировавших доменное имя на три подгруппы – «частные люди», «организации» и «неизвестно». Стоит сразу отменить, что хостинги типа PJSC "Rostelekom" мы относили в категорию «неизвестно», т. к. действительно whois-клиент нам предоставляет информацию не о том, кто владеет конкретным сайтом, а скорее о том, посредством

[1] Здесь и далее каждый раз когда говорится о регистрационных данных имеются в виду «домены», а когда говорится о текстах, расположенных по адресу, имеются в виду «сайты»

кого была оформлена регистрация сайта.

Рис. 2: Распределение по типам владельцев сайтов на малых языков для каждого языка



На графике видно, что сайтов, зарегистрированных организациями и частными лицами примерно поровну. Регистрации сайтов частными лицами преобладают для ингушского, кумыкского, лакского, лугово-восточно марийского, тувинского, удмуртского, чеченского и чукотского. Тогда как сайты на аварском, башкирском, бурятском, калмыцком, коми-зырянском, коми-пермяцком, даргинском, рутульском, татарском, татском, хакасском, эвенкийском и якутском зарегистрированы в основном организациями.

С помощью простых текстовых эвристик мы расклассифицировали организации владельцов сайтов и отнесли их к одному из следующих видов деятельности:

- СМИ – газеты, журналы, радио, телевидение
- издательства – издательские дома
- государственные – организации, в названии которых упоминаются министерства, главы республик, региональные и муниципальные администра-

ции

- образование – государственные университеты
- искусство – театры, библиотеки, культурные центры
- другое – по названию организации сложно определить, к какому виду деятельности её можно отнести

Так как данные для каждого языка немногочисленны, разбиение по типам организации для них получается очень разреженным. В таблице 1 представлена сводная информация по всем сайтам (включая омонимичные). Для полноты картины таблица дополнена данными по блогам и сайтохостингам, которые были убраны из данных по построению всех графиков в этом обзоре.

Таблица 1: Владельцы сайтов на малых языках

Частные лица			203
Организации	СМИ	46	197
	Издательства	10	
	Государственные	34	
	Образование	7	
	Искусство	8	
	Другие	92	
Блоги			9
Хостинги			54
Неизвестно			19

В таблице 2 приведены примеры распределения по типам организаций для 5 языков, для которых больше всего регистраций сайтов было сделано именно организациями. Можно отметить, что в татарском превалируют сайты, созданные СМИ, – 13 из 25 сайтов, зарегистрированных организациями, и при этом совсем нет сайтов, созданных образовательными учреждениями, издательствами и культурными центрами. В башкирском же имеются сайты всех 5 типов организаций, большая часть из них создана СМИ, государственными учреждениями

и издательскими домами. Якутский лидирует по количеству сайтов, созданных университетом (это 4 сайта, созданных в домене Якутского государственного университета, из нашей спорной группы сайтов). У удмуртского так же, как и у башкирского, заняты все слоты приведённых пяти типов организаций. Аналогичная картина наблюдается и у неприведённого в таблице чувашского. Распределение по типам организаций бурятского языка характерно также для луговосточно марийского и чеченского. Данные по остальным языкам крайне малочисленны.

Таблица 2: Распределение по типам организаций для 5 языков с большим количеством сайтов, зарегистрированных организациями

	СМИ	Издательства	Гос.	Образование	Искусство	Другие
Башкирский	9	5	8	1	1	12
Татарский	13	0	3	0	0	9
Якутский	3	0	2	4	0	10
Удмуртский	1	1	2	1	1	6
Бурятский	3	0	2	0	1	4

Из обеих таблиц, представленных выше, видно, что сайты на малых языках прежде всего создают государственные учреждения и СМИ.

На рис. 3 изображено общее распределение сайтов по годам регистрации. Для 5 из 432 сайтов нам не удалось получить дату регистрации, остальные сайты разбиты по годам регистрации и владельцам сайтов. По этому графику видно, что пик регистраций новых сайтов на малых языках приходится на 2012 год. По данным Российского рынка доменов (*Российский рынок доменов. Регистрация новых доменов* б.г.) всплеск регистраций в 2012 году характерен и для всего Рунета (см. рис. 4). Однако, при более близком рассмотрении оказывается, что большинство регистраций 2012-го года в Рунете было сделано в декабре (см. рис. 5), тогда как всплеск регистраций доменов на малых языках приходится скорее на август (см. рис. 6). Дополнительные графики для доменов верхнего уровня .РФ и .SU приведены в Приложении В. Отметим, что тенденции аналогичны домену .RU.

Рис. 3: Общее распределение сайтов по году регистрации

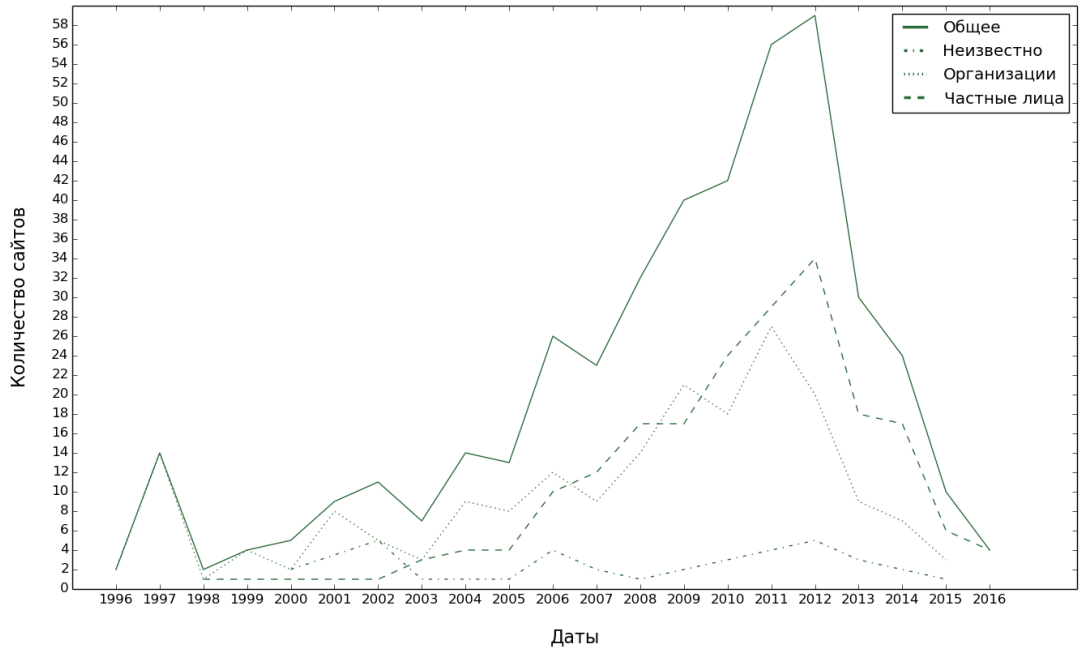


Рис. 4: Общее распределение регистраций новых доменов в зоне .RU. Скриншот от 27.04.2016 с сайта stat.nic.ru

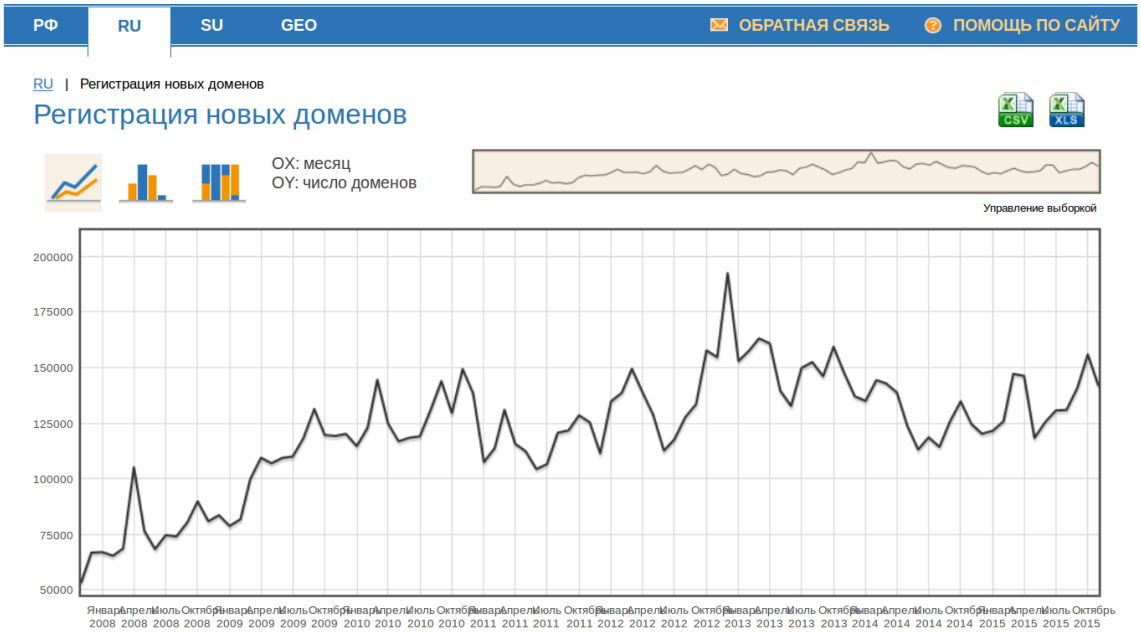
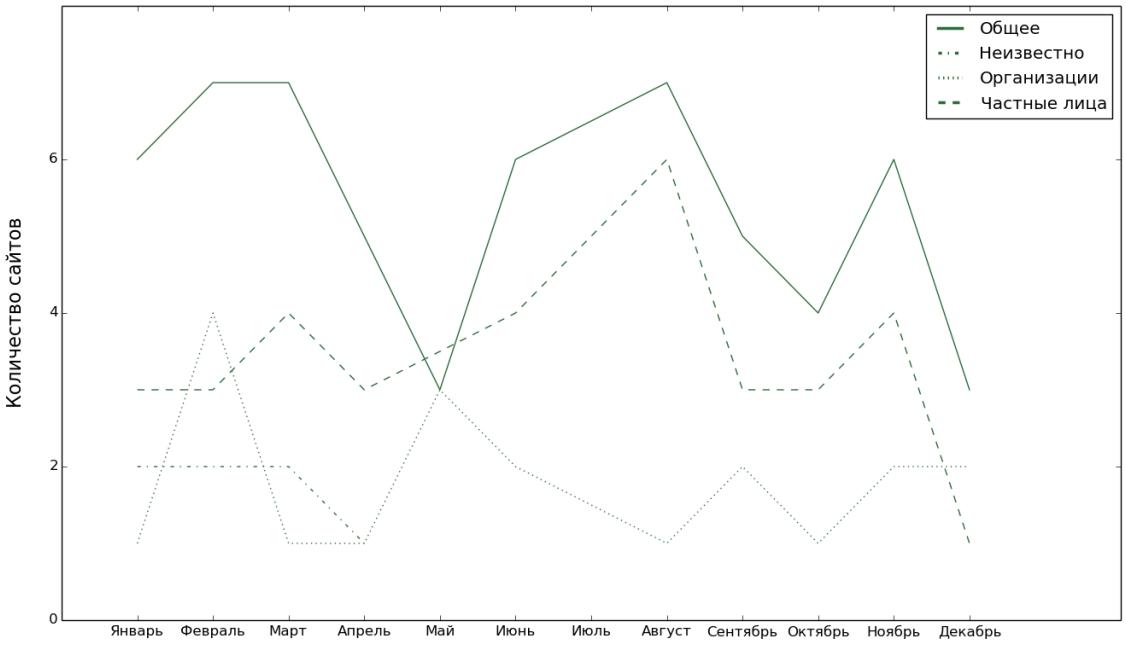


Рис. 5: Распределение регистраций новых доменов в зоне .RU за 2012 год. Скриншот от 27.04.2016 с сайта stat.nic.ru

Рис. 6: Общее распределение регистраций сайтов по месяцам 2012-го года



Если принять во внимание регистраторов, через которых оформлялись новые домены на интересующих нас языках, то для 2012-го года получится следующий упорядоченный список (см. таблицу 3):

Таблица 3: Топ-5 регистраторов 2012-го года для сайтов на малых языках

Регистратор	Количество сайтов
REGRU-RU	20
RU-CENTER-RU	10
R01-RU	9
NAUNET-RU	7
REGTIME-RU	3

С официального сайта REG-RU можно узнать, что в 2012-м году этот регистратор получил аккредитацию в ICANN (с англ. *Корпорация по управлению доменными именами и IP-адресами*), а также «осуществил планы по территориальной экспансии, увеличив количество офисов компании в два раза» (*REG.RU подводит итоги 2012 года* 2012). Второй же по количеству сайтов на малых языках регистратор активно открывал представительства в регионах РФ, например, (*В Туле приступил к работе региональный представитель RU-CENTER* 2012), и проводил, так называемый, Всероссийский доменный марафон, основной целью которого было заявлено «выявление проблем, связанных с регистрацией доменов и предоставлением услуг хостинга в регионах, проведение партнерских семинаров, обмен опытом, презентации проектов по созданию доменов верхнего уровня для городов и культурно-лингвистических сообществ России» (*Доменный марафон шагает по российским регионам* 2012). Насколько можно понять с официального сайта доменного марафона, впервые марафон был проведён в 2011-м году (*Всероссийский доменный марафон* б.г.), после чего повторен в 2012-м, о более поздних марафонах ничего не известно.

Исходя из таких данных можно предположить, что пик регистраций сайтов на малых языках в 2012-м году связан с активной экспансией регистраторов в регионы и старательным разъяснением тонкостей, связанных с владением и регистрацией доменов, регистратором RU-CENTER. На возможное влияние маркетинговых акций регистраторов в личной переписки указывает и один сетевых деятелей А. Венедюхин[2]. Однако, эксперт отмечает, что «эффекты от мар-

[2] страница А.Венедюхина в галерее видных сетевых деятелей ФРИ – http://ezhe.ru/fri/410/

кетинговых акций со снижением цены, повлекшие всплески регистрации имён, следует с большой осторожностью включать в причинно-следственные связи при анализе эволюции сайтов на малых языках, даже если что-то вдруг совпало по времени» из-за малого количества сайтов в выборке и того, что эти маркетинговые акции имеют отложенный эффект (т.е. сайт не мгновенно размещается под зарегистрированным доменным именем, для его создания необходимо время). Дальнейшее же падение регистраций новых доменов может быть связано с насыщением «рынка малых языков», закончившимися акциями от регистраторов и/или перераспределением интернет-аудитории в социальные сети.

На общем распределении сайтов по году регистрации (рис. 3) также видно, что в наши данные попало 3 сайта, созданных в 2016 году. Это обусловлено задержкой во времени между получением списков сайтов на малых языках и получением регистрационных данных. Про эти 3 экземпляра уже нельзя быть уверенными в том, что под этим доменным именем расположен сайт, написанный на малом языке.

Рассматривая графики регистраций владельцами под общим графиком, можно заметить, что до 1998 года сайты на малых языках создавались только организациями, в 2009-ом и 2011-ом годах ещё можно наблюдать пики по количеству сайтов, созданных организациями, однако, примерно с 2005-го года тенденция изменилась и сайты на малых языках стали создавать в основном частные лица. По данным Регионального сетевого информационного центра (RU-CENTER) эти тенденции характерны для всего Рунета в целом. Такие выводы можно сделать, исходя из статистики по типам владельцов за каждый конкретный год, представленной на сайте Российского рынка доменов[3]. Если в 2003-ем году доля регистраций организациями составляла 57%, то начиная с декабря 2004-го года регистрации частными лицами начинают преобладать и эта тенденция не меняется впоследствии. Разве что в брошюре 2010-го года (*Российский рынок доменов. IP-адресация, хостинг, сервисы. Итоги 2010 года* 2010) упоминается, что в начале 2011-го года «соотношение впервые начало

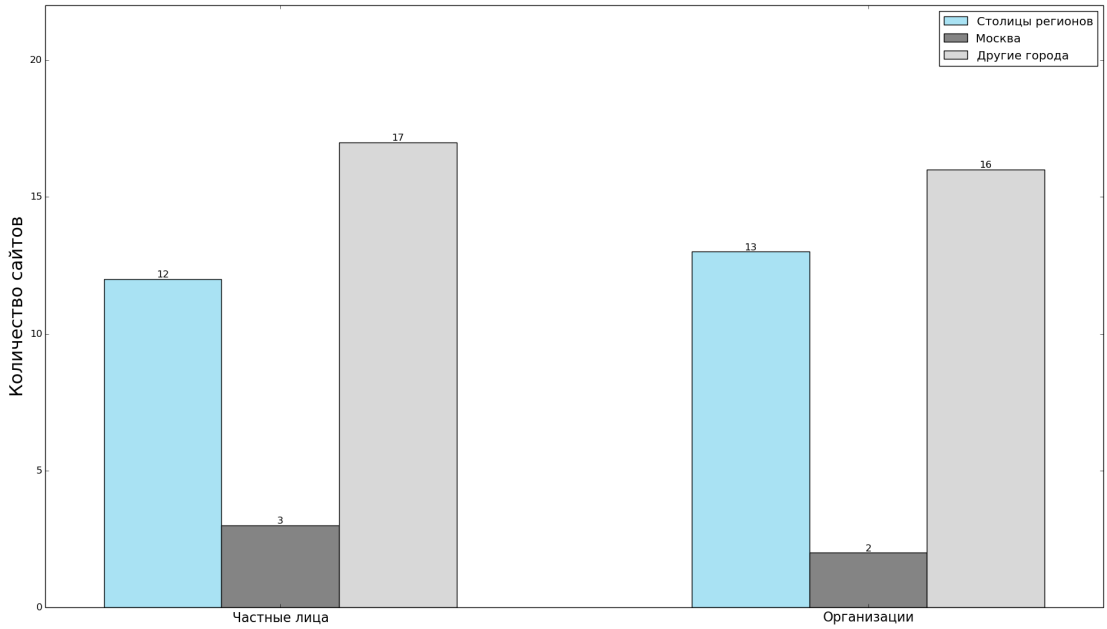
[3] Распределение доменов в зоне RU по типам владельцев (рассчитано на 2003-05-01) – URL: http://stat.nic.ru/2003/05/31/owner-20030531.shtml

изменяться в пользу юридических лиц — их доля стала увеличиваться и поднялась до 33%», чем может быть объяснён небольшой пик регистраций сайтов на малых языках организациями в 2011-ом году.

На рис. 7 и 8 представлена информация по городам регистрации сайтов. К сожалению, нам удалось получить такую информацию всего для 75 сайтов, при этом для 11 из них неизвестно, кем зарегистрирован сайт, и ещё 1 сайт зарегистрирован хостингом, поэтому на графиках представлена информация только для 63 сайтов.

На графиках отдельно представлены сайты, зарегистрированные в столицах титульных регионов распространения языка, столице Российской Федерации и в других городах. Отметим, что в список других городов входят столицы регионов, в которых не говорят на рассматриваемых языках, – Саратов и Саранск; столицы других стран – Стамбул, Вашингтон, Лондон, Брюссель, Бишкек, Астана, Нассау; и другие административные центры – Касумкент, Нижний Новгород, Инжич-Чикун. Подробная информация по каждому типу владельцев с указанием конкретных городов доступна в Приложении С.

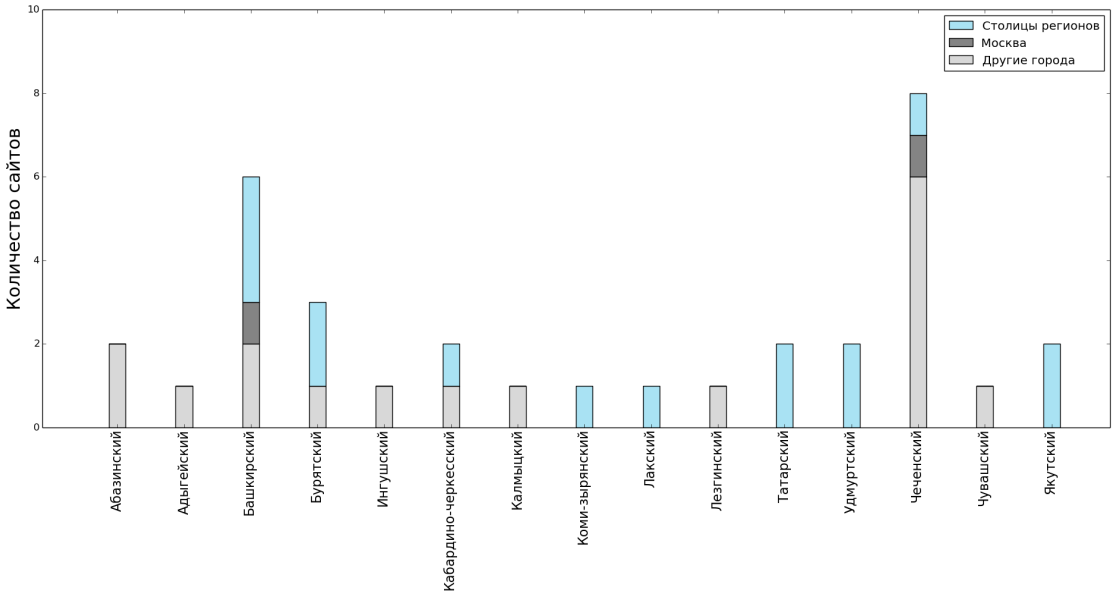
Рис. 7: Общее распределение сайтов по городам



На рис. 8 данные по городам представлены отдельно по каждому малому

языку, для которого известна информация о городе регистрации сайта. Видно, что все сайты, для которых было возможно получить данные о городе с помощью whois, на коми-зырянском, лакском, татарском, удмуртском и якутском были созданы в столицах соответствующих регионов – в Сыктывкаре, Махачкале, Казани, Ижевске и Якутске.

Рис. 8: Распределение сайтов по городам для каждого языка



Больше всего данных по городам известно для башкирского, бурятского и чеченского. Рассмотрим эти языки отдельно.

Все сайты, указанные в таблице 4 имеют адреса в доменах верхнего уровня .com, .org, .net, .info, .mn, .fm, поэтому сравнивать эти сайты со статистикой Рунета не имеет смысла. Сайты, для которых не известен владелец, но известен город регистрации, убраны из таблицы 4 и рис. 8, т. к. зачастую оказывается, что для таких сайтов информация о городе была искажена с помощью одного из сервисов защиты whois-записей. По всей видимости, аналогичная процедура защиты whois-данных была применена и для башкирских сайтов в Нобби бич (замена сделана аналогично примеру на специализированном сайте защиты данных (*About Privacy Protection* б.г.)) и Сан-Матео (*Domain Privacy* б.г.). Однако, при этом по этим искажённым данным понятно, кем был зарегистрирован сайт. Та же схема была применена и для чеченского сайта, зарегистрированно-

Таблица 4: Регистрации сайтов по городам для башкирского, бурятского и чеченского

Башкирский		Бурятский		Чеченский	
Уфа	3	Улан-Удэ	2	Стамбул	2
Москва	1	Уб	1	Москва	1
Нобби бич	1			Грозный	1
Сан-Матео	1			Лондон	1
				Вашингтон	1
				Орем	1
				Саннивейл	1
Неизвестно	60	Неизвестно	16	Неизвестно	12

го в Ореме. Бурятский же сайт в Убе относится к тем трём сайтам, которые были зарегистрированы в 2016-м году, поэтому тоже не представляет особого интереса.

Получается, что та небольшая толика сайтов, о которых известен более менее достоверный город регистрации, зарегистрированы либо в столицах титульных регионов распространения языка, либо в Москве. Исключение составляют два одинаковых чеченских сайта, зарегистрированные одним и тем же человеком, но в разное время и под разными доменами первого уровня (dinulislam.org и dinulislam.net); один сайт чеченской газеты, зарегистрированный в Лондоне (thechechenpress.com); да два сайта из США (chechenews.com и radiomarsho.com). Если верить интернет-энциклопедии «Википедия» , в которой говорится, что «с начала XX века чеченские беженцы образовали заметные диаспоры в крупных городах США и Канаде, куда многие из них прибыли в качестве беженцев через территории других, дружественных НАТО стран (например, Турцию)» (*Чеченская диаспора* б.г.), то можно предположить, что сайты на чеченском в других странах были созданы чеченскими диаспорами, образовавшимися после эмиграции из России.

Отметим ещё раз, что для большинства сайтов информация о городе регистрации неизвестна, а об известных городах регистрации данные могут быть

искажены специальными сервисами защиты whois-записей. Поэтому на основе этих данных можно разве что предположить, что чаще всего не скрывают информацию о регистрации владельцы сайтов из России. Однако, это не значит, что мы знаем про все регистрации, сделанные в России.

Сервисы, собирающие статистику по доменам, такие как (*Domaintools* б.г.) и (*Российский рынок доменов* б.г.), смотрят также на распределение доменов по городам на основе информации об IP-адресе, к которому привязан домен. С помощью свободно распространяемого ip-api (*ip-api* б.г.) можно получить такую информацию и для наших сайтов.

На рис. 9 и 10 представлена информация о географическом распределении сайтов на малых языках по разным городам на основе IP-адресов, к которым привязаны сайты. В связи с большим количеством сайтов с IP-адресом в Санкт-Петербурге эти данные были вынесены в отдельный столбец. Информация о городах, полученная с помощью IP-адресов, по большей части не совпадает с информацией, указанной в whois-записи. Так, например, сайт на бурятском buryad.fm, зарегистрированный по адресу г. Улан-Удэ, согласно данным по IP находится в Чебоксарах, а сайты sabantuy.net и www.chuvash.org находятся не в Казани и Нижнем Новгороде, а в Санкт-Петербурге.

Рис. 9: Общее распределение доменов по городам на основе IP-адресов

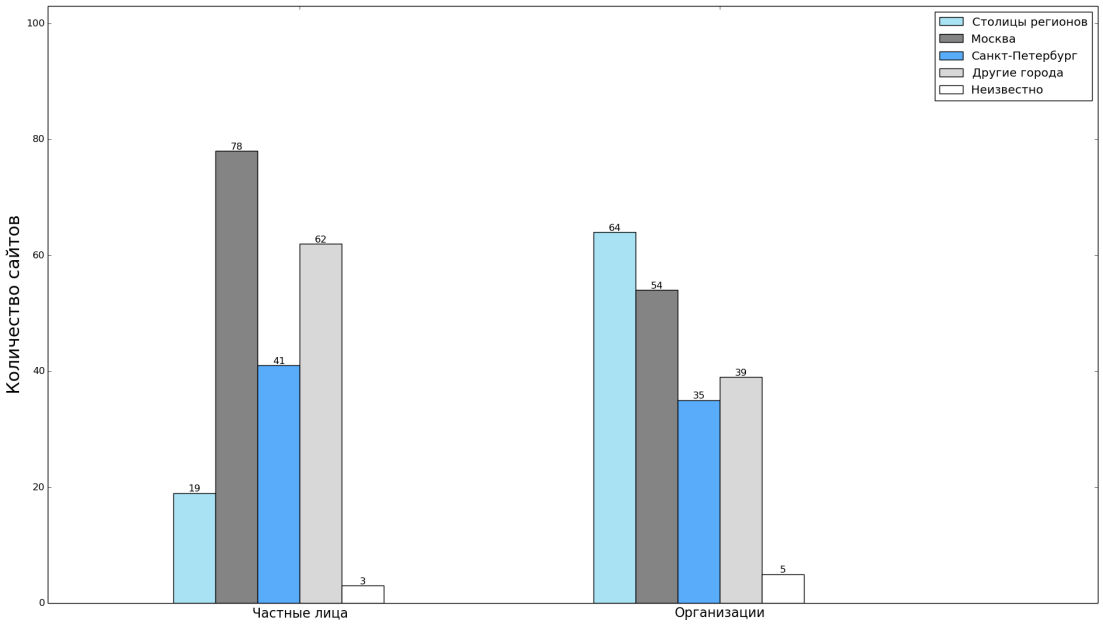
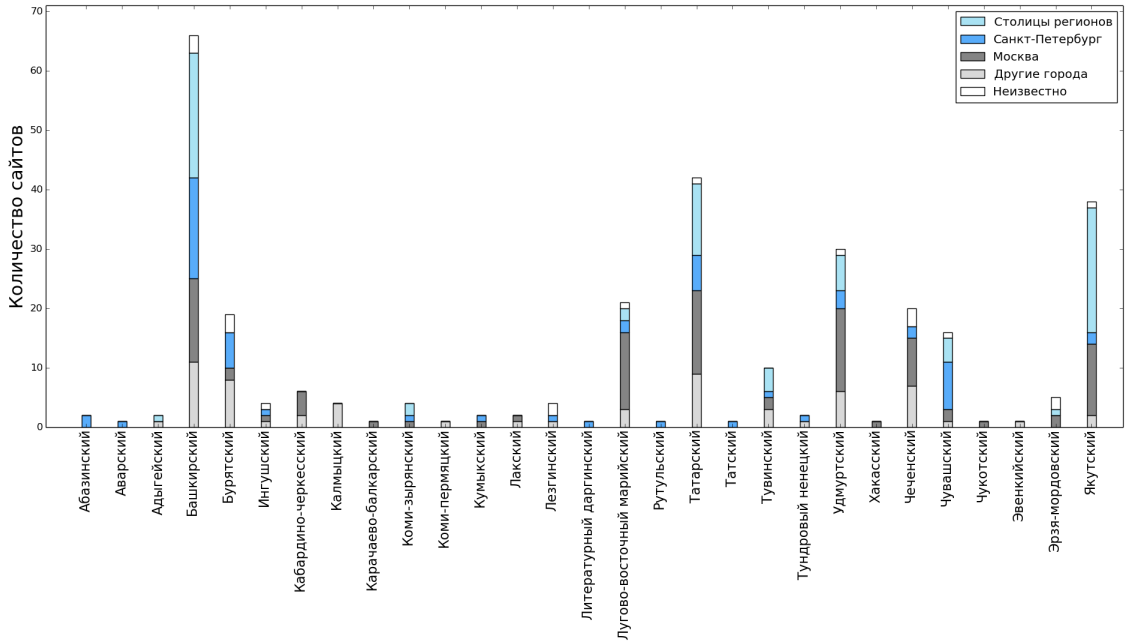


Рис. 10: Распределение доменов по городам на основе IP-адресов



Подавляющее большинство сайтов определяются по IP как находящиеся в одной из трёх столиц: столице титульного региона распространения языка, Москве или Санкт-Петербурге («второй столице РФ»). Остальные сайты расположены на серверах в США (34), Германии (19), Украине (6), Эстонии (6). Такая статистика в целом совпадает с данными RU-CENTER как по странам[4] – лидирующие позиции России ($\sim$ 90%), следующие за ней Украина ($\sim$ 1%), Германия ($\sim$ 1%), США ($\sim$ 1%)), так и по регионам[5] – лидирующие позиции Москвы ($\sim$ 36%) и Санкт-Петербурга ($\sim$ 7%)). В отличие от рассмотренной выше ситуации с башкирским, бурятским и чеченским, данные по IP более полные и найдены для 420 из 432 сайтов, доля же сайтов в домене .ru на малых языках составляет 77%, поэтому наши данные допустимо сравнивать с общей статистикой по Рунету.

Отметим, что среди других городов в небольшом количестве встречаются не только зарубежные города, но и города России. В списке городов можно найти административные центры регионов РФ, которые не являются титульными

[4] Распределение зарегистрированных доменов в реестре домена RU по странам (рассчитано на 2010-12-01) – URL: http://stat.nic.ru/2010/12/31/country-20101231.shtml

[5] Распределение зарегистрированных доменов в реестре домена RU по субъектам Российской Федерации (рассчитано на 2010-12-01) – URL: http://stat.nic.ru/2010/12/31/region-20101231.shtml

для рассматриваемого подмножества малых языков (в скобках указано количество сайтов, отнесённых по IP-адресу к этому городу): Иркутск (8), Новосибирск (2), Вологда (1), Саратов (1), Оренбург (1), Нарьян-Мар (1), Красноярск (1), Когалым (1), Калининград (1); города московской области: Электросталь (2), Реутов (1), Одинцово (1); и некоторые другие: Хорошёвская (3), Волга (1), Посёлок N1 (1), посёлок Луховка (1), Первоуральск (1), Арзамас (1). Полный перечень городов, определённых по IP-адресу, разбит по типам владельцов сайтов и приведён в Приложении D.

По рис. 9 можно заметить, что организации гораздо чаще частных лиц располагают свои сайты на серверах в столицах регионов. Можно предположить, что у организаций чаще всего есть свой сервер, а частные лица пользуются услугами сторонних компаний, предлагающих разместить сайт на их сервере, который может быть расположен заграницей или в одном из городов федерального значения – Москве или Санкт-Петербурге.

На основе наших данных можно составить топ-5 хостингов с сайтами на малых языках. По таблице 5 видно, что владельцы сайтов на башкирском и татарском предпочитают локальных хостинг-провайдеров, тогда как владельцы сайтов на остальных языках пользуются просто повсеместно популярными хостингами: Timeweb входит в топ-10 лучших хостингов по мнению (*Лучшие хостинги* б.г.), REG.RU является одним из самых популярных регистраторов доменов (см. таблицу 3), поэтому кажется логичным, что часть сайтов предпочитают и их хостинги, РТКомм.ру, по всей видимости, удерживает свои позиции с 2008-го года (*Рунет в марте 2008 года: домены, хостинг, география сайтов* 2008).

Итак, в этом подразделе мы изучили регистрационные данные сайтов на малых языках, а также ознакомились с гео-информацией, полученной на основе IP-адресов этих сайтов. Мы работали с 432 сайтами, 124 из которых мы затрудняемся отнести к какому-то конкретному малому языку. Количество сайтов, созданных частными лицами (203), незначительно превышает количество сайтов на малых языках, созданных организациями (197). Кажется, что сайты на малых языках прежде всего создают СМИ и государственные учреждения. Пик

Таблица 5: Топ-5 хостинг-провайдеров для сайтов на малых языках

Хостинг-провайдер	Количество сайтов	Город	Примечание
Timeweb	17	Санкт-Петербург	сайты на 11 малых языках
Эр-телеком	15	Уфа	все сайты на башкирском
Tatmedia	10	Казань	все сайты на татарском
REG.RU	8	Москва	сайты на башкирском, лезгинском, чеченском, чувашском, якутском
РТКомм.ру	8	Москва	все сайты на удмуртском

регистраций сайтов на малых языках приходится на 2012-ый год, что может быть связано с активной территориальной экспансией регистраторов доменов в регионы, а также с проведением ими сопроводительных разъяснительных мероприятий. С 2012-го года наблюдается спад в регистрации новых сайтов на интересующих нас языках, что предположительно может быть связано связано с насыщением «рынка малых языков», закончившимися акциями от регистраторов и/или перераспределением интернет-аудитории в социальные сети. Отметим также, что с 2005-го года частные лица начинают активнее организаций регистрировать сайты на малых языках, что в целом совпадает с общей тенденцией Рунета. По нашим малочисленным сведениям о городах регистрации сайтов известно, что больше трети сайтов (40%) на малых языках создаются в столицах титульных регионов распространения малого языка. Т. к. большинство контактных данных (в т. ч. адрес) в whois-записях скрыто или искажено, на основе этой информации можно разве что предположить, что в регионах России реже скрывают адрес регистрации. Впрочем такая статистика в целом подтверждается и полученными географическими данными по IP-адресам сайтов: сайты в столицах регионов составляют 30% от общего числа сайтов на малых языках. Большая часть таких сайтов была зарегистрирована организациями, которые

возможно располагают своими собственными серверами непосредственно в столицах регионов распространения языка или предпочитают хостинг-провайдеров с местными серверами. Остальные сайты в основном «хостятся» либо в Москве и Санкт-Петербурге, либо зарубежом, что может быть связано скорее с популярностью и распространённостью столичных хостинг-провайдеров, чем с чем-либо другим.

Эксперты, однако, утверждают, что «на единичных выборках, сопоставляющих значение IP-адреса, полученное для веб-сайта, с географическими данным, такой метод слишком неточен, чтобы делать на его основе содержательные выводы о георасределении контента сайтов»[6], а понятия «регистрация доменов» и «регистрация сайтов» не одно и тоже, поэтому к представленным данным и выводам следует относиться с осторожностью.

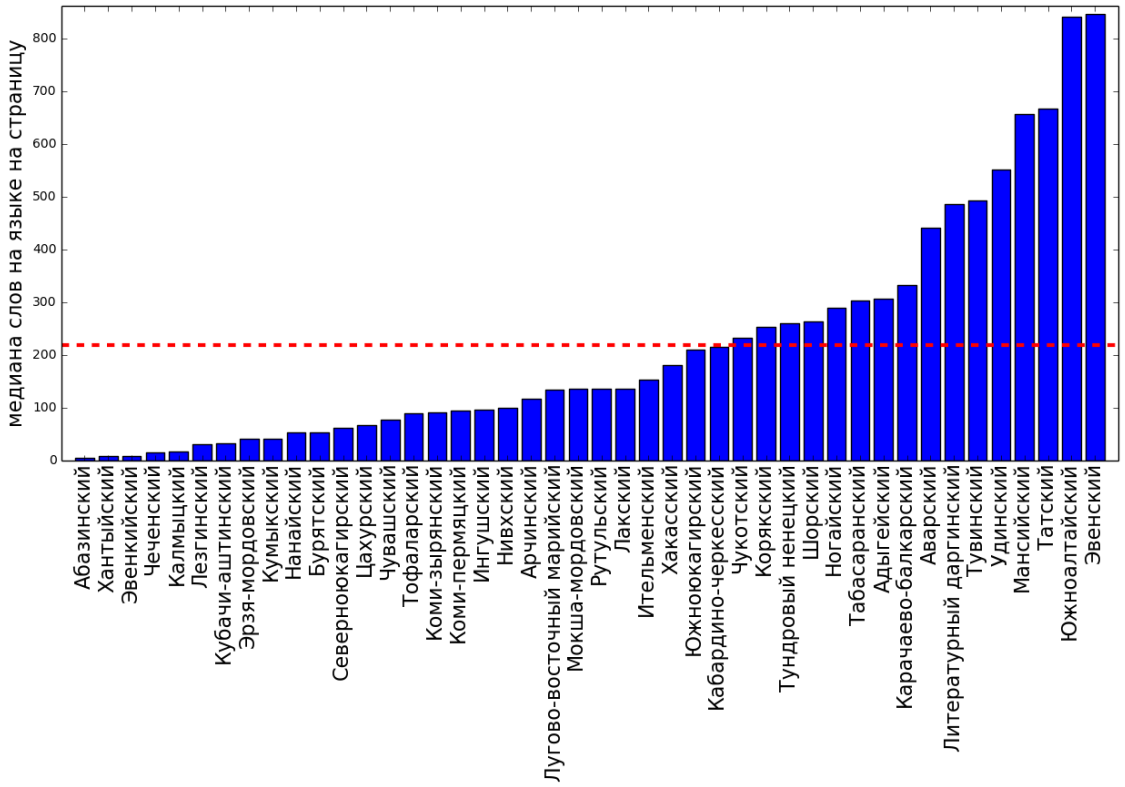
3.2. Медианное значение токенов на страницу текста на малом языке

Медианное значение токенов на страницу текста на малом языке является, на наш взгляд, ещё одним показателем того, насколько язык «уверенно себя чувствует» в интернете. Самая актуальная по количеству токенов и страниц информация опубликована на сайте (*Языки России* б.г.). Ниже показан график распределения медианного значения токенов (далее – медиана токенов) по языкам (см. рис. 11). Видно, что у этого значения достаточно большой разброс – маленькое минимальное и на 4 порядка больше его максимальное. На графике языки упорядочены по медианному значению слов на страницу, а красной пунктирной линией обозначена среднее данного значения по всем языкам.

Как видно из рис. 11, распределение языков по медиане токенов не совсем соответствует нашим ожиданиям – почти все самые большие по количеству токенов и страниц языки находятся ниже среднего медианного значения. Это может быть объяснено методикой создания текстовой коллекции. Дело в том, что порог количества страниц, после которого мы выкачивали сайт целиком был статическим, ввиду сложности технической реализации мы никак не учитывали общее количество страниц сайта на этапе поиска и сортировки сайтов по груп-

[6] из личной переписки с А. Венедюхином

Рис. 11: Распределение медианного значения токенов на веб-страницу



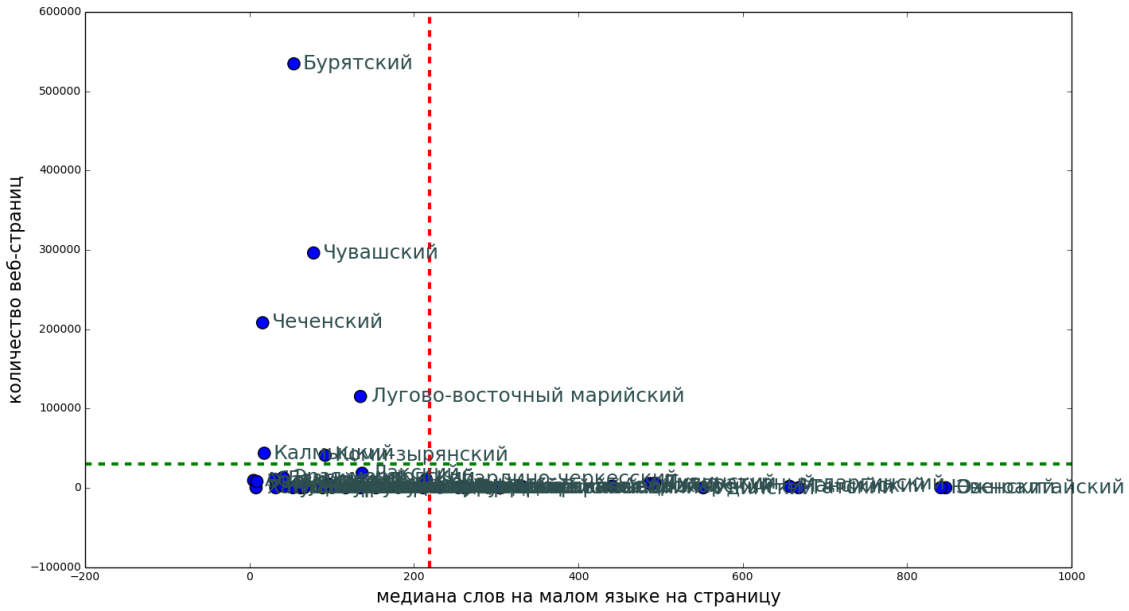
пам. Конечно, при подсчёте скачанного количества веб-страниц на малом языке мы учитывали только те страницы, на которых есть хотя бы один абзац на малом языке, но зачастую эти абзацы могли быть просто краткими заголовками новости на соседней странице. Так, например, на бурятском сайте «Россия Бурятия» мы нашли с помощью поискового движка «Яндекса» 86 страниц на бурятском.Вот эта конкретная страница http://bgtrk.ru/tv/archive/syuzhety/85633/ была учтена в общем количестве страниц на бурятском, что верно с точки зрения методики, ведь в правой колонке представлены названия программ бурятского телевидения по-бурятски, но оказывает влияние на расчёт медианы токенов на страницу. Можно предположить, что конкретно на этом сайте таких страниц много, а ведь всего с сайта было выкачано больше сорока тысяч страниц.

Вероятно, для того чтобы можно было сделать более содержательные выводы, полезно будет рассмотреть медианное значение токенов на веб-страницу относительно общего числа веб-страниц на этом языке. На рис. 12 представлена диаграмма рассеяния по двум интересующим нас параметрам. Красной пунктирной линией снова отмечено среднее медианного значения токенов, а зелёной пунктирной линией обозначено среднее значение количества веб-страниц.

Из диаграммы рассеяния становится понятно, что для языков, для которых в коллекции много веб-страниц, медиана стремится к некоторому определённому значению, если же страниц мало, то медиана токенов может принимать любые значения (от минимального до максимального). Вероятно, принимать какое-то решение об «уверенности языка в интернете» на основе медианного значения токенов можно только при наличии достаточного количества скачанных веб-страниц на малом языке (не меньше 10000 страниц, возможно, больше 30000 – значение у средней линии). Уже сейчас можно предположить, где будут на осях располагаться невыкачанные ещё бурятский, татарский, якутский и удмуртский – в верхнем левом квадрате.

Отметим также, что метод определения языка текста, использовавшийся в проекте, работает не со 100% точностью, поэтому часть текста могла быть отнесена к малому языку ошибочно и таким образом учитывались страницы

Рис. 12: Отношение общего количества веб-страниц и медианнго значения токенов на веб-страницу



и токены, которые при более тщательном (скорее ручном) составлении коллекции могли бы быть вычеркнуты.

4. Анализ данных

В этом разделе с помощью нескольких методов анализа данных мы попробуем выявить связи и зависимости для различных параметров малых языков. Для этого в нашем распоряжении следующие наборы данных:

- интернет-данные (посчитаны по интернет-коллекции)
  - количество всех сайтов, с которых была выкачана хотя бы одна страница на малом языке
  - количество веб-страниц на малом языке
  - количество токенов на малом языке
  - количество найденных сайтов на малом языке (см. раздел «Сайты на малых языках»)
  - количество выкачанных сайтов на малом языке

  - медианное значение токенов на веб-страницу

- ВКонтакте-данные (данные, полученные из социальной сети ВКонтакте, предоставлены Л. Я. Зайдельман)
  - количество сообществ на малом языке
  - общее количество постов
  - количество постов на малом языке
  - количество токенов на малом языке

- вики-данные (данные из соответствующих языковых разделов Википедии; данные за 2014 и 2015 года предоставлены Б. В. Ореховым)
  - количество статей за 2014, 2015 и 2016 года
  - количество активных пользователей на апрель 2016-го года
  - количество админов на апрель 2016-го года
  - наличие википедии или инкубатора википедии для малого языка

- данные о носителях (эти и все перечисленные ниже данные предоставлены Б. В. Ореховым)
  - число носителей за 2002 и 2010 года (по переписи наслеения)
  - прирост носителей на 2010 год (с 2002 года)
  - средний возраст носителей
  - религия

- данные о титульном регионе распространения (люди)
  - процент интернет-пользователей за 2010 и 2011 года
  - количество компьютеров на семью за 2010 год
  - плотность населения
  - количество детских садов
  - показатели рождаемости и смертности за 2014-ый год

  - прирост населения на 2014 год
  - процент горожан в регионе (?)
- данные о титульном регионе распространения (финансы)
  - дотации и расходы региона за 2010, 2011 и 2014 года
  - средняя месячная зарплата за 2010 и 2015 года
  - среднедушевой доход за 2002 и 2010 года
- данные о титульном регионе распространения (статус)
  - федеральный округ
  - позиция рейтинга региона за 2013 и 2014 года
  - индекс рейтинга региона за 2013 и 2014 года
- печатные издания на малом языке
  - количество книг за 2009 год, газет и журналов за 1996 год
  - тираж книг за 2009 год, газет и журналов за 1996 год
- данные о статусе языка
  - наличие iso кода для языка
  - статус языка в регионе (официальный, государственный, отсутсвует)
  - языковая семья
  - титульный регион распространения

Каждый набор параметров (в разделе синонимично используется с «данными» и «признаками»), на наш взгляд, в той или иной степени может характеризовать статус жизнеспособности языка и влиять на его представленность в интернете. Все онлайн-данные дают представление о том, насколько язык присутствует в цифровом мире. Данные о носителях и их возрасте обычно используются для составления шкалы жизнеспособности языка (EGIDS), поэтому могут

быть полезны и нам. Так как наше исследование в первую очередь ориентировано на онлайн-существование языка, то кажется важным учесть и количество интернет-пользователей и компьютеров в регионе. Немаловажны и финансовые показатели региона, так как при низком финансировании и малых зарплатах носители не смогут позволить купить себе компьютер, а значит и «выйти общаться в сеть», а администрация региона будет тратить деньги не на повышение компьютерной грамотности населения и поддержания языка, а на более насущные проблемы офлайн-жизни. Отметим, что в нашем исследовании каждый малый язык отнесён к одному региону Российской Федерации, несмотря на то, что на этом языке могут разговаривать и в других регионах. Наличие и количество печатных изданий на малом языке могут свидетельствовать о том, что язык существует не только в устной форме, но и поддерживается письменно. Статус языка может также оказаться важным параметром. Наличие или отсутствие iso-кода свидетельствует о том, замечен ли язык международными организациями стандартизации (правда, в наших данных есть всего один язык без iso-кода – кубачи-аштинский). Официальный статус языка может способствовать его более быстрому продвижению и развитию как в реальной жизни, так и в онлайн-среде.

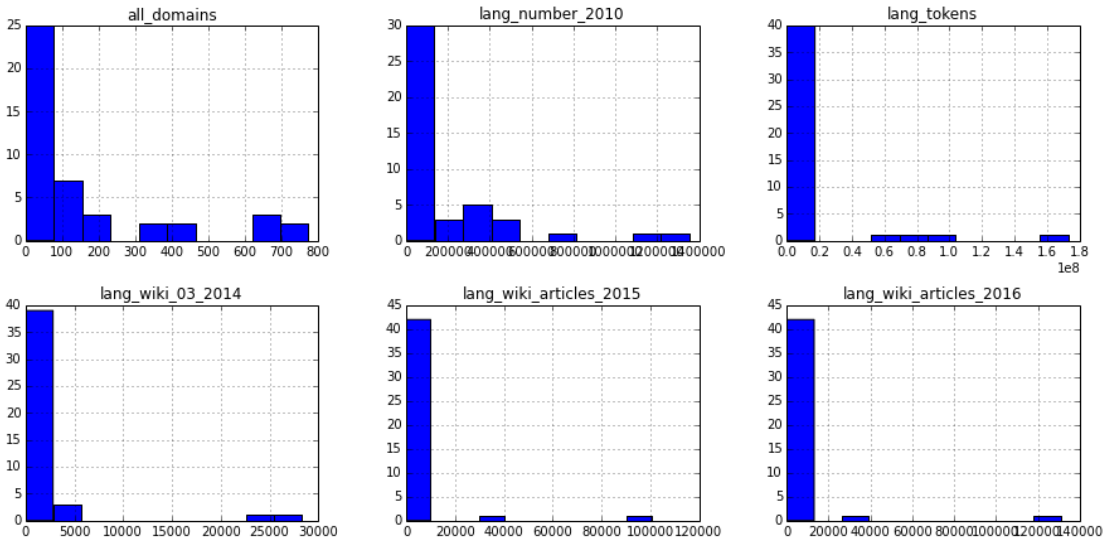
В качестве методов анализа данных были выбраны анализ корреляций для поиска сильносвязанных параметров, регрессионный анализ для выявления линейной зависимости интернет-параметров от других параметров и кластерный анализ для упорядочивания малых языков на основе имеющихся онлайн и офлайн параметров. Подробнее обработка данных и результаты каждого метода описаны в подразделах ниже. В заключительном подразделе мы исследуем характеристики графов национальных интернетов и в частности вопрос их связности.

4.1. Подготовка данных

Из книг по статистике (например, (Бослаф 2015)) известно, что большинство статистических методов хорошо работают на нормально распределённых данных. Построив гистограммы по всем нашим данным, мы убедились в том, что наши данные не удовлетворяют этому требованию. В связи с этим было

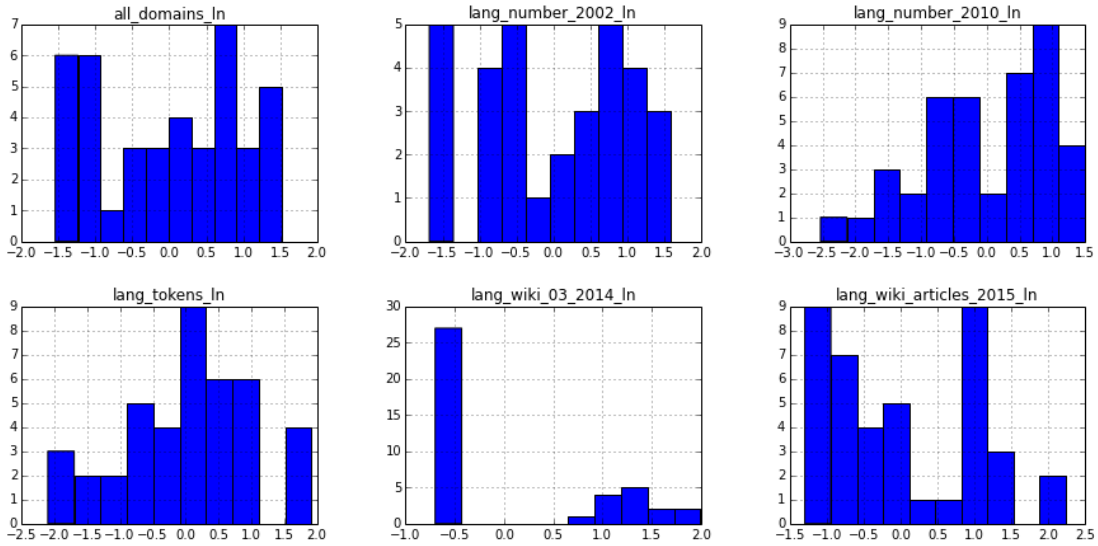
решено взять натуральный логарифм от тех данных, распределение которых наиболее далеко от нормального, и нормировать все параметры. Т.к. логарифм от 0 равен бесконечности, а в наших данных довольно много 0 и это значение является ценным, то мы использовали логарифм со смягчением: каждое значение было увеличено на 1. Пример исходных данных приведён на рис. 13. Эти же данные после нормирования показаны на рис. 14. Ввиду большого количества признаков кажется нецелесообразным приводить их все непосредственно в тексте работы, поэтому для более подробного ознакомления в репозиторий нашего проекта выложен ipython-notebook с графиками для всех признаков (см. Приложение E).

Рис. 13: Гистограммы исходных данных по всем сайтам (all_domains), по числу носителей за 2010 год (lang_number_2010), по количеству токенов на малом языке в интернете (lang_tokens), число статей в Википедии за 2014, 2015 и 2016 года (lang_wiki_03_2014, lang_wiki_articles_2015, lang_wiki_articles_2016)



Видно, что после логарифмирования распределение наших данных стало ближе к нормальному. Далее будем работать с нормированными данными. Отметим также, что в данных присутствуют и категориальные признаки, которые были отобраны отдельно и использовались только в регрессионном анализе. Для регрессионного анализа каждый категориальный признак переведён в набор векторов из 0 и 1. Это стандартная методика использования кате-

Рис. 14: Гистограммы логарифмированных и нормированных данных по тем же признакам, что и на рис. 13



гориальных признаков в моделях, т. к. мы не знаем, чем, например, в случае с категориями-регионами один регион лучше другого. А если даже знаем, что наличие википедии статусно должно иметь более высокое значение, чем наличие лишь инкубатора или полное отсутствие языкового раздела в википространстве, то не можем наверняка знать порядок этих различий, т. е. нужно ли эти данные изменить на 2 – 1-0 или 4 – 2-0.

4.2. Анализ корреляций

Анализ корреляций, как правило, используют для первичного выявления связей между различными параметрами. Чаще всего используется корреляционный коэффициент Пирсона. Однако, известно, что корреляция Пирсона использует допущение о нормальности распределения (*Корреляция* б.г.), поэтому мы дополнительно рассчитали корреляционный коэффициент Спирмена и сравнили полученные значения. Сравнение значений показывает, что несмотря на то, что для большинства пар признаков коэффициенты различаются на сотые, для некоторых признаков различие идёт на десятки (например, для признаков рейтинга и позиции региона за 2013-ый год разница составляет 0.64). Как мы писали выше, корреляция Пирсона в большей степени ориентирует-

ся на нормальность распределения, а распределение наших данных не удалось свести к достаточно схожему с нормальным, поэтому считаем уместным ориентироваться в анализе и выявлении сильных связей между признаками на корреляцию Спирмена.

В (Бослаф 2015) говорится, что определение сильных и слабых связей зависит от области исследования. В компьютерной лингвистике традиционно миниально сильным значением коэффициента считают значение коэффициента корреляции больше 0.5, а о сильных связях говорят при коэффициенте больше 0.7. Таблица значений коэффициентов корреляции Спирмена для всех пар признаков приводится в ipython-notebook, ссылка на который есть в Приложении Е. В notebook приведены только значения по модулю большие 0.7.

Анализ корреляций показывает, с одной стороны, довольно очевидные связи, например, между статьями википедии за разные годы, числом носителей за разные годы, между дотациями и расходами титульного региона распространения языка за разные годы, между сообществами, постами и количеством токенов ВКонтакте, между сайтами, веб-страницами и токенами на малом языке в интернете. С другой стороны, мы видим довольно сильную корреляцию между всеми имеющимися у нас данными о текстовой представленности языка в интернете: статьи в Википедии за 2015 и 2016 года с текстами из интернета и ВКонтакте – сайтами, веб-страницами, сообществами, постами. Также сильна связь разных печатных изданий (книги, журналы, газеты) с числом носителей за 2002-ой, но не за 2010-ый года. Отметим, что число носителей за 2002-ой год коррелирует (т. е. значение коэффициента корреляции больше 0.7) с гораздо большим количеством интернет-данных, чем число носителей за 2010-ый год. Возможно это говорит о том, что малый язык в веб-пространстве развивается по своим отдельным законам и не так сильно привязан к увеличению или уменьшению числа носителей языка. Также это может быть объяснено тем, что данные из интернета, как правило, какое-то время никуда не пропадают (до момента окончания регистрации домена) в то время как число носителей малых языков может уменьшаться. Из любопытных связей, касающихся офлайн-жизни языка, можно отметить связь процентного числа интернет-пользователей

региона с количеством компьютеров на семью по региону, а также со средними зарплатами по региону и просто с процентным числом горожан. Однако, нельзя сказать, что эти офлайн-связи самые неожиданные.

4.3. Регрессионный анализ данных

В качестве методов регрессионного анализа мы использовали однофакторную и многофакторную линейные регрессии. Регрессионная модель рассчитывалась по методу наименьших квадратов статистическим пакетом (*StatsModels* б.г.) для языка Python. Для построения модели были созданы две таблицы признаков: с исключением языков без интернет-данных и полностью без столбцов вконтакте-данных (интернет-таблица) и таблица с исключением всех языков, для которых нет либо интернет-данных, либо вконтакте-данных (онлайн-таблица). Первая таблица была сформирована для того чтобы не терять те языки, для которых есть интернет-данные, но нет вконктакте-данных. Вторая таблица позволяет использовать полный массив онлайн-данных, но лишь для тех языков, для которых эти данные есть.

В нашем распоряжении 51 числовой признак и ещё 7 категориальных для 41 языка в онлайн-таблице; и 47 числовых с 7 категориальными признаками для 44 языков в интернет-таблице. Нас прежде всего интересует, есть ли линейная зависимость интернет-данных о малых языках от каких-нибудь офлайн-признаков по языку и титульному региону его распространения. Для выяснения этой зависимости построим сначала простую линейную регрессию от числа носителей за разные года, а потом попробуем группами добавлять другие факторы и смотреть на то, как изменяется модель линейной регрессии и какие связи становятся значимыми.

В качестве представляющего интернет и онлайн-жизнь языка параметра выберем количество токенов на малом языке. Этот выбор обоснован тем, что в общем количестве сайтов учитываются все те, на которых была найдена хотя бы одна страница на языке, а значит этот параметр может характеризовать скорее «размазанность» языка по разным сайтам. В количество страниц языка включены только те страницы, на которых есть хотя бы один токен на этом языке,

но т. к. данные не подвергались дополнительной очистке полученных текстовых коллекций, то в это значение включены и страницы с крайне малым количеством слов на миноритарном языке, в том числе могут попадаться и страницы просто с названиями книг или однословными комментариями. Мы считаем, что количество токенов наиболее подходящая величина для иллюстрации онлайн-жизни языка.

При оценке получившейся модели будем прежде всего смотреть на F-статистику (далее в таблицах: F-stat.) и её значимость (P(F-stat)), а также на значение R-квадрат (R-Squared) и скорректированный R-квадрат (Adj. R-Squared). Насколько понятно из книги (Бослаф 2015), не так важно само значение F-статистики как то, насколько оно значимо. Чем меньше показатель P(F-stat), тем меньше вероятность принять модель, когда она не верна. Возьмём стандартный уровень значимости равный 0.05, т.е. значимой считаем модель при значении p-value (P(F-stat)) < 0.05. Значение R-квадрат показывает то, насколько модель хорошо соответсвует набору данных (т. е., например, R-квадрат $= 0.521$, означает, что модель смогла аппроксимировать 52% наблюдений – объяснить больше 50% дисперсии набора данных). Однако, в (Frost 2013) утверждается, что при увеличении количества независимых переменных R-квадрат будет скорее всего только возрастать и перестанет адекватно отражать качество модели. Для понимания того, насколько модель хорошо будет предсказывать данные не из выборки, следует использовать скорректированный R-квадрат. Этот параметр принимает во внимание влияние новых факторов на построенную модель, т. е. именно на него следует смотреть при подборе факторов для модели множественной регрессии методом включения.

В таблицах ниже мы также будем приводить значения полученных коэффициентов для независимых переменных: стандартизованный коэффициент регрессии (coef), стандартную ошибку (std err), t-статистику (t) и её значимость (P>|t|). Мы приводим все значения, которые показывает наш статистический пакет, однако, в анализе в первую очередь будем учитывать только значимость независимых переменных.

Для того чтобы убедиться в том, что в данных действительно есть линей-

ная зависимость, бывает полезно оценить нормальность и независимость ошибок (остатков). Для оценки нормальности мы использовали, так называемый, qq-plot – точки, расположенные практически на прямой свидетельствуют о нормальности ошибок. Для оценки независимости ошибок мы рассчитывали статистику Дарбина-Уотсона, значения которой лежат в интервале от 0 до 4, а значение близкое к 2 показывает абсолютную независимость. Мы не приводим эти параметры в тексте работы, однако, рассчитали их для всех результирующих моделей и можем сказать, что ошибки модели достаточно независимы и нормальны. Графики и значения статистики Дарбина-Уотсона приводятся в ipython-notebook, ссылка на который есть в Приложении Е.

Линейная регрессия нормированного логарифма количества токенов от нормированного логарифма числа носителей за 2002 и 2010 года показала, что количество токенов на языке зависит и от числа носителей 2002-го года и от числа носителей 2010-го года. Полученные значения и модель значимы, регрессия для значения числа носителей за 2002-ой год подходит чуть большему количеству данных, однако значимость выше у второй модели. Параметры обеих линейных регрессий приведены в таблице 6. Значение константы в обеих моделях не значимо, поэтому в таблице не представлено.

Таблица 6: Параметры линейной регрессии ln(числа токенов) от ln(числа носителей)

Параметры моделей				
Независимая переменная	R-squared	Adj. R-Squared	F-stat.	P(F-stat)
ln(число носителей за 2002г)	0.525	0.511	36.50	8.54e-07
ln(число носителей за 20010г)	0.510	0.499	43.80	5.12e-08
Значения коэффициентов				
	coef	std err	t	P>|t|
ln(число носителей за 2002г)	0.7444	0.123	6.041	0.000
ln(число носителей за 20010г)	0.7145	0.108	6.618	0.000

С помощью множественной линейной регрессии проверим, какие факторы ещё могут влиять на представленность языка в интернете. Будем добавлять

факторы группами, при этом из двух сильно коррелированных факторов про один и тот же тип данных (например, тираж книг и количество книг) будем выбирать какой-то один и за самый последний из известных годов.

Мы перебрали около 70 различных наборов независимых переменных[7] и лучшая модель линейной регрессии была достигнута на наборе из 4 признаков: ln(число носителей за 2010 год), прирост населения на 2014 год, расходы региона за 2014 год и ln(число газет на 1996 год). Оказалось, что с помощью этих признаков можно построить хорошую модель не только для количества токенов, но и для количества веб-страниц, количества всех сайтов и лучше всего для количества сайтов на малых языках. При этом значимыми параметрами в модели оказываются только число носителей и число газет, остальные параметры не значимы, но при их удалении модель ухудшается. Параметры полученных моделей для указанных выше зависимых переменных приведены в таблице 7.

Таблица 7: Параметры многофакторной линейной регрессии разных зависимых переменных от ln(числа носителей за 2010 год), прироста населения на 2014 год, расходов региона за 2014 год и ln(числа газет на 1996 год)

Параметры моделей				
Зависимая переменная	R-squared	Adj. R-Squared	F-stat.	P(F-stat)
ln(количество токенов)	0.629	0.591	16.52	5.36e-08
ln(количество веб-страниц)	0.639	0.602	17.27	3.14e-008
ln(количество всех сайтов)	0.634	0.597	16.91	4.05e-008
ln(количество сайтов на м.я.)	0.666	0.632	19.45	7.21e-009
Значения коэффициентов (для модели с ln(число токенов))				
	coef	std err	t	P>|t|
ln(число носителей за 2010г)	0.5634	0.136	4.151	0.000
прирост населения на 2014г	-0.1654	0.110	-1.510	0.139
расходы региона за 2014г	-0.1859	0.111	-1.671	0.103
ln(число газет за 1996г)	0.2368	0.119	1.987	0.054

Мы также проверили на линейную зависимость интернет-данных от других

[7] Параметры моделей для этих признаков см. в репозитории: файлы params_for_ols.tsv и params_for_ols_with_online.tsv

онлайн-параметров. Модель с наилучшими коэффициентами (скорректированный R-квадрат = 0.658, значимость F-статистики = 5.31e-010) получилась при использовании логарифма от количества статей в языковом разделе википедии за 2016-ый год и количества токенов в социальной сети ВКонтакте в качестве независимых переменных.

Однако, лучшая модель линейной регрессии, которую нам удалось построить, достигается при комбинировании указанных выше офлайн и онлайн-признаков: ln(число носителей за 2010 год), прирост населения на 2014 год, расходы региона за 2014 год, ln(число газет на 1996 год), ln(количество статей в википедии за 2016 год) и ln(количества токенов на языке во ВКонтакте). Модель описывает $65-75\%$ данных при выборе разных зависимых параметров и показывает неплохую предсказательную способность для количества сайтов на малом языке. Для проверки предсказания мы исключали из набора данных конкретный язык, заново нормировали их и переобучали модель. Для языков со средними значениями признаков (т. е. для языков, чьи значения признаков не максимальны и не минимальны) модель ошибается всего на $1-2$ пункта, а для калмыцкого модель точно предсказывает количество найденных сайтов на калмыцком языке. Параметры моделей для разных зависимых переменных на основе онлайн и офлайн признаков представлены в таблице 8. Мы приводим значения коэффициентов для независимых переменных только модели с зависимой ln(количество сайтов на малом языке), потому что сами значения для моделей с разными зависимыми меняются несущественно.

Регрессионный анализ показывает, что возможно подобрать набор офлайн-параметров и построить значимую модель многофакторной линейной регрессии, которая будет объяснять до 63% интернет-данных. Однако, такая модель сильно улучшается с добавлением онлайн-параметров из других источников (социальных сетей и Википедии) и начинает объяснять уже до 75% интернет-данных. Последняя также показывает неплохую предсказательную способность и может быть использована для приблизительной оценки количества сайтов на малом языке. В исследовании были использованы данные за разные года, т. е. одновременно в модель не попадали одни и те же признаки за разные года,

Таблица 8: Параметры многофакторной линейной регрессии разных зависимых переменных от ln(числа носителей за 2010 год), прироста населения на 2014 год, расходов региона за 2014 год, ln(числа газет на 1996 год), ln(количество статей в википедии за 2016 год) и ln(количества токенов на языке во ВКонтакте)

Параметры моделей				
Зависимая переменная	R-squared	Adj. R-Squared	F-stat.	P(F-stat)
ln(количество токенов)	0.711	0.661	13.97	6.03E-008
ln(количество веб-страниц)	0.706	0.654	13.61	8.15E-008
ln(количество всех сайтов)	0.710	0.659	13.89	6.47E-008
ln(количество сайтов на м. я.)	0.792	0.756	21.61	2.76E-010
Значения коэффициентов (для модели с ln(количество сайтов на м. я.))				
	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

но разные признаки для языков посчитаны не за один и тот же год. Мы предполагаем, что модель может измениться при использовании обновлённых данных за один и тот же год, а возможно, даже улучшиться.

4.4. Кластерный анализ данных

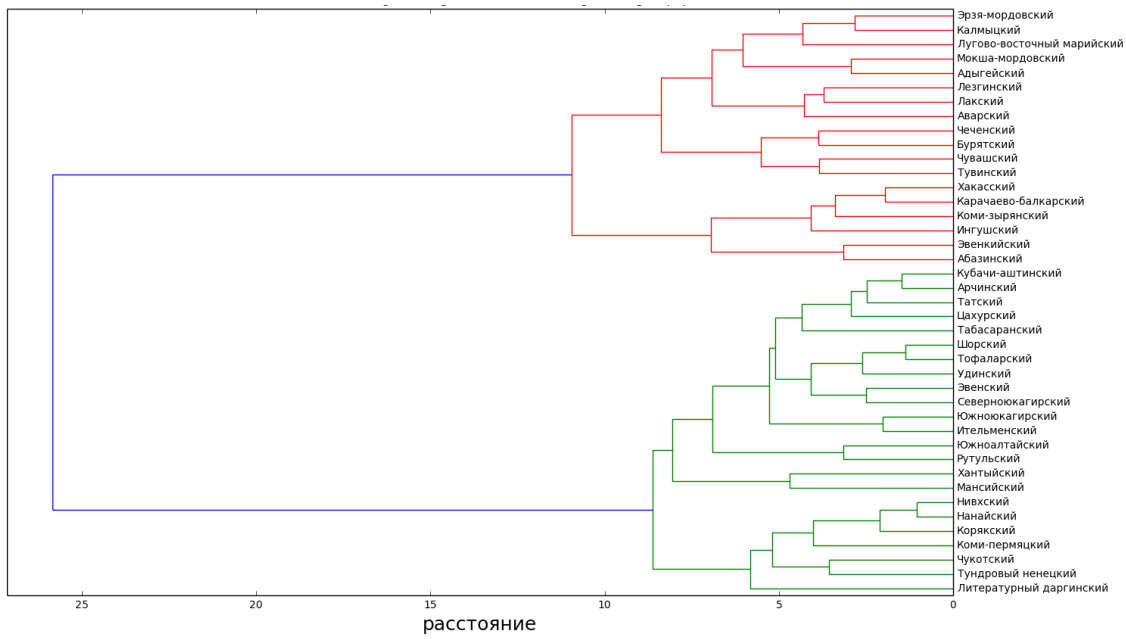
Кластерный анализ позволяет выявить группы объектов в данных, а также создать иерархические деревья, которые показывают систематические связи между группами и их прототипами. Для иерархической кластеризации важно выбрать тип расчёта расстояния между объектами. Наиболее распространённым является расстояние Уорда, его же советует использовать и профессор К. В. Воронцов в лекциях по машинному обучению (Воронцов 2015). Для оценки качества иерархической кластеризации был рассчитан кофенетический коэффициент корреляции, который является мерой того, насколько результирующая дендрограмма сохраняет попарные расстояния между оригинальными немоделированными объектами. Чем выше этот коэффициент, тем лучше модель (Teknomo 2009).

Наибольшего кофенетического коэффициента корреляции со значением 0.70 нам удалось достичь при использовании нормированных интернет-данных, вконтакте-данных (всех кроме общего числа постов), финансовых данных по титульному региону распространения за самые последние года, всех данных по печатным изданиям на малом языке. Получившаяся дендрограмма показана на рис. 15.

По дендрограмме (см. рис. 15) видно, что все языки разбились на две большие группы в 18 и 23 малых языка. Примечательно, что близкие генетически коми-зырянский и коми-пермяцкий попали в две разные группы. Это объясняется тем, что по параметрам онлайн-представленности (интернет и вконтакте данные) коми-пермяцкий «сильно проигрывает» коми-зырянскому – в 20 раз меньше токенов в интернете, в 4.5 раза меньше сообществ и в 5.5 раз меньше токенов ВКонтакте, примерно такая же разница между языками и по печатным изданиям, однако почти по всем финансовым параметрам коми-пермяцкий в 2 раза опережает коми-зырянский.

В целом, это и есть главный принцип различия двух кластеров: в верхнем

Рис. 15: Дендрограмма иерархической кластеризации малых языков на основе онлайн и оффлайн параметров



красном кластере собраны языки с большими значениями онлайн-параметров и параметров по печатным изданиями, а в нижнем зелёном кластере наоборот языки, у которых низкие значения по этим параметрам. Чем меньше расстояние между двумя языками, тем более схожи значения онлайн-параметров у этих языков. При этом может срабатывать своеобразная «компенсация» одних онлайн-параметров в другие. Например, тувинский и чувашский – лидеры среди выкачанных языков по количеству сообществ «ВКонтакте», и то, что всех сайтов (350) и сайтов на малом языке (6) на тувинском почти в два раза меньше, чем на чувашском (661, 13, соответственно), «компенсируется» тем, что сообществ на тувинском (136) почти в два раза больше, чем на чувашском (87). У бурятского с чеченским одни из самых высоких значений по количеству всех сайтов (679, 730) и одинаковые значения по количеству сайтов на малом языке (17), различие значений по выкаченным сообщетсвам ВКонтакте (60 и 17, соответственно) кажется не таким большим, если принимать во внимание количество выкачанных постов (68860, 49401) и токенов (531115, 472301) на малом языке. Всё то же самое может быть сказано и про нижний зелёный кластер: кубачи-аштинский и арчинский, имеющие наименьшее расстояние между друг

другом, являются одними из наименее представленных в интернете языков (4 сайта, 12 веб-страниц и 2 сайта, 3 веб-страницы), совсем не присутствуют в социальной сети, у нас нет информации о каких-либо печатных изданиях на этих языках, а также у этих языков полностью совпадают финансовые показатели для титульного регионааа распространения, т. к. на обоих языках разговаривают в Дагестане.

В целом, результаты иерархической кластеризации, отражённые на дендрограмме на рис. 15, совпадают с нашей интуицией, и 41 малый язык, для которого собрана информация об онлайн-представленности с использованием дополнительных признаков о количестве печатной продукции и финансовых показателях титульного региона распространения, можно разделить на два кластера чуть более представленных в онлайн и офлайн среде языков и чуть менее представленных.

4.5. Веб-графы интернетов малых языков

Отличительной особенностью интернет-данных являются ссылки. Получив с каждой скачанной страницы гиперссылки, можно построить, так называемый, веб-граф – граф, вершинами которого являются сайты, а рёбрами – гиперссылки между этими сайтами. На основе имеющихся данных был построен ориентированный взвешенный веб-граф для национальных интернетов по следующим принципам:

- вершинами графа являются малые языки, для которых был найден хотя бы 1 сайт на малом языке.
- ребро проставляется в том случае, если в выкачанных данных есть ссылка с одного сайта на другой сайт из списка сайтов на малом языке, т. е. оба сайта должны входить в один из списков сайтов на малом языке, но при этом сайт, на который ссылаются, может быть не выкачан. Поэтому в граф попадают языки типа башкирского и татарского.
- в графе разрешено ссылаться на себя, например, учитываются ссылки с лезгинского сайта www.gusarpress.com на лезгинский же сайт alamjurnal.com

(ссылки на тот же сайт, конечно, запрещены)

- в весе графа отражено количество ссылок на сайты из списка, при этом если www.gusarpress.com ссылается на alamjurnal.com, а alamjurnal.com на www.gusarpress.com, то это засчитывается как вес равный 2.

Получившийся граф показан на рис. 16. Синим цветом шрифта отмечена вершина с максимальным коэффициентом центральности. В граф не вошли лакский и карачаево-балкарский интернеты, т.к. оказалось, что с сайтов на этих языках не ссылаются ни на сайты на других языках, ни на сайты на этом языке и более того на сайты на лакском и карачаево-балкарском не ссылаются с сайтов на других языках.

Любопытным кажется и устройсво национального интернета каждого языка в отдельности. Для национального интернета малого языка мы построили ориентированные невзвешенные графы. Пример графа для эрзя-мордовского приведён на рис. 17, графы для остальных языков можно посмотреть в Приложении F.

Для каждого графа были рассчитаны следующие метрики: количество вершин и рёбер, среднее количество входящих и исходящих рёбер, вершины с максимальной степенью центральности (учитываются входящие и исходящие рёбра), коэффициент ассортативности и количество сильно связных и слабо связных компонент. Значения всех метрик приведены в Приложении F.

Оказалось, что ни один построенный граф не является сильно связным, зато почти все (7 из 10) графы национальных интернетов слабо связные. Однако, общий граф национальных интернетов не является слабо связным, т. к. лезгинский национальный граф образовал отдельную, не связанную с другими национальными интернетами, компоненту. Также не являются слабо связными графы для лезгинского, чеченского и тувинского. Графы лезгинского и тувинского интернетов распались на две примерно равные по количеству вершин компоненты, граф чеченского состоит из одной компоненты на 8 вершин и двух компонент поменьше.

Коэффициент ассортативности показал, что те графы, для которых можно

Рис. 16: Веб-граф национальных интернетов

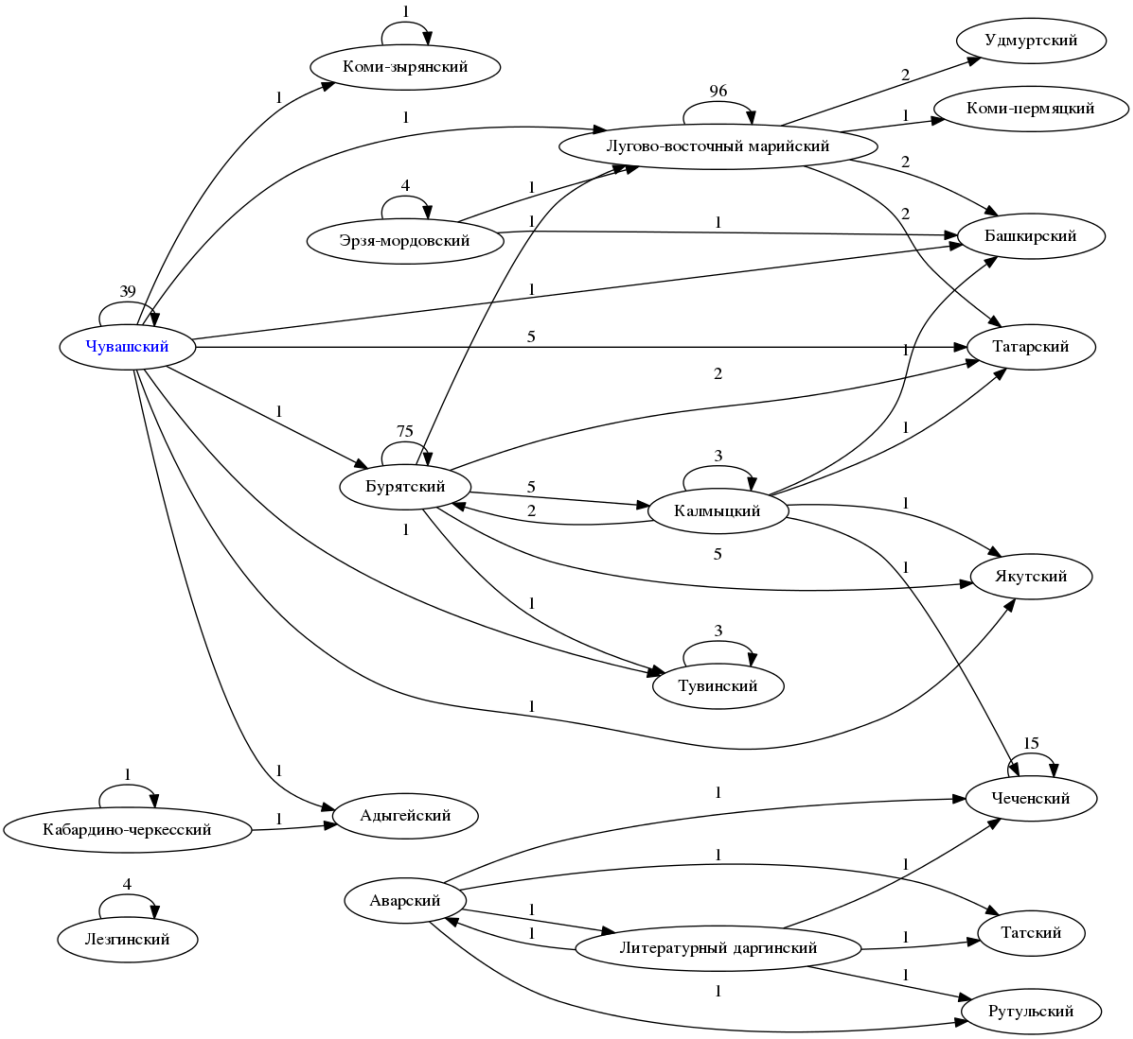


Рис. 17: Веб-граф эрзя-мордовского интернета

рассчитать этот коэффициент, дисассортативны (значение коэффициента указано в скобках). Это графы бурятского (-0.28), чувашского (-0.51), чеченского (-0.73) и лугово-восточного марийского (-0.15) национальных интернетов. Общий граф национальных интернетов ассоративен (0.21), а граф эрзя-мордовского национального интернета не имеет выраженной ассортативности (0.0). В графах национальных интернетов не так много вершин и отрицательный коэффициент ассортативности в данном случае показывает, что, как правило, в графе есть один какой-то сайт, с которого ссылаются на другие и на который ссылаются с других сайтов, и «конкурентов» у этого сайта нет. Это хорошо видно на графе чувашского интернета, в котором степень вершины chuvash.org равна 20, а следующая по количеству связей вершина chgign.ru имеет лишь степень 10. Общий же граф национальных интернетов ассортативен и это выражено в том, что языки, которые больше всего ссылаются на другие языки, ссылаются и на друг друга тоже. Отметим, что в наших данных нет самых больших языков – бурятского, татарского, якутского, удмуртского. Можно ожидать, что с добавлением этих языков, коэффициент ассортативности общего графа лишь увеличится.

Вершины с максимальной степенью центральности (далее – центральные вершины) на всех графах обозначены синим цветом шрифта. Любопытным оказалось, что не всегда сайт-вершина с максимальной степенью центральности является самым большим по количеству выкачанных страниц и токенов на малом языке. Так для чувашского (chuvash.org), лугово-восточного марийского (mariuver.wordpress.com) и тувинского (shyn.ru) графов центральные вершины самые большие и по страницам, и по токенам среди всех сайтов на этом малом языке. Однако, для бурятского (minkultrb.ru), чеченского (chechnyatoday.com) и эрзя-мордовского (erzia.saransk.ru) графов центральные узлы оказываются лишь в топ-5 по количеству выкачанных страниц и токенов. Удивительным оказалось, что и на общем графе национальных интернетов центральным узлом является чувашский язык – язык, который не является самым представленным в интернете ни по общему количеству сайтов, ни по количеству токенов, веб-страниц или сайтов на малом языке. Примечательно, что в основном на сайты

на других языках ссылаются с чувашского сайта chuvash.org, тогда как в чувашском национальном интернете много ссылок ведут на этот сайт.

Отметим, если рассчитывать «важность» вершины по алгоритму pagerank (*pagerank numpy* б.г.), то совпадать эти вершины с центральными будут лишь в половине случаев – для чеченского, чувашского и бурятского графов. Для общего графа «важным» будет чеченский язык, по рис. 16 видно, что это вершина, которая связывает два самых крупных подграфа общего графа. Для графов тувинского и эрзя-мордовского интернетов алгоритм pagerank счёл важными самые крайние вершины, которые ни на кого не ссылаются, но на которые ведёт ссылка с других вершин (goloserzi.ru для эрзя-мордовского, baitaiga.ru для тувинского). Для лугово-восточного марийского важной была выбрана вершина mari-el.name – вершина, у которой много входящих рёбер, но мало исходящих.

Рассмотрев веб-графы интернетов малых языков, мы узнали, что большинство национальных интернетов представляют из себя слабо связный граф. Это говорит о том, что если бы интернет был неориентированным графом, то с любого сайта национального интернета можно было бы попасть на любой другой сайт интернета на том же малом языке. По большей части веб-графы национальных интернетов дисассортативны, что объясняется скорее не тем, что «важные» сайты национальных интернетов ссылаются друг на друга через «посредников», а тем, что зачастую существует один «самый важный» сайт, степень которого может в два раза превышать степени всех остальных вершин-сайтов. Однако, общий граф национальных интернетов является ассортативным, что говорит о том, что языки, которые больше всего ссылаются на другие языки, ссылаются и друг на друга. Отметим также, что вершины с максимальной степенью центральности не всегда являются самыми большими сайтами национального интернета, что не позволяет заранее угадать, какой сайт окажется для национального интернета «самым главным».

5. Заключение

В работе мы рассмотрели представленность миноритарных языков в интернет-пространстве. Наиболее «самостоятельными», т.е. имеющими наибольшее

количество сайтов на малом языке являются башкирский, татарский, якутский, удмуртский и лугово-восточный марийский языки. В то же время 19 из 49 рассмотренных языков не имеют сайта на малом языке вовсе. Регистрационные данные, полученные по доменным именам, на которых располагаются такие сайты, свидетельствуют о том, что частные лица и организации регистрируют примерно равное количество доменов, а среди организаций преобладают СМИ и государственные структуры. Наблюдения по конкретным языкам приведены в тексте работы.

Нам не удалось выяснить достоверные причины всплеска и дальнейшего падения регистраций доменов малых языков в 2012-м году. По данным на сайтах крупенейших регистраторов доменов можно предположить влияние территориальной экспансии и различные маркетинговые акции. Однако, эксперты утверждают, что на выборке меньше 1000 доменов сложно говорить о причинно-следственных связях.

К сожалению, географических данных в регистрационных записях доменных имён крайне мало, более того часть из них искажена специальными сервисами защиты данных, а данные, полученные из баз георапределения IP-адресов зачастую крайне не точны и предоставляют лишь информацию о физическом расположении сервера хостинг-провайдера. Поэтому никаких содержательных выводов о георасположении сайта на основе этой информации сделать нельзя, можно только иметь в виду, что согласно регистрационным данным сайты на малых языках за редким исключением для чеченского языка располагаются либо в столицах титульных регионов распространения языка, либо в Москве. А согласно георапределению IP-адресов подавляющее большинство сайтов (точнее серверов, на которых «хостятся» эти сайты) определяются как расположенные в одной из «трёх столиц»: столице титульного региона распространения языка, Москве или Санкт-Петербурге.

Для анализа данных и поиска связей и зависимостей был использован достаточно широкий набор параметров о малых языках: от количества веб-страниц, сообществ и токенов до финансовых показателей титульного региона распространения языка. Анализ корреляций помимо очевидных связей между призна-

ками за разные года или полученных из одного источника (например, Википедии) показал наличие сильных связей между всеми имеющимися у нас данными о текстовой представленности языка в интернете, что, возможно, говорит об одинаковом уровне интернет-представленности малых языков в разных онлайн-источниках. Также было замечено наличие сильной связи интернет-данных с числом носителей за 2002-ой год и то, что число носителей за 2002-ой год коррелирует с гораздо большим числом онлайн-параметров, чем число носителей за 2010-ый год, что может говорить о том, что в цифровом мире язык развивается по своим, отличным от реального мира законам.

Регрессионный анализ показал, что с помощью многофакторной линейной регрессии можно построить значимую модель интернет-данных от офлайн-данных, объясняющую до 63% наблюдений. Однако, такая модель сильно улучшается с добавлением онлайн-параметров из других источников и начинает объяснять уже до 75% интернет-данных. Такая модель показывает неплохую предсказательную способность и может быть использована для приблизительной оценки количества сайтов на малом языке. В качестве офлайн-параметров в модели были использованы число носителей за 2010 год, прирост населения на 2014 год, расходы региона за 2014 год, число газет на 1996 год. Из онлайн-параметров использовались количество статей в языковом разделе интернет-энциклопедии «Википедия» за 2016 год и количество токенов на языке в социальной сети «ВКонтакте». Мы предполагаем, что при использовании данных за один и тот же срез времени модель может улучшиться.

С помощью схожих параметров (онлайн-данные, финансовые показатели региона и данные по печатным изданиям) нам удалось иерархически раскластеризовать малые языки и представить результаты кластеризации в виде дендрограммы, на которой отчётливо видны два кластера более и менее представленных в реальной и цифровой среде малых языков, а языки с наиболее схожими параметрами расположены «рядом» (в соседних листьях дендрограммы).

В работе также были изучены свойства веб-графов национальных интернетов. Оказалось, что в большинстве своём национальные интернеты являются слабо связными дисассортативными графами. Из первого свойства следует,

что если бы ссылки в интернете вели в обоих направлениях (т.е. с каждого сайта, на которой сослались, была бы ссылка на сайт, с которого сослались), то с любого сайта национального интернета можно было бы попасть на любой другой сайт на этом малом языке. Дисассортативность графов, однако, объясняется тем, что в графе каждого национального интернета обычно присутствует ровно один «важный сайт», у которого нет «конкурентов». Что, пожалуй, должно заставить сообщество задуматься об обеспечении сохранности таких сайтов. Удивительным оказалось и то, что вершины графов с максимальной степенью центральности, т. е. наиболее значимые сайты, не всегда являются самыми большими сайтами конкретного национального интернета.

Общий же граф национальных интернетов не является ни слабо связным, ни дисассортативным. Есть языки (лакский и карачаево-балкарский), чьи интернеты вообще не были представлены в общем графе, т. к. они никак не контактируют (не ссылаются) с национальными интернетами других языков. Отличился так же и лезгинский, ссылки из национального интернета которого ведут только обратно в лезгинский национальный интернет. В остальном же ассортативность графа указывает на то, что национальные интернеты в некотором смысле «держатся» на нескольких крупных языках, которые ссылаются и на интернеты на других языках, и на друг друга. Самым «важным» языком оказался чувашский, которому ведомо о существовании интернетов на 8 разных языках. Однако, мы не отрицаем возможности того, что лидер может смениться с добавлением более «крупных игроков» (удмуртского, башкирского, татарского или якутского).

Подводя итог скажем, что мы создали количественное описание национальных интернетов для более чем 40 языков, построили модель на основе офлайн и онлайн данных по малому языку, с помощью которой можно приблизительно предсказать, будет ли присутствовать малый язык не из выборки в интернете, проанализировали регистрационные данные сайтов на малых языках, а также изучили графовые свойства интернетов миноритарных языков и упорядочили языки по представленности в онлайн и офлайн среде. В заключении отметим, что самые крупные, по нашим данным, национальные интернеты уже успели

привлечь внимание авторов как качественных, так и количественных исследований.

Список использованных источников и литературы

About Privacy Protection. — URL: http://privacyprotect.org/about-privacyprotection/.

Arkhangeskiy T., *Belyaev O.*, *Vydrin A.* The Creation of Large-Scale Annotated Corpora of Minority Languages using UniParser and the EANC platform. // COLING (Posters). — 2012. — С. 83—92.

Baroni M., *Bernardini S.*, *Ferraresi A.*, *Zanchetta E.* The WaCky wide web: a collection of very large linguistically processed web-crawled corpora // Language resources and evaluation. — 2009. — Т. 43, № 3. — С. 209—226.

Comrie B. The languages of the Soviet Union. — CUP Archive, 1981.

Das B. Social media rescues dying Indian languages. — 2013. — URL: http://www.aljazeera.com/indepth/features/2013/11/social-media-rescues-dying-indian-languages-201311277047252312.html.

Domain Privacy. — URL: https://www.dynadot.com/domain/privacy.html.

Domaintools. — URL: http://research.domaintools.com/statistics/ip-addresses/.

Endangered Languages. — URL: https://www.ethnologue.com/endangered-languages.

Frost J. Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables. — 2013. — URL: http://blog.minitab.com/blog/adventures-in-statistics/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables.

Indigenous Blogs. — URL: http://indigenoustweets.com/blogs/.

Indigenous Tweets. — URL: http://indigenoustweets.com/.

ip-api. — URL: http://ip-api.com/docs/api:json.

Kornai A. Digital language death // PloS one. — 2013. — Т. 8, № 10. — e77056.

pagerank numpy. — URL: https://networkx.github.io/documentation/networkx-1.9.1/reference/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank_numpy.html.

REG.RU подводит итоги 2012 года. — 2012. — URL: https://www.reg.ru/company/year_summary/2012.

Scannell K. P. The Crúbadán Project: Corpus building for under-resourced languages // Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop. T. 4. — 2007. — С. 5—15.

StatsModels. — URL: http://statsmodels.sourceforge.net/devel/.

Teknomo K. Cophenetic Correlation Coefficient. — 2009. — URL: http://people.revoledu.com/kardi/tutorial/Clustering/Cophenetic.htm.

Zaydelman L., *Krylova I.*, *Orekhov B.*, *Stepanova E.* Languages of Russia: Using Social Networks to Collect Texts //. T. 573. — St. Petersburg, в печати. — С. 1—8.

Бослаф С. Статистика для всех / пер. П. А. Волкова, И. М. Флямер, М. В. Либерман, А. А. Галицына. — Москва : ДМК Пресс, 2015. — С. 586.

Бурятский корпус. — 2012. — URL: http://web-corpora.net/BuryatCorpus/search/index.php.

В Туле приступил к работе региональный представитель RU-CENTER. — 2012. — URL: https://www.nic.ru/news/2012/regional_tula.html.

Воронцов К. В. Методы кластеризации. — 2015. — URL: http://www.machinelearning.ru/wiki/images/2/28/Voron-ML-Clustering-slides.pdf.

Всероссийский доменный марафон. — URL: http://marafon.domeniny.ru/.

Доменный марафон шагает по российским регионам. — 2012. — URL: https://www.nic.ru/news/2012/marafon.html.

Замятин К., *Пасанен А.*, *Саарикиви Я.* Как и зачем сохранять языки народов России // Хельсинки: Vammalan Kirjapaino Oy, Vammala. — 2012.

Калмыцкий корпус. — URL: http://web-corpora.net/KalmykCorpus/search/index.php.

Корпус лезгинского литературного языка. — 2011. — URL: http://dag-languages.org/LezgianCorpus/search/index.php.

Корпус удмуртского языка. — 2015. — URL: http://web-corpora.net/UdmurtCorpus/search/.

Корпусы ИЭА РАН. — URL: http://corpora.iea.ras.ru/corpora/.

Корреляция. — URL: https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D1%8F.

Лучшие хостинги. — URL: http://hosting.effective-it.ru/luchshie-hostingi.html.

Малые языки Сибири: наше культурное наследие. — URL: http://siberian-lang.srcc.msu.ru/ru.

Махмутов З. А. К вопросу о функциональном взаимодействии русского и национального языков в татарском сегменте интернета // Русский язык и новые технологии. — Москва, 2014. — С. 214—224.

Орехов Б. В., *Галлямов А. А.* Башкирский интернет: лексика и прагматика в количественном аспекте // Компьютерная лингвистика и Интеллектуальные технологии: Труды Международной Конференции «Диалог 2012». — Бекасово, 2013. — С. 502—509.

Орехов Б. В., *Махмутов З. А.* Методологические и практические аспекты изучения национальных интернетов в России // Русский язык и новые технологии. — Москва, 2014. — С. 189—199.

Орехов Б. В., *Решетников К. Ю.* К оценке Википедии как лингвистического источника // Современный русский язык в интернете: Языки славянской культуры. — Москва, 2014. — С. 310—321.

Пишлёгер К. Бесермяне в интернете: социальные сети как шанс для сохранения родного языка? — 2013a. — URL: http://www.academia.edu/8306998/%D0%91%D0%B5%D1%81%D0%B5%D1%80%D0%BC%D1%8F%D0%BD%D0%B5_%D0%B2_%D0%B8%D0%BD%D1%82%D0%B5%D1%80%D0%BD%D0%B5%D1%82%D0%B5_%D1%81%D0%BE%D1%86%D0%B8%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B5_%D1%81%D0%B5%D1%82%D0%B8_%D0%BA%D0%B0%D0%BA_%D1%88%D0%B0%D0%BD%D1%81_%D0%B4%D0%BB%D1%8F_%D1%81%D0%BE%D1%85%D1%80%D0%B0%D0%BD%D0%B5%D0%BD%D0%B8%D1%8F_%D1%80%D0%BE%D0%B4%D0%BD%D0%BE%D0%B3%D0%BE_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0.

Пишлёгер К. «Бурановские бабушки» на YouTube: роль социальных сетей в сохранении удмуртской идентичности и языка в условиях глобализации //. — Ижевск : УИИЯЛ УрО РАН, 2013b. — С. 171—175.

Пишлёгер К. Запис(к)и из Муржол Underground: Super удмурты в Cyberspace //. — Глазов, 2014. — С. 56—59.

Российский рынок доменов. — URL: http://stat.nic.ru/.

Российский рынок доменов. IP-адресация, хостинг, сервисы. Итоги 2010 года. — 2010. — URL: http://stat.nic.ru/bull_10.pdf.

Российский рынок доменов. Регистрация новых доменов. — URL: http://stat.nic.ru/reports/whist-ru/newreg.html.

Рунет в марте 2008 года: домены, хостинг, география сайтов. — 2008. — URL: http://www.rukv.ru/runet-2008.html.

Самойленко И. Ю. Автоматическое определение языка по тексту: языки России и ближнего зарубежья. — 2015.

Сахарных Д. М. Удмуртский национальный интернет: что мешает развитию? — 2006. — URL: http://udmurt.info/pdf/texts/udmint.pdf.

Сахарных Д. М. Новое в развитии удмуртского национального интернета: Зима–осень 2008 г. — 2008. — URL: http://udmurt.info/pdf/texts/udmurtnet-zima-osenj-2008.pdf.

Татарский национальный корпус «Туган тел». — URL: http://web-corpora.net/TatarCorpus/search/index.php.

Центр мультимедийных языковых архивов LangueDOC. — URL: http://www.philol.msu.ru/~languedoc/rus/index.php.

Чеченская диаспора. — URL: https://ru.wikipedia.org/wiki/%D0%A7%D0%B5%D1%87%D0%B5%D0%BD%D1%81%D0%BA%D0%B0%D1%8F_%D0%B4%D0%B8%D0%B0%D1%81%D0%BF%D0%BE%D1%80%D0%B0.

Языки народов России. Красная книга / под ред. В. П. Нерознак. — Москва : Academia, 2002. — С. 392.

Языки России. — URL: http://web-corpora.net/minorlangs.

Яндекс.Переводчик. — URL: https://translate.yandex.ru/about.

Приложения

А. Список больших сайтов, с которых скачивались только конкретные страницы

salda.ws
song5.ru
www.stihi.ru
stihi.ru
proza.ru
www.gumer.info
lib.ru
lib.aldebaran.ru
royallib.com
www.big-library.info
www.lib.mn
www.livelib.ru
www.epubbooks.ru
www.electroniclibrary21.ru
ihtik.lib.ru
www.krelib.com
libshara.com
electrolib.com
www.publiclibrary.ru
mexalib.com
ModernLib.ru
TheLib.ru
librs.net
library.ua
lib.align.ru
youtube.com
www.youtube.com
ytube.com.ua
livejournal.ru
livejournal.com
www.jooov.net
otvet.mail.ru
www.proza.ru
www.chitalnya.ru
www.livejournal.com
umorina.info
filka.info
myzlo.info
webkind.ru
flatik.ru
multiurok.ru
musicov.net
muzpoisk.net
www.adamalla.com
mp3-pesnja.com
lyricstranslate.com
pesenok.ru
get-tune.net
mp3ll.net
www.ashkimsin.ru
iplayer.fm
online-song.net
music.bibib.kz
OneSong.ru

musicaplays.ru
www.mus3.ru
muzozo.ru
tekstovnet.ru
songspro.ru
kniga.seluk.ru
x-minus.org
prezi.com
presentacid.ru
websta.me
androidmafia.ru
maidanvideo.ru
wikibit.me
mozhgatv.ru
vkinotik.net
www.ruclip.com
oprb.ru
teksty-pesenok.ru
annot.ru
festival.1september.ru
www.openclass.ru
instaliga.com
video-pro-koshak.ru
kot-saymona.ru
grammio.com
mp3-pesni.net
Mozhgatv.ru
forum.na-svyazi.ru
abunda.ru
risovach.ru
textpesni2.ru
sportonline.biz
instagr.in
tv-show-best.ru
song-txt.ru
videolike.org
vinteresah.com
muzlan.com
www.megalyrics.ru
Zebroid.tv
knigobolen.ru
www.ps3youtube.com
picbi.com
igbox.co
ilovegram.com
crabme.pp.ua
forum.meta.ua
www.litmir.me
sprosi.name
www.podslyshano.ru
ww.realtorspb.com
www.edinros21.ru
www.archives21.ru
forum.cheb.ru
j-vaynah.ru
fcterek.com
tubewatch.ru
transcriptase3.rssing.com
eocshool.ucoz.ru
pinhere.me
tolkuy.ru
vopy.ru

www.instawebgram.com

prevodi-na-pesni.com

myphone.info

www.bible.is

lib.rus.ec

B. Дополнительные графики распределения регистраций доменов

Рис. 18: Общее распределение регистраций новых доменов в зоне .SU

Рис. 19: Распределение регистраций новых доменов в зоне .SU за 2012 год

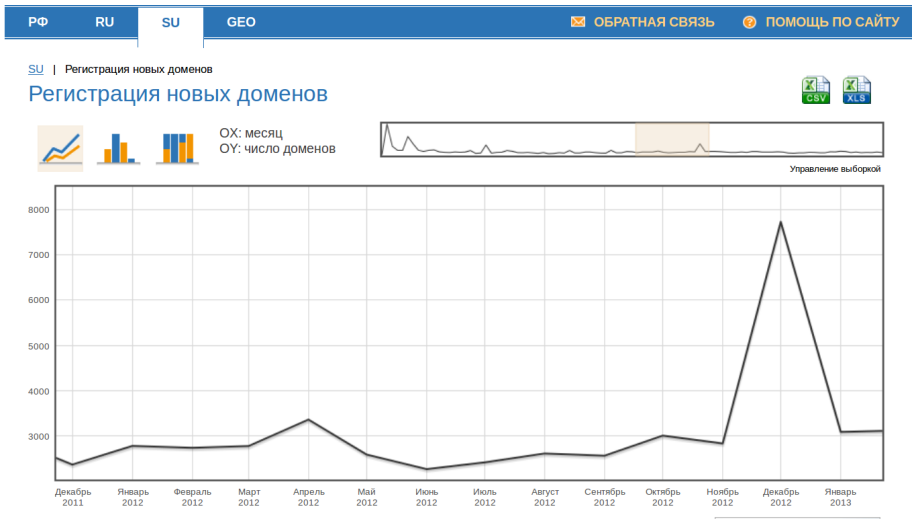


Рис. 20: Общее распределение регистраций новых доменов в зоне .RF

Рис. 21: Распределение регистраций новых доменов в зоне .RF за 2012 год

C. Таблицы регистраций доменов по городам

Таблица 9: Организациями

none	166
washington	3
kazan	3
istanbul	2
bishkek	2
moscow	2
yakutsk	2
groznyi	2
makhachkala	2
ufa	1
tbilisi	1
kasumkent	1
saransk	1
london	1
saratov	1
astana	1
nalchik	1
st. petersburg	1
syktyvkar	1
örebro	1
brussel	1
nobby beach	1

Таблица 10: Частными лицами

none	171
moscow	3
ufa	3
ulaan-ude	2
istanbul	2
orem	2
izhevsk	2
perm	1
brea	1
helsinki	1
starnberg	1
syktyvkar	1
maykop	1
stavropol	1
nijniy novgorod	1
tamra	1
inzhich-chukun	1
france	1
kazan	1
ub	1
yablonovskiy	1
sunnyvale	1
san mateo	1
makhachkala	1
shali	1

D. Таблицы распределения сайтов по городам на основе IP-адресов

Организациями

moscow	54
saint petersburg	35
yakutsk	16
kazan	13
ufa	12
izhevsk	8
gunzenhausen	7
cheboksary	7
none	5
kyzyl	3
jõhvi	3
irkutsk	2
yoshkar-ola	2
novosibirsk	2
syktyvkar	2
roubaix	1
poselok № 1	1
naryan-mar	1
nuremberg	1
orenburg	1
västerås	1
odesa	1
krasnoyarsk	1
falkenstein	1
maykop	1
tallinn	1
kyyiv	1

Частными лицами

moscow	78
saint petersburg	41
ufa	11
san francisco	8
provo	5
irkutsk	4
khoroshëvskaya	3
amsterdam	3
gunzenhausen	3
brisbane	3
yakutsk	3
none	3
brea	2
falkenstein	2
jõhvi	2
kazan	2
ulan-ude	2
'kharkiv	2
reutov	1
elista	1
houston	1
kyzyl	1
arzamas	1
elektrostal	1
cedar knolls	1
mountain view	1
scottsdale	1

Организациями

odintsovo	1
volga	1
scottsdale	1
pervouralsk	1
washington	1
ulan-ude	1
charlotte	1
bishkek	1
vologda	1
provo	1
brea	1
elektrostal	1
austin	1
пос. луховка	1

Частными лицами

kaliningrad	1
newcastle upon tyne	1
burlington	1
kogalym	1
thalwil	1
cheboksary	1
london	1
saratov	1
dusseldorf	1
austin	1
copenhagen	1
petah tikva	1
bourlens	1
chicago	1
nuremberg	1
ashburn	1
atlanta	1

E. Код и документация

Весь код, использовавшийся в работе, в том числе ipython-notebook с анализом данных и комметариями на русском, закоммичены в репозиторий: https://github.com/irinfox/minor_langs_internet_analysis

Краткая документация по использованию представлена в README репозитория.

F. Веб-графы национальных интернетов

Метрики для общего веб-графа национальных интернетов (рис. 22)

Количество вершин: 20

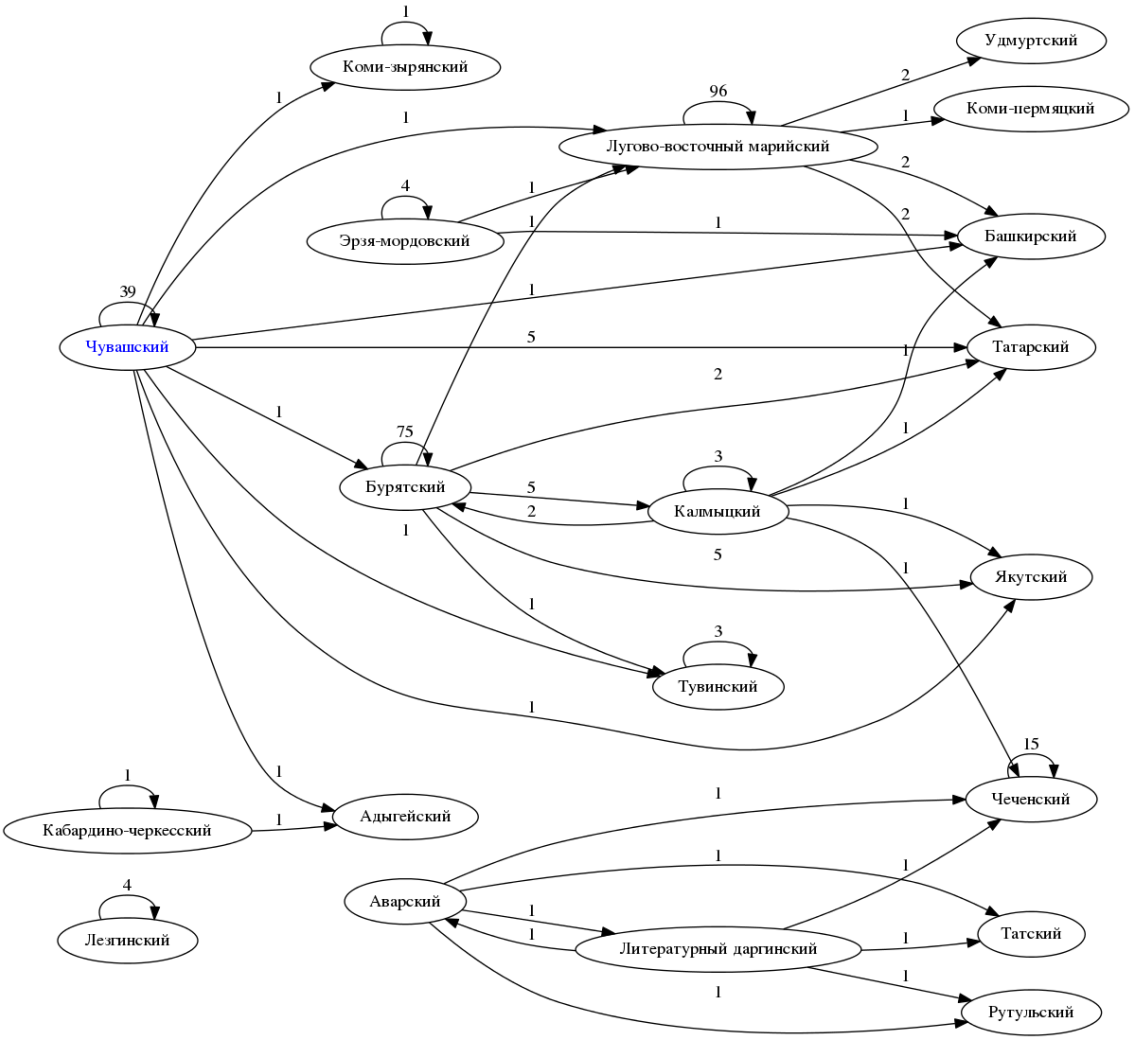
Количество рёбер: 43

Средняя полустепень захода: 2.150

Средняя полустепень исхода: 2.150

Вершина с максимальной степенью центральности: Чувашский (0.526)

Рис. 22: Общий веб-граф национальных интернетов



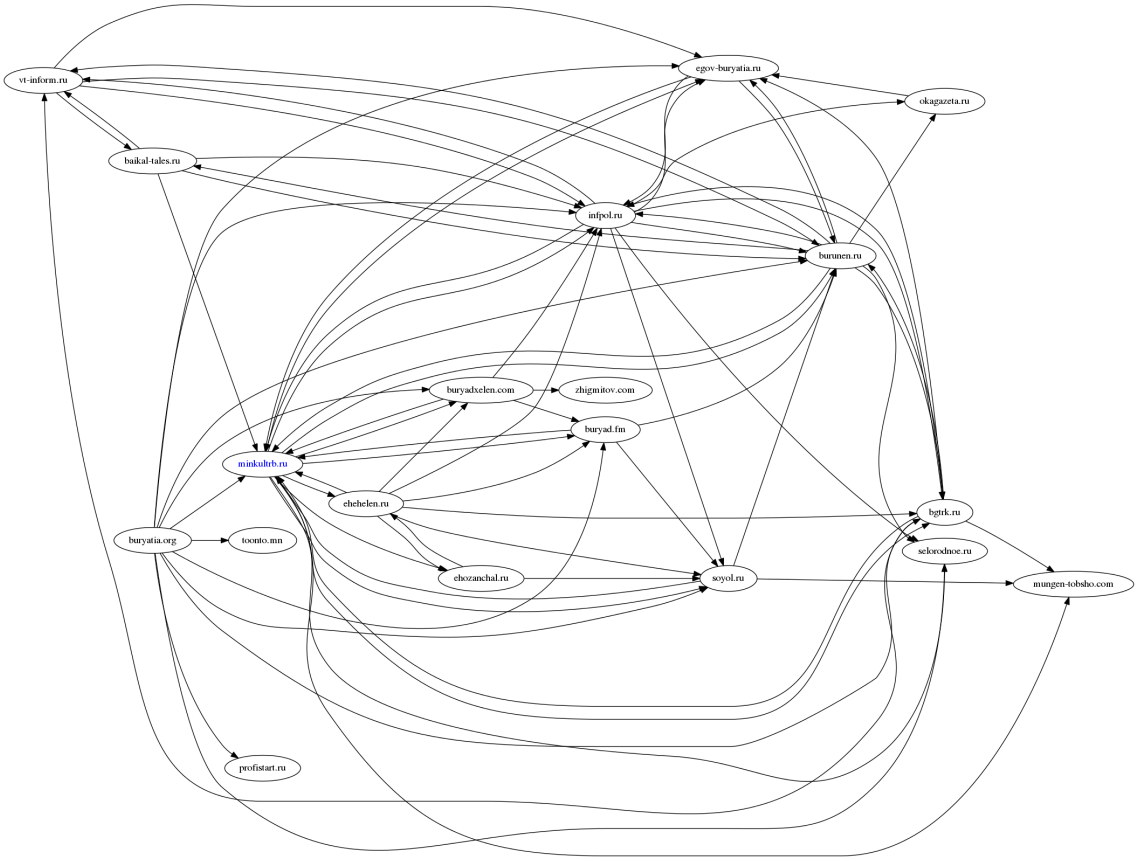
Коэффициент ассортативности: 0.208

Количество слабо связных компонент: 2

Количество сильно связных компонент: 18

Топ-5 вершин по алгоритму pagerank: Чеченский (0.169), Тувинский (0.109), Коми-зырянский (0.104), Лугово-восточный марийский (0.102), Лезгинский (0.099)

Рис. 23: Веб-граф бурятского интернета



Метрики для веб-графа бурятского интернета (рис. 23)

Количество вершин: 19

Количество рёбер: 75

Средняя полустепень захода: 3.947

Средняя полустепень исхода: 3.947

Вершина с максимальной степенью центральности: minkultrb.ru (1.166)

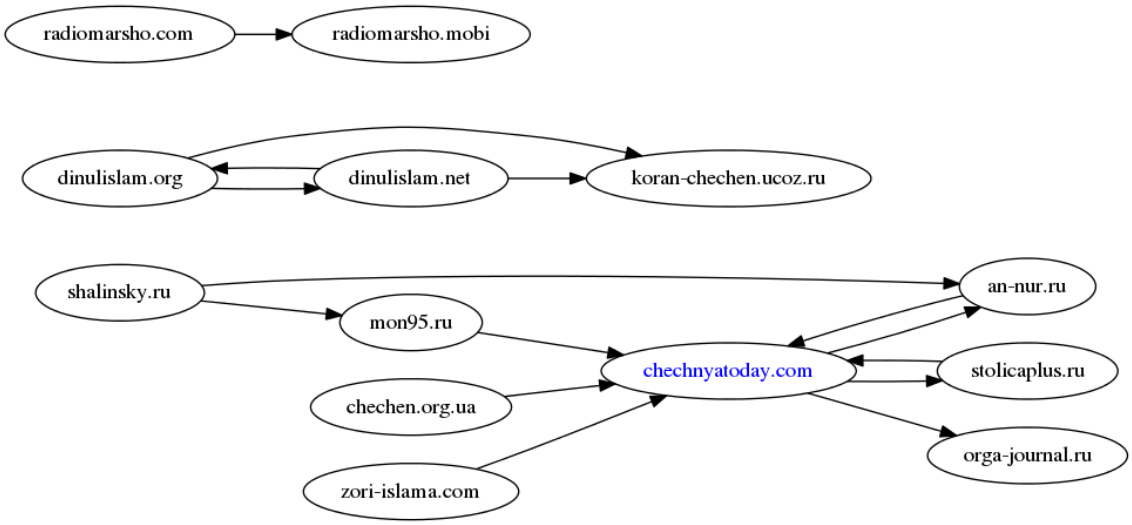
Коэффициент ассортативности: -0.285

Количество слабо связных компонент: 1

Количество сильно связных компонент: 7

Топ-5 вершин по алгоритму pagerank: minkultrb.ru (0.124), burunen.ru (0.121), infpol.ru (0.105), egov-buryatia.ru (0.102), soyol.ru (0.063)

Рис. 24: Веб-граф чеченского интернета



Метрики для веб-графа чеченского интернета (рис. 24)

Количество вершин: 13

Количество рёбер: 15

Средняя полустепень захода: 1.154

Средняя полустепень исхода: 1.154

Вершина с максимальной степенью центральности: chechnyatoday.com (0.666)

Коэффициент ассортативности: -0.726

Количество слабо связных компонент: 3

Количество сильно связных компонент: 10

Топ-5 вершин по алгоритму pagerank: chechnyatoday.com (0.303), an-nur.ru (0.124), stolicaplus.ru (0.112), orga-journal.ru (0.112), koran-chechen.ucoz.ru (0.065)

Метрики для веб-графа чувашского интернета (рис. 25)

Количество вершин: 16

Количество рёбер: 39

Средняя полустепень захода: 2.438

Средняя полустепень исхода: 2.438

Вершина с максимальной степенью центральности: chuvash.org (1.333)

Коэффициент ассортативности: -0.505

Количество слабо связных компонент: 1

Количество сильно связных компонент: 10

Топ-5 вершин по алгоритму pagerank: chuvash.org (0.168), ntrk21.ru (0.075), chgign.ru (0.074), suvar.su (0.065), chuvbook.ru (0.064)

Метрики для веб-графа кабардино-черкесского интернета (рис. 26)

Рис. 25: Веб-граф чувашского интернета

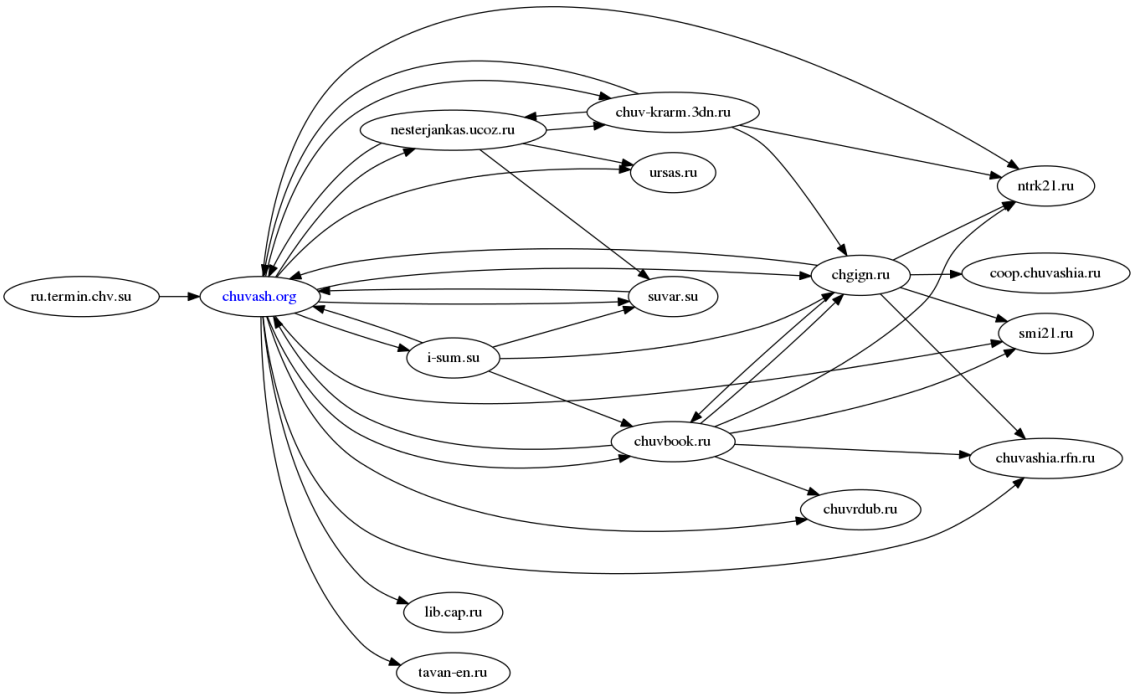
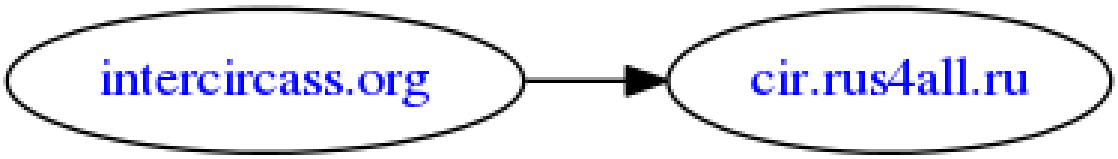


Рис. 26: Веб-граф кабардино-черкесского интернета



Количество вершин: 2

Количество рёбер: 1

Средняя полустепень захода: 0.500

Средняя полустепень исхода: 0.500

Вершины с макс. степенью центральности: cir.rus4all.ru (1.0), intercircass.org (1.0)

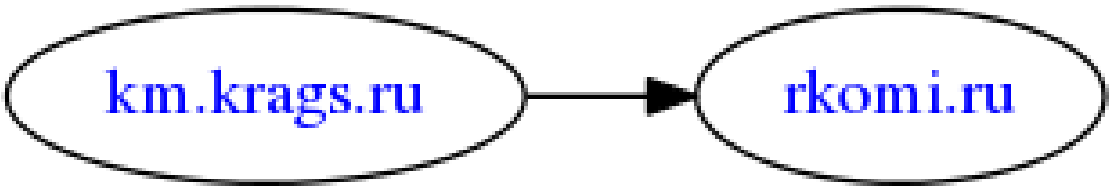
Коэффициент ассортативности: nan

Количество слабо связных компонент: 1

Количество сильно связных компонент: 2

Топ-5 вершин по алгоритму pagerank: cir.rus4all.ru (0.649), intercircass.org (0.351)

Рис. 27: Веб-граф коми-зырянского интернета



Метрики для веб-графа коми-зырянского интернета (рис. 27)

Количество вершин: 2

Количество рёбер: 1

Средняя полустепень захода: 0.500

Средняя полустепень исхода: 0.500

Вершины с макс. степенью центральности: km.krags.ru (1.0), rkomi.ru (1.0)

Коэффициент ассортативности: nan

Количество слабо связных компонент: 1

Количество сильно связных компонент: 2

Топ-5 вершин по алгоритму pagerank: rkomi.ru (0.649), km.krags.ru (0.351)

Метрики для веб-графа лезгинского интернета (рис. 28)

Количество вершин: 5

Количество рёбер: 4

Средняя полустепень захода: 0.800

Средняя полустепень исхода: 0.800

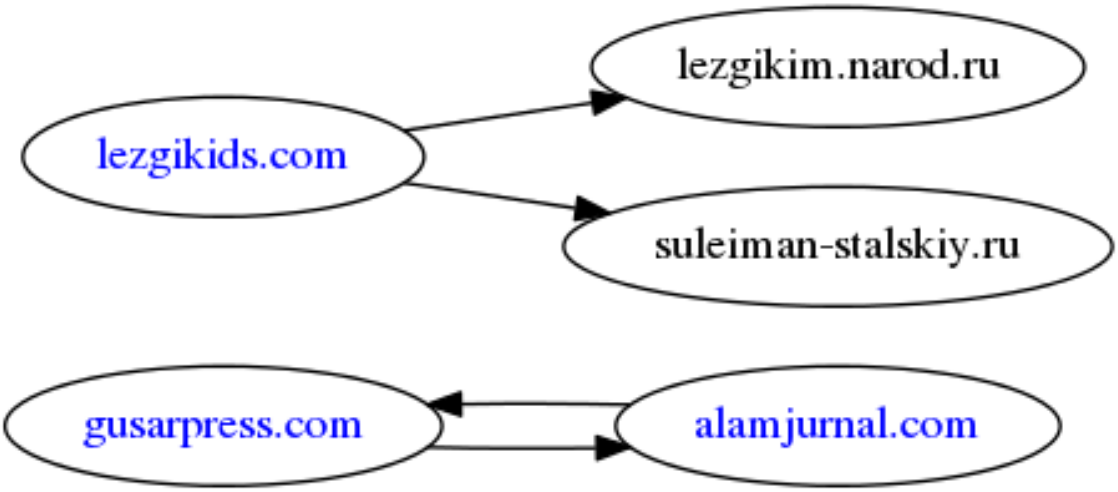
Вершина с макс. степенью центральности: gusarpress.com (0.5), lezgikids.com (0.5), alamjurnal.com (0.5)

Коэффициент ассортативности: nan

Количество слабо связных компонент: 2

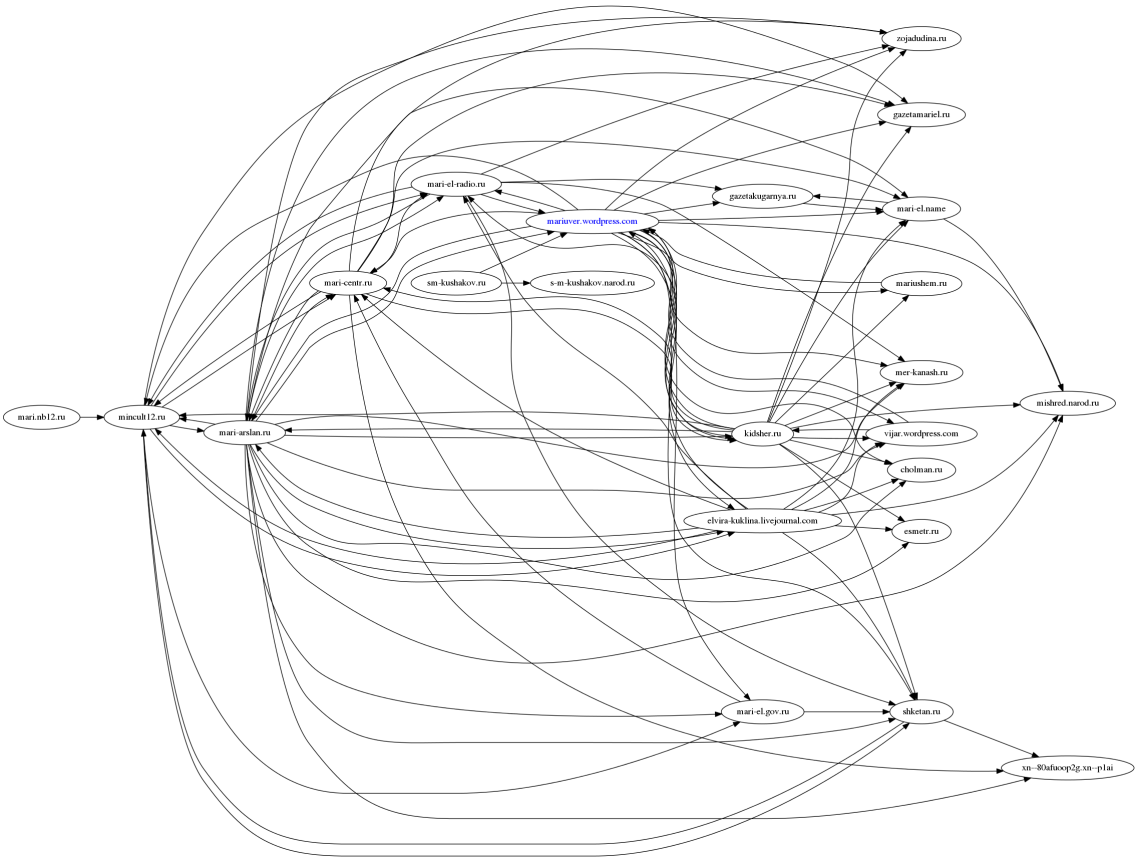
Количество сильно связных компонент: 4

Рис. 28: Веб-граф лезгинского интернета



Топ-5 вершин по алгоритму pagerank: gusarpress.com (0.388), alamjurnal.com (0.388), lezgikim.narod.ru (0.083), suleiman-stalskiy.ru (0.083), lezgikids.com (0.0582)

Рис. 29: Веб-граф лугово-восточного марийского интернета



Метрики для веб-графа лугово-восточного марийского интернета (рис. 29)

Количество вершин: 23

Количество рёбер: 96

Средняя полустепень захода: 4.174

Средняя полустепень исхода: 4.174

Вершина с макс. степенью центральности: mariuver.wordpress.com (1.090)

Коэффициент ассортативности: -0.149

Количество слабо связных компонент: 1

Количество сильно связных компонент: 12

Топ-5 вершин по алгоритму pagerank: mari-el.name (0.088), mincult12.ru (0.079), mariuver.wordpress.com (0.069), mishred.narod.ru (0.066), gazetakugarnya.ru (0.064)

Рис. 30: Веб-граф эрзя-мордовского интернета

Метрики для веб-графа эрзя-мордовского интернета (рис. 30)

Количество вершин: 4

Количество рёбер: 4

Средняя полустепень захода: 1.000

Средняя полустепень исхода: 1.000

Вершина с максимальной степенью центральности: erzia.saransk.ru (1.0)

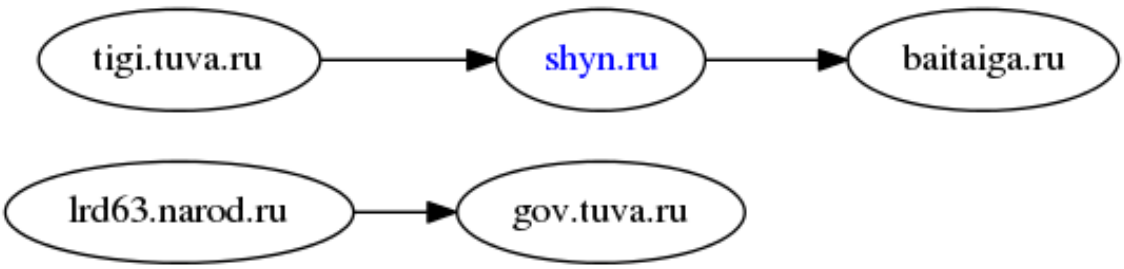
Коэффициент ассортативности: 0.0

Количество слабо связных компонент: 1

Количество сильно связных компонент: 4

Топ-5 вершин по алгоритму pagerank: goloserzi.ru (0.390), erzia.saransk.ru (0.318), erzianj-jurnal.livejournal.com (0.172), erziana.my1.ru (0.120)

Рис. 31: Веб-граф тувинского интернета



Метрики для веб-графа тувинского интернета (рис. 31)

Количество вершин: 5

Количество рёбер: 3

Средняя полустепень захода: 0.600

Средняя полустепень исхода: 0.600

Вершина с максимальной степенью центральности: shyn.ru (0.5)

Коэффициент ассортативности: nan

Количество слабо связных компонент: 2

Количество сильно связных компонент: 5

Топ-5 вершин по алгоритму pagerank: baitaiga.ru (0.311), gov.tuva.ru (0.224), shyn.ru (0.224), lrd63.narod.ru (0.121), tigi.tuva.ru (0.121)

Рис. 32: Веб-граф калмыцкого интернета

Метрики для веб-графа калмыцкого интернета (рис. 32)

Количество вершин: 4

Количество рёбер: 3

Средняя полустепень захода: 0.750

Средняя полустепень исхода: 0.750

Вершины с макс. степенью центральности: forum.freekalmykia.org (0.666), forum.kalmykia.ru (0.666)

Коэффициент ассортативности: nan

Количество слабо связных компонент: 1

Количество сильно связных компонент: 4

Топ-5 вершин по алгоритму pagerank: freekalmykia.org (0.365), vesti-kalmykia.ru (0.235), forum.freekalmykia.org (0.235), forum.kalmykia.ru (0.165)