ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

Государственное образовательное учреждение
высшего профессионального образования
Омский государственный университет
им. Ф. М. Достоевского

Исторический факультет
Кафедра социологии

Нагорный Олег Станиславович

Использование метода интеллектуального анализа текста для изучения Интернет-СМИ Омской области

ДИПЛОМНАЯ РАБОТА

Научный руководитель
кандидат социологических наук
К.В. Павленко

Омск 2014 г.

Оглавление

Заметки по тексту

Введение

Последние несколько десятилетий наука анализа данных претерпевает существенные изменения.

Появление глобальной сети Интернет и распространение персональных компьютеров привело к тому, что информации стало больше и производится она намного быстрее, чем раньше. Значительная часть человеческой коммуникации переместилась в виртуальную сферу. Практически у каждой газеты или журнала имеется электронная версия или веб-сайт, где постоянно появляются новые материалы, происходит коммуникация пользователей между собой и с редакцией, проводятся голосования и прямые трансляции. Некоторые СМИ и вовсе отказываются от бумаги и полностью перебираются в электронный формат. Предоставляя более удобные средства потребления, хранения и поиска информации, чем традиционные печатные СМИ, Интернет становится новым центром притяжения как для издателей, так и для их аудитории.

К тому же, благодаря развитию технических средств и совершенствованию алгоритмов оперировать информацией стало проще. Обычный персональный компьютер теперь способен обрабатывать миллионы строк текста за считанные секунды.

Эти изменения открывают перед исследователями невиданные ранее перспективы. На основе наработок в области искусственного интеллекта, машинного обучения, статистики и проектировании баз данных в 80-х гг. XX века сформировалась новая междисциплинарная область знания — Data Mining или интеллектуальный анализ данных. Особенность методов, объединяемых данным понятием, заключается в их способности извлекать из «сырых» данных ранее неизвестные нетривиальные знания. Системы Data Mining сейчас находятся на острие исследований и разработок в области анализа, моделирования и практического использования информации и знаний, создавая новую культуры анализа данных.

Сфера применения данных методов практически ничем не ограничена — их можно применять везде, где имеются какие-либо данные [1, стр. 81]. Одной из таких сфер применения является интеллектуальный анализ данных — прежде всего текста — в социальных науках. Группа методов data mining, предназначенная для интеллектуального анализа неструктурированного текста объединяется под названием text mining.

В социологии анализ текстов обычно осуществляется следующими традиционными методами: дискурс-анализ, контент-анализ, когнитивное картирование и т.п. Однако, как уже говорилось, виртуальное пространство является хранилищем огромного количества текстов. Поэтому обрабатывать и анализировать их обычными, привычными для социологов методами не представляется

возможным. Здесь на помощь социальному исследователю могут прийти методы text mining. С их помощью можно получить результаты, недоступные классическим методам анализа данных – с высокой точностью спрогнозировать результаты выборов [2] или предсказать популярность фильма до выхода в прокат на основе его обсуждения в сети [3].

Однако по некоторым оценкам, многие российские социологи не знакомы с данными методами, что нельзя признать нормальным, поскольку «отбрасывает» отечественную социологию на 20-30 лет назад. Отсутствие соответствующей подготовки в области анализа данных приводит к поверхностному анализу эмпирических данных, в то время как важные и полезные неочевидные закономерности в данных «ускользают» от внимания исследователя [4]. Такое игнорирование современных методов анализа данных вполне может стать «фатальной ошибкой» [5] и привести к возникновению «чёрной дыры» [6] в российской социологии. Сказанное позволяет считать, что работа, показывающая перспективы применения методов data mining в социологических исследованиях, является **актуальной**.

В данном исследовании мы ставим цель рассказать о таком методе анализа данных как text mining и на практическом примере показать его актуальность для социологического анализа.

Проблема исследования заключается недостаточности наработок в области применения методов Data Mining в социологии.

Объект исследования — методы Text Mining в социологическом исследовании.

Предмет исследования — возможности применения Text Mining для задач обработки естественного языка, моделирования тем, анализа настроений и неструктурированного текста в социологическом исследовании.

Глава 1

Теоретическая часть. Text mining как метод анализа данных

1.1. Место text mining в структуре исследовательских методов

1.1.1. Дуальность статистики

> Если данные говорят с вами,
> значит вы — байесовец.
>
> Филип А. Шродт [7, стр. 11]

> Формально, теорема Байеса – это просто математическая формула. Однако её значение гораздо глубже. Теорема Байеса подводит нас к тому, что необходимо иначе взглянуть на процесс выдвижения и проверки идей.
>
> Нэт Сильвер[1] [8]

Для определения того места, которое занимают методы text mining, следует сказать о двух основных направлениях, в которых развивалась математическая статистика и понимание понятия вероятности. Дуальность статистики берёт своё начало из философского спора Аристотеля и Платона [9, стр. 7]. Аристотель считал, что реальность может быть познана только эмпирически и что исследователь должен тщательно изучать вещественный мир вокруг себя. Он пришёл к убеждению, что можно разложить сложную систему на элементы, детально описать эти элементы, соединить их вместе и, затем, понять целое. Именно таким механистичным путём долгое время

[1] Американский статистик, давший самые точные прогнозы президентских выборов в США в 2008 и 2012 гг. Входит в 100 самых влиятельных людей в мире по версии журнала Times.

следовала наука. Однако в дальнейшем стало понятно, что не всегда целое можно представить как простую сумму частей, его составляющих. Часто, будучи соединёнными вместе, совокупность этих частей приобретает новое качество.

В отличии от своего ученика, Платон считал что свойством подлинного бытия обладают только идеи, а человек может лишь воспринимать и воплощать в вещах их смутные очертания. Для Платона идея (целое) была большим, чем сумма её материальных проявлений.

Эта дихотомия восприятия реальности проявляется во многих аспектах человеческой мысли, в том числе и в сфере статистического знания, в котором с XVIII в. существует две основных философских позиции относительно того, как применять вероятностные модели. Первая определяет вероятность как нечто, заданное внешним миром. Вторая утверждает, что вероятность существует в головах людей. [10, стр. 18]. В русле первого подхода возникли вначале классическая и затем развивающая её частотная концепции вероятности. Второй подход нашёл выражение в концепции байесовской вероятности.

Сторонники классического подхода исходят из того, что истинные параметры модели не случайны, а аппроксимирующие их оценки случайны, поскольку они являются функциями наблюдений, содержащих случайный элемент. [11, стр. 5-6] Параметры модели считаются не случайными из-за того, что классическое определение вероятности исходит из предположения равновозможности как объективного свойства изучаемых явлений, основанного на их реальной симметрии [12, стр. 24]. На такое представление о вероятности повлияло то, что в начале своего развития теория вероятности применялась прежде всего для анализа азартных игр. Суждение вида «Вероятность выпадения шестёрки при бросании игрального кубика равняется 1/6» основывается на том, что любая из шести граней при подбрасывании на удачу не имеет реальных преимуществ перед другими, и это не подлежит формальному определению. Таким образом, вероятностью случайного события A в её классическом понимании будет называться отношение числа несовместимых (не могущих произойти одновременно) и равновозможных элементарных событий m к числу всех возможных элементарных событий n:

$$P(A) = \frac{m}{n} \tag{1.1}$$

Однако такое определение наталкивается на некоторые непреодолимые препятствия, связанные с тем, что не все явления подчиняются принципу симметрии. Например, из соображений симметрии невозможно определить вероятность наступления дождливой погоды. Для преодоления подобных трудностей был предложен статистический или частотный способ приближённой оценки неизвестной вероятности случайного события, основанный на длительном наблюдении над проявлением или не проявлением события A при большом числе независимых испытаний и поиске устойчивых закономерностей числа проявлений этого события. Если в результате достаточно многочисленных наблюдений замечено, что частота события A колеблется около некоторой постоянной, то мы скажем, что это событие имеет вероятность. Данный тип вероятности был

выражен Р. Мизесом в следующей математической формуле:

$$p = \lim_{x \to \infty} \frac{\mu}{n}, \tag{1.2}$$

где μ — количество успешных испытаний, n — количество всех испытаний [12, стр. 46-47]. Вероятность здесь понимается как частота успешных исходов я является чисто объективной мерой, поскольку зависит лишь от точного подсчёта отношения количества успешных и неуспешных событий.

Основываясь на этом подходе, статистика занималась созданием вероятностных моделей, которые включали в себя параметры, которые, как предполагалось, связаны с характеристиками исследуемой выборки. Параметры никогда не могут быть известны с абсолютной точностью до тех пор, пока мы не исследуем всю генеральную совокупность [10, стр. 1]. До тех пор всегда существует вероятность отклонить гипотезу, когда она на самом деле верна, т. е. совершить ошибку первого рода. Для обозначения вероятности такой ошибки частотники используют понятие уровня значимости α. Именно вероятность ошибки первого рода частотники ставят во главу анализа, определяя вероятность события. После каждого своего утверждения они обычно добавляют «... на доверительном уровне в 95%», подразумевая, что исследователь допускает вероятность ошибки в пяти процентах случаев (при $\alpha = 0{,}05$) [9, стр. 10-11].

Иногда параметры вообще не возможно интерпретировать применительно к реальной жизни, поскольку модели редко бывают абсолютно верными. Модели, как мы надеемся, — это некоторые полезные приближения к истине, на основании которых можно делать прогнозы. Тем не менее прежде всего классическое статистическое исследование сосредоточено на оценке параметров, а не на предсказании [10, стр. 1].

Частотный подход доминировал в XX веке, придя на смену другому пониманию вероятности, связанном с именем английского математика Томаса Байеса [13, стр. 2]. Сущность байесовского подхода составляют три элемента: априорная вероятность, исходные статистические данных данные, постаприорная вероятность.

Байесовская статистика начинает построение своей модели при помощи понятия априорной вероятности, с помощью которой описывается текущее состояние наших знаний, относительно параметров распределения [10, стр. 18]. Априорная вероятность, такие образом, — это степень нашей уверенности в том, что исследуемый параметр примет то или иное значение ещё до начала сбора исходных статистических данных. На этом основании байесовское понимание вероятности относят к группе субъективистских трактовок вероятности. Чаще всего предполагается, что для оценки степени уверенности необходимо привлечь экспертов, чьё субъективное свидетельство

позволит избежать действительной многократной реализации интересующего нас эксперимента[2] [15, стр. 34].

Следующий элемент — это исходные статистические данные. По мере их поступления статистик пересчитывает распределение вероятностей анализируемого параметра, переходя от априорного распределения к апостериорному, используя для этого формулу Байеса:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1.3}$$

где $P(A)$ — априорная вероятность гипотезы A, $P(A|B)$ — вероятность гипотезы A при наступлении события B (апостериорная вероятность), $P(B|A)$ — вероятность наступления события B при истинности гипотезы A, $P(B)$ — полная вероятность наступления события B. Суть формулы в том, что она позволяет переставить причину и следствие: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной. Эту формулу также называют формулой обратной вероятности. Процесс пересмотра вероятностей, связанных с высказываниями, по мере поступления новой информации составляет существо **обучения на опыте**[3] [11, стр. 21-22] и является одним из возможных способов формализации и операционализации следующего тезиса: «*степень нашей разумной уверенности в некотором утверждении (касающемся, например, неизвестного численного значения интересующего нас параметра) возрастает и корректируется по мере пополнения имеющейся у нас информации относительно исследуемого явления*» [16, стр. 93]. В частотном подходе данный тезис интерпретируется в свойстве состоятельности оценки неизвестного параметра: чем больше объём выборки, на основании которой мы строим свою оценку, тем большей информацией об этом параметре мы располагаем и тем ближе к истине наше заключение. Специфика байесовского подхода к интерпретации этого тезиса основана на том, что вероятность, понимаемая как количественное значение степени разумной уверенности в справедливости некоторого утверждения, пересматривается по мере изменения информации, касающейся этого утверждения. Поэтому в данном подходе вероятность всегда есть условная вероятность, при условии нынешнего состояния информации (в русле классического подхода исследователь скорее склонен рассматривать совместную вероятность [9, стр. 5]).

Дискуссии вокруг того, какой же метод предпочтительней, ведутся уже не одно столетие, породив великое множество книг и статей на эту тему [17], [13], но к однозначному выводу прийти

[2] Не следует путать субъективный характер байесовской вероятности в целом с внутренним разделением сторонников данного подхода на объективистов и субъективистов, основанном на различном отношении к роли рациональных ограничений при определении априорной вероятности. В качестве примера различного подхода к определению априорной вероятности рассмотрим ситуацию, где событием является изъятие мячика из урны, наполненной красными и чёрными мячиками — и это всё, что нам известно об урне. Зададим вопрос: какова априорная вероятность (до изъятия мячика), что изъятый мячик будет чёрного цвета? Субъективисты, считающие роль рациональных ограничений относительно небольшой, ответят, что любая вероятность от 0 до 1 может быть рациональной, так как по их мнению наша оценка априорной вероятности зависит большей частью от нерациональных факторов — социализации, свободного выбора и др. Объективисты же будут настаивать, что априорная вероятность в данном случае равняется 1/2, поскольку именно такая вероятность в соответствии с принципом неопределённости Джейнса инвариантна к к размерам и трансформациям мячиков [14].

[3] Понятие «обучение на опыте» ещё не раз встретится в данной работе, поскольку именно оно составляет суть машинного обучения — подраздела науки искусственного интеллекта, методы которого используются в text-mining.

не удалось. Острота дискуссии объясняется тем, что спор сторонников байесовского и частотного подхода к статистическому выводу отражает два различных взгляда на способ добычи научного знания. Именно поэтому от ответа на этот, казалось бы, локальный вопрос математической статистики зависит развитие всей науки.

Так или иначе, в 1980-х годах, стало ясно, что частотный подход к статистическому выводу не достаточно хорошо подходит для анализа нелинейных отношений в больших объёмах данных, производимых сложными системами при моделировании процессов реального мира [9, стр. 10]. Для преодоления этих ограничений частотники создали нелинейные версии параметрических методов, такие как множественный нелинейный регрессионный анализ.

В то время как в частотном подходе происходили изменения, немногочисленные сторонники байесовского подхода упрямо продвигали свою точку зрения на модель статистического вывода. Как оказалось, байесовская модель лучше подходит для поиска ответов на некоторые практические вопросы, поскольку полнее учитывает прошлую информацию и располагает к предсказаниям. Например, намного важнее минимизировать вероятность ложноотрицательного диагностирования некоторой опухоли как раковой, чем вероятность её ложноположительного определения (ошибка первого рода).

Продемонстрируем на примере различия в рабоче частотных и байесовских методов проверки гипотез. Предположим, некоторый стрелок утверждает, что точность его стрельбы составляет 75%. Когда стрелка попросили продемонстрировать свои навыки, он попал в мишень только 2 раза из 8. Какова вероятность, что стрелок сказал правду о своих навыках.

Решение задачи в частотном подходе. Гипотеза H_0 — стрелок сказал правду. Испытание — стрельба по мишени. Событие A — попадание в мишень. $P(A)$ постоянная и равна 0,75. Для расчёта вероятности того, что событие A наступило не более 2 раз в 8 независимых испытаниях, применим формулу Бернулли для количества успешных испытаний k = 0, 1, 2 и получим, что $P(A \leq 2) = 0{,}0042$. Следовательно, при уровне значимости α = 0,05 следует признать невероятным, что точность стрелка составляет 75%, гипотеза H_0 отвергается.

Отметим некоторые особенности данного решения. Во-первых, для решения задачи мы фактически использовали только умение рассчитывать совместную вероятность, ведь формула Бернулли является сокращённым видом расчёта совместной вероятности успешных комбинаций. Во-вторых, мы решили, что если гипотеза верна, то вероятность отклонить гипотезу, когда она на самом деле верна должна быть не менее 5%, т. е. нам важно, чтобы вероятность ложноположительного ответа была ниже определённой границы. Вероятность ложноотрацательного ответа не рассматривается.

Решение задачи в байесовском подходе. В данном подходе мы не проверяем гипотезу, а рассчитываем условную вероятность события A (точность стрелка составляет 75%) при условии события B (стрелок попал в мишень не более 2 раз из 8). Прежде всего нам нужно оценить априорную вероятность события A. Это можно сделать посмотрев статистику стрельбы остальных стрелков. Предположим, мы выяснили, что 70% стрелков имеют точность в 75%. Следователь-

но, $P(A) = 0{,}7$. $P(B|A)$ мы уже рассчитали в частотном подходе. $P(B)$ легко рассчитывается по формуле полной вероятности. По формуле Байеса $P(A|B) = 0{,}0301$.

Как видно из этого примера, в байесовском подходе другая логика расчёт вероятности: на основании данных рассчитывается вероятность того, что H_0 верна, в то время как раньше мы рассчитывали вероятность того, что стрелок поразил мишень не более 2 раз в 8 независимых испытаниях. Данные, полученные с помощью данного метода, данные можно использовать более продуктивно. Предположим, что мы рассчитываем не вероятность того, что стрелок с определённым умениями поразил мишень какое-то количество раз, а вероятность наличия тяжёлого заболевания у человека с каким-то количеством положительных тестов. В случае частотного подхода мы узнаем, какова вероятность того, что больной человек получит н-ое количество положительных тестов. Байесовский же подход позволяет узнать именно то, что нам надо — вероятность того, что человек, получивший н-ое количество положительных тестов, болен. Другой плюс данных методов — они работают даже если размер выборки равен нулю. В таком случае байесовская вероятность равна априорной.

Проведение тестирования на статистическую значимость оценивает лишь вероятность получения похожего результата с другим набором данных при сохранении тех же самых условий. Однако оно предоставляет ограниченную картину такой вероятности, поскольку в расчет принимается ограниченное количество информации относительно исследуемых данных. И оно само по себе не способно вам сказать, являются ли основные положения исследования верными и будут ли подтверждены полученные результаты в различных условиях [18]. Уровень p говорит только о вероятности получения результата при (обычно) совершенно нереалистичных условиях нулевой гипотезы. А это совсем не то, что мы хотим узнать, — обычно мы хотим знать величину эффекта независимой переменной с учетом имеющихся данных. Это байесовский вопрос, а не частотный. Вместо этого значение p часто интерпретируется так, будто бы оно показывало силу ассоциации [7, стр. 11].

С другой стороны у и байесовского метода имеются несколько недостатков. Одним из них является необходимость привлекать для расчёта априорные данные, которые могут быть недоступны. А если они и доступны, то, как отмечалось выше, часто носят субъективный характер. Другой недостаток — сложность вычислений. В вышеописанном примере для вычисления байесовской вероятности нам необходимо было вычислить частотную вероятность, полную вероятность, и, наконец, собственно байесовскую вероятность. Сложность байесовских вычислений частично объясняет тот факт, что байесовские методы вновь обрели популярность с развитием вычислительной техники. Следующий недостаток байесовского метода — неинтуитивность, непонятность его результатов для обыденного сознания. Именно на этой неинтуитивности построен знаменитый парадокс Монти Холла, который легко решает с помощью формулы Байеса.

1.1.2. Data mining как объединение подходов

В грамм добыча, в годы труды.
Изводишь единого слова ради
Тысячи тонн словесной руды.

В. В. Маяковский

Дальнейшее развитие статистических методов, особенно в их байесовском варианте, привело к возникновению следующего поколения методов статистического анализа, а именно методов машинного обучения. Первоначально эти методы развивались в двух направлениях, первое из которых представлено искусственными нейронными сетями, а второе — деревьями принятия решений [9, стр. 11-12].

Развитие методов машинного обучения в свою очередь привело к созданию статистической теории обучения (Statistical Learning Theory), которая направлена на решения проблемы предсказания на основе имеющихся данных [9, стр. 12-13].

Вышеуказанные сферы знания, пересекаясь друг с другом образуют новую дисциплину – интеллектуальный анализ данных или data mining. Data mining — это междисциплинарная область знания, находящая на пересечении традиционного статистического анализа, искусственного интеллекта, машинного обучения и развития больших баз данных [9, стр. 5]. Можно даже сказать, что data mining — это новая философия, новый взгляд на анализ данных.

Хотя как самостоятельная дисциплина data mining окончательно оформился в 1990-х гг. [9, стр. 15], о важности ухода от чистой математической статистики в пользу анализа реальных данных говорил ещё Джон Тьюки, который в 1962 году написал статью под названием «Будущее анализа данных» (The future of data analysis), в которой изложил основные идеи новой тенденции. Тьюки говорил о том, что излишняя сосредоточенность на математических теориях в статистики не помогает в решении реальных жизненных проблем. Он был убеждён, что анализ данных — это работа, схожая с работой следователя и что надо дать данным говорить самим за себя. Однако эти идеи тогда не были восприняты приверженцами чистой математической статистики, которые утверждали, что правильная процедура статистического анализа прежде всего предполагает выдвижение научных гипотез, а затем уже из проверку, на основе полученных данных. Попытка анализа данных до выдвижения гипотезы категорически отвергалась, поскольку считалось, что это приведёт к смещению гипотезы в сторону того, что показали данные. Такая позиция привела к тому, что термин «Data mining» стали использовать в уничижительном значении [19, стр. 788].

Развитие информационных технологий и вычислительной техники с одной стороны привело к появлению огромного количества данных, а с другой — предоставило инструменты для их удобного сбора, хранения и обработки. Эти процессы также изменили течение академических споров, поскольку учёные осознали перспективы новой парадигмы анализа данных. Почему же data mining стал популярен в сложившихся условиях?

Суть философии data mining частично выражена в названии этой области знания, которое состоит из двух понятий: поиск ценной информации в большой базе данных (data) и добыча гор-

ной руды (mining). Именно в просеивании через сито своих инструментов огромного количества «сырых», часто неструктурированных данных в поисках самородков, т. е. осмысленной, нетривиальной информации — знаний. Более верным названием для этого процесса было бы «knowledge mining from data» (добыча знаний из данных) [20, стр. 5]. Как видно, строки Маяковского, вынесенные в эпиграф, как нельзя лучше характеризуют интеллектуальный анализ данных.

Исходное определение термина, которое дал наш бывший соотечественник Григорий Пятнецкий-Шапито, звучит следующим образом: «Data mining — это процесс обнаружения в сырых данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности» [1, стр. 78].

В статистике data mining часто иногда отождествляют с таким процессом как Knowledge Discovery in Databases, в то время как компьютерщики (computer scientists) предпочитают рассматривать первое определённую как часть второго.

1.2. Методология text mining

По аналогии с термином data mining термину text mining можно дате следующее определение – это нетривиальный процесс обнаружения действительно новых, потенциально полезных и понятных шаблонов в неструктурированных текстовых данных [21, стр. 211].

Главная цель text mining состоит в обработке неструктурированного текста и, если это требует решаемая с помощью данного метода проблема, слабоструктурированных и стуктурированных данных, с тем, чтобы извлечь новое, значимое и применимое знание для лучшего принятия решений [22, стр. 78].

Так как по сравнению с остальными устоявшимися статистическими методами text mining является относительно новой и неустоявшейся областью знания, сложно говорить, о наличии единой и общепринятой совокупности методов, направленных на получение устойчивого результата, т. е. о методологии. Во многом, исследователи, использующие методы text mining, руководствуются собственным опытом, приобретённым методом проб и ошибок, и создают собственную методологию. Наиболее значимые причины такого волюнтаризма включают следующее [22, стр. 74]:

- Разные исследователи вкладывают в понятие text mining разные значения. Данное определение ещё только формируется.
- Неструктурированный характер данных открывает широкие возможности для действий исследователя.
- Существует несколько форматов неструктурированных данных, некоторые из которых могут быть классифицированы как полуструктурированные (HTML, XML, JSON и другие).
- Огромные объёмы данных часто требуют сокращения и упрощения.

Самым популярным вариантом методологии data mining является CRISP-DM (CRoss Industry Standard Process for Data Mining) – Стандартный межотраслевой процесс data mining. Так как главное отличие text mining от data mining заключается в том, что первый специализируется на определённом типе данных, с небольшими изменениями CRISP-DM можно применить и для анализа текстовых данных. Весь цикл обработки данных это методологии представлен шестью последовательными этапами [22, стр. 74].

Этап 1. Определение целей исследования. С этого начинается практически любая осмысленная деятельность. Грамотная постановка цели требует глубокого понимания всех аспектов ситуации, в которой проводится исследование и чёткого определения результата, который мы хотим получить. Для этого необходимо изучить проблему, на решение которой направлено исследование.

Этап 2. Оценка доступности и характера данных. Данный этап включает в себя следующие задачи:

- Определение источников текста. Текст может иметь цифровую форму или написан на бумаге, находится внутри или за пределами исследуемой организации.
- Оценка доступности и применимости данных.
- Сбор первичных данных.
- Оценка содержательности данных (содержится ли в них необходимая для исследования информация).
- Оценка количества и качества данных.

После того, как разведывательная часть исследования успешно завершена, можно приступить к сбору данных из различных источников.

Этап 3. Подготовка данных. Подготовка данных – необходимый для text mining этап, ведь специфика данного метода по сравнению с data mining заключается в более трудоёмких стадиях сбора и обработки данных [22, стр. 77]. Это следствие неструктурированности или слабой структурированности данных. Этап подготовки данных состоит из следующих фаз:

Создание корпуса. В лингвистике корпус – это большой структурированный набор текстов. На данном этапе необходимо собрать все текстовые документы, относящиеся к исследуемой проблеме. Исследователю предстоит решить, какие данные и в каких объёмах необходимо собрать и проанализировать, чтобы решить поставленную задачу. Следует помнить, что все методы data mining сильно зависимы от точности полученных результатов от их количества.

После того, как документы будут собраны, их необходимо трансформировать таким образом, чтобы они были представлены в единой форме (например в базе данных или текстовом файле) для компьютерной обработки.

Предварительная обработка данных. На данном этапе мы должны решить одну из главных проблем анализа текстов, а именно большое количество лишних слов в документе [21, стр. 213], которые только создают помехи при включении их в анализ. Таким образом целью данного этапа будет удаление несущественных и вносящих помехи данных и преобразование данных к удобному для анализа виду. При подготовке данных обычно используют следующие приёмы:

– Удаление стоп-слов. Стоп-словами называются слова, которые являются вспомогательными и несут мало информации о содержании документа. Обычно заранее составляются списки таких слов, и в процессе предварительной обработки они удаляются из текста. Типичным примером таких слов являются вспомогательные слова и артикли, например: «так как», «кроме того» и т. п.

– Стемминг или лемматизация терминов, т. е. приведение их к простейшей форме, чаще всего к корню или к начальной форме слова (1-е лицо, единственное число, именительный падеж). Например, слова «социолог», «социологический» и «социология» различны, но относятся к одной и той же теме. Вследствие процедуры стемминга, основанного на приведении к корню, всё они будут приведены к одному термину «социолог». Это позволит сократить количество терминов и увеличить их частоту.

– Вышеописанные приёмы значительно уменьшают количество терминов в корпусе. Обычно после этого на основе корпуса обычно создаётся матрица терминов (document-term matrix), строками которой являются отдельные документы корпуса, а колонками – уникальные термины. Соответственно в ячейках матрицы записывается число повторений терминов в документах. Представленные в таком виде текстовые данные удобно использовать для дальнейшего анализа.

Этап 4. Разработка и калибровка модели. На этом этапе происходит применение методов извлечения знаний. В text mining используется четыре основных метода: классификация, кластеризация, ассоциация, анализ трендов.

Классификация. Вероятно, наиболее распространённым методом, использующимся в интеллектуальном анализе данных является распределение объектов по классам согласно каким-либо важным признакам. В отношении к text mining эта задача известна как *категоризация текста* и заключается в нахождении верной темы или понятия для каждого документа из корпуса. Сегодня автоматическая категоризация текста применяется в контекста различных задач, включая фильтрацию от спама, определение жанра, категориацию веб-страниц в иерархических каталогах и многое другое.

Существует два основных подхода к классификации текста. В первом подходе знания экспертов о категориях кодируются в правила, на основе которых объект относится к тому или иному классу. Второй подход, пришедший из машинного обучения, построен на работе определённого алгоритма, который обучившись на уже классифицированном наборе данных, способен в дальнейшем с некоторой вероятностью определять класс остальных объектов.

Кластеризация. Кластеризация – это упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к классу задач обучения без учителя. Это означает, что в процессе кластеризации не используется какая-либо предварительная информация о характеристиках групп, которые должны получится в итоге. В этом отличие кластеризации от классификации, где для определения класса объекта используется обучающая выборка или знания экспертов (происходит обучение с учителем).

Создание правил ассоциации. Ассоциация – это процесс поиска повторяющихся образцов в группе объектов. Этот метод используется в интернет магазинах, чтобы на основании выбранных пользователем товаров предложить ему другие варианты. Главная идея этого метода в том, чтобы определить, правила, на основании которых определённые и часто непохожие между собой объекты объединяются в единый набор.

В text mining данный метод используется чтобы измерить отношения между понятиями или группами понятий. В правиле ассоциации $X \Rightarrow Y$

Этап 5. Проверка результатов. После того, как модель создана и проверена, мы должны произвести общую проверку всех действий. Например, необходимо убедиться, что выборка произведена правильно. Также случается, что в процессе построения исследования теряется основная цель, для достижения которой оно начиналось. На данном этапе следует проверить, решает ли модель сформулированную проблему и служит ли, таким образом, достижению цели. Если что-то упущено, необходимо вернуться назад к этапу, породившему рассогласованность между целью и результатом.

Этап 6. Внедрение. В случае, если по итогам проверок было решено, что модель решает поставленную проблему, её можно применять. В самом простом случае внедрение может принимать форму написания отчёта о результатах исследования. В сложном – построение интеллектуальной системы на основе построенной модели с тем, чтобы она могла быть повторно использована для принятия решений.

1.3. Область применения и примеры использования методов text mining

Интеллектуальный анализ текста находит своё применение во многих областях. В экономике с его помощью можно установить, как настроения в СМИ влияют на котировки фондового рынка [23], имеется ли связь между отзывами о продукте в Интернет-магазине и его продажами [24], как макроэкономические показатели могут быть измерены поисковыми запросами [25] и текстами из социальных медиа.

В психологии этот метод позволяет узнать, как психическое состояние человека выражается в его языке [26] и правда ли, что суточные и сезонные циклы настроения носят надкультурный характер [27].

Одним из самых известных и ранних примеров применения методов text mining в исторических исследованиях является установление авторства сборника статей «Федералист» [28]. Здесь text mining принял форму стилометрии. Другое исследование в области text mining продемонстрировало, что в XVIII понятием «литература» объединялся более широкий класс явлений, чем сегодня [29].

Социолингвисты использовали text mining для идентификации географически зависимых лингвистических переменных и, на основании этого, предсказания местоположения пользователя на основе написанного им текста [30].

Text-mining также можно использовать в качестве вспомогательного метода, уточняющего результаты традиционных опросов [31].

Рассматриваемый метод активно используется в политологических и социологических исследованиях. Так как в данной работе будет представлено исследование именно такого вида, рассмотрим из подробней.

В 2012 году было опубликована работа, посвящённая выявлению политических предпочтений бельгийских Интернет-СМИ в ситуации политического кризиса [32]. Суть кризиса состояла в том, что на протяжении более чем полутора лет ведущие валлонские и фламандские партии не могли договориться о составе федерального правительства. Корпус документов, используемых в исследовании, составили 68 000 статей, опубликованные в онлайн версиях восьми крупнейших фламандских газет в период с начала 2011 года до завершения политического кризиса в октябре того же года. Помимо даты публикации, критерием выбора статьи для анализа служило наличие в ней ключевых слов. Такими ключевыми словами считались названия фламандских политических партий, имеющих, по крайней мере, одно место в парламенте, и имена их важнейших представителей.

Первичная обработка данных включала удаление дубликатов. Затем на основе тонального словаря из более чем 3000 прилагательных, которые чаще всего встречались в отзывах на товары и которые были вручную проранжированы по шкале полярности (1 – позитивное, -1 – негативное) и субъективности (0 – объективное, 1 – субъективное), в каждой статье был произведён анализ

тональности упоминаний выбранных политических партиях и политиках. Для этого подсчитывалась полярность каждого прилагательного в пределах двух предложений до и двух предложений после упоминания партии. Для уменьшения шума исключались прилагательные набравшие меньше 0,1 и больше -0,1 очка по шкале полярности. В результаты было выделено 360 613 оценок.

Следующий шаг в данном исследовании – определение степени представленности и популярности политической партии. Степень представленности $coverage(e, s)$ политического субъекта e в газете s определялась как отношение количества статей газеты, где упоминалась данная партия, к количеству всех статей данной газеты A_s:

Популярность $popularity(e)$ политического субъекта e определялась через относительное количество голосов, отданных за неё в результате голосования в 2010 году.

Популярность использовалась в качестве априорного распределения для расчёта степени склонности газеты к освещению определённой политической партии. Данная склонность определялась как разность между представленностью партии в газете и её реальной популярностью, определённой в результате выборов.

Таким образом было выявлено, какие политические субъекты пользуются популярностью электронных СМИ в большей или меньшей степени, чем среди населения в целом.

Следующий шаг – выявление тональности упоминания политических партий и их представителей. Для каждого субъекта было подсчитано количество положительных и отрицательных отзывов, составлен график изменения тональности во времени.

В результате исследования при помощи методов анализа текстов были выявлены политические предпочтения главных фламандских новостных сайтов во время политического кризиса.

Другие методы были применены для выявления различий в освещении событий, приведших к восстанию 2011 года в Египте, египетскими государственными и негосударственными СМИ [33]. Материал для анализа составили более 29 000 новостных статей, вышедших в 2010–2011 годах. В методологической части работы был использован такой метод тематического моделирования как латентное размещение Дирихле (LDA), с помощью которого можно выполнить задачу категоризации документов (такой же метод будет использован в данной работе).

Было показано, что правительственные СМИ при освещении таких событий акцентировали внимание на угрозе дестабилизации и терроризма и старались рассказывать проведении реформ в стране. Независимые же СМИ наоборот были нацелены на мобилизацию в целях противостоянию режиму и фактически игнорировали действия правительства. Таким образом, было доказано, что режим Хосни Мумбарака потерял контроль на медиадискурсом ещё до начала активной фазы протестов.

Существуют примеры использование методов text-mining и в отечественных исследованиях. Дальше всего в этой сфере продвинулась сотрудники НИУ-ВШЭ, в частности заведующая Лабораторией интернет-исследований Кольцова Елена Юрьевна. Исследовательский коллектив под её руководством в рамках проекта «Разработка методологии сетевого и семантического анализа блогов для социологических задач» поставил перед собой задачу выявления на больших массивах данных русскоязычной блогосферы тематические кластеры постов (о чем говорят?) и сообществ,

основанные на комментировании (кто с кем говорит?), а также выяснения того, совпадают ли комментовые сообщества с тематическими кластерами (т.е. основана ли общность комментирования на общности темы?).

Тестовой тематикой являлась тема Ислама. Эмпирический материал исследования составили 7941 статей топовых блогеров Живого Журнала за период 21-23 и 24-26 декабря 2011 года и комментарии к ним, собранные с помощью программы «Blogminer». Выбор записей с таким временем написания был обусловлен тем, что именно в это время ожидалась реакция со стороны «населения» российской блогосферы на выборы в Государственную Думу, состоявшиеся 4 декабря.

После операции по выявлению сообществ, которая разделила полную сеть постов на отдельные подмножества исследователи отобрали несколько групп постов для качественного анализа. Его целью было установить, связаны ли посты, входящие в одну группу по смыслу (тематически) или каким-либо другим образом (принадлежат перу одного или нескольких авторов).

По результатам качественного изучения постов из автоматически составленных групп был сделан вывод, что гипотеза исследования не подтвердилась: не были найдены доказательства того, что комментовые сообщества интегрированы общими темами в Живом Журнале.

Несмотря на неподтверждение гипотезы исследования, участие в проекте дало исследователям богатый опыт в организации Интернет-исследований, в результате чего была написана статья «К методологии сбора Интернет-данных для социологического анализа» [34].

1.4. Отличие text mining от контент анализа

После поверхностного взгляда на методы text mining может сложиться впечатление, что они повторяют уже хорошо известный социологом контент-анализ. И правда, из-за своего широкого распространения термин «контент-анализ» иногда используют как обобщающий для всех методов систематического и претендующего на объективность анализа политических текстов и текстов, циркулирующих в каналах массовой коммуникации. Однако такое расширительное понимание контент-анализа неправомерно, поскольку существует ряд исследовательских методов которые не могут быть сведены к стандартному контент-анализу даже при максимально широком его понимании [35]. Обладая большим количеством общих черт, эти методы тем не менее имеют существенные отличия, что оправдывает их выделение в отдельные группы.

Что интересно, хотя методы контент-анализа и text mining имеют много общего, исследователи, работающие в каждой из этих двух сфер, редко ссылаются друг на друга. В литературе о text mining почти никогда не упоминаются методы контент-анализа и наоборот. Ещё сложнее найти источники, где сравниваются два данных метода. Как нам видится, причина такой ситуации прежде всего кроется 1) в недостаточной осведомлённости исследователей о методах интеллектуального анализа текста, 2) отсутствии однозначного определения как контент-анализа [36, стр. 156], так и (в ещё большей степени) интеллектуального анализа текста, 3) и, о чём уже говорилось выше, привычкой называть любой метод анализа текстов контент-анализом.

Итак, как соотносятся такие понятия как контент-анализ и text mining?

Если рассматривать временной критерий, то контент-анализ возник раньше text miming. В советской социологической литературе происхождение контент-анализа связывалось с именами У. Томаса и Ф. Знанецкого. Сейчас же многие зарубежные [37] и отечественные [38] исследователи отмечают, что он возник сто и более лет тому назад, упоминая опыт использования метода, очень близкого к этому, когда в XIII веке в Швеции был осуществлён анализ сборника из 90 церковных гимнов, прошедших государственную цензуру и приобретших большую популярность, но обвинённых в несоответствии религиозным догматам. Наличие или отсутствие такого соответствия и определялось подсчётом в текстах этих гимнов религиозных символов и сравнения их с другими религиозными текстами, в том числе тех, которые считались еретическими. Частота использования определённых заранее собранных слов и тем позволяла судить о том, насколько корректен текст с точки зрения официального учения церкви. Как сформировавшийся метод анализа контент-анализ впервые был использован Максом Вебером в 1910 году для анализа освещаемости прессой политических акций в Германии [36, стр. 155].

История методов text mining насчитывает гораздо меньше времени, поскольку тесно связана с развитием вычислительной техники и сопутствующих ей дисциплин, таких как искусственный интеллект, обработка естественного языка. Развитие информационных технологий привело к взрывообразному росту количества информации. Этот рост иллюстрирует тот факт, что количество веб-страниц в Интернете возросло с 10 миллионов в 2001 г. до 150 миллиардов в 2009 [22, стр. 4]. Для описания нового характера данных, которые отличаются большим объёмом, высокой скоростью роста и значительным многообразием свои форм, в специальном номере журнала «Nature» от 3 сентября 2008 года был введён термин «большие данные» (big data). Феномен «больших данных» создал потребность в новых методах обработки и анализа, способных извлекать полезное знание из ранее невиданных объёмов неструктурированной текстовой информации. Можно сказать, что методы text mining предназначены в первую очередь для анализа «больших данных».

Указать станицы, где это написано. Для этого надо взять рабочую книгу социолога

Другое различие между контент-анализом и интеллектуальным анализом текста заключается различии их методологии. Из «Рабочей книги социолога» [39] мы знаем, что контент-анализ относится к формализованным методам анализа документов, суть которых «сводится к тому, чтобы найти такие легко подсчитываемые признаки, черты, свойства документа (например, частота употребления определённых терминов), которые с необходимостью отражали бы определённые существенные стороны содержания с тем, чтобы сделать его доступным точным вычислительным операциям. В настоящее время, программы, ориентированные на контент-анализ используют основанные на классической статистике алгоритмы [40, стр. 735]. Контент-анализ не занимается собственно смыслом, а исключительно частотным распределением смысловых единиц в тексте, или по другому - анализом статических закономерностей частотного распределения смысловых единиц в тексте, и не более того [41, стр. 15].

В процедуре контент-анализа можно выделить несколько этапов [42, стр. 12-13]. Основа контент-анализа – это подсчёт встречаемости некоторых компонентов в анализируемом информа-

ционном массиве, дополняемый выявлением статистических взаимосвязей и анализом структурных связей между ними, а также снабжением их теми или иными количественными или качественными характеристиками. Таким образом, на первом этапе контент-анализа необходимо выбрать то, что необходимо считать, т. е. определить единицы текста. Этими единицами могут быть слова, абзацы, статьи или квадратные сантиметры площади, которую занимает текст. Следующий этап – составление кодировочной инструкции и кодирование, т. е. трансформация и агрегация исходных данных в категории, которые позволяют точно описать характеристики текста, релевантные для исследования. На этом этапе исследователь подсчитывает количество появлений слова в тексте или решает, относится ли текст к определённой категории (например, определяет наличие в его содержании эротики). Контент-анализ заканчивается статистической обработкой полученных количественных данных (обычно используются процентные и частотные распределения, разнообразные коэффициенты корреляций) и их интерпретацией.

С другой стороны, для text mining используются методы анализа основанные главным образом на байесовском подходе к статистике. Цель интеллектуального анализа текста состоит не в подсчёте частоты некоторых выделенных единиц, а в получении нового, ранее неизвестного знания. Его результатом скорее всего будет сложная математическая модель, распределяющая документы по заданным категориям или объединяющая их в кластеры.

Рассматриваемые методы различаются также в источниках входных данных [40, стр. 735]. Исторически контент-анализ применялся главным образом для анализа социологических, политологических или психологических данных, в то время как с помощью text-mining сейчас анализируют любые текстовые данные. Однако, что касается типа данных, то контент-анализ является более универсальным методом, поскольку используется для анализа документов самого разнообразного типа – это могут быть визуальные изображения, устная речь, невербальное поведение и т. д.[4] Text mining же по определению является сферой data mining, ограниченной анализом текстовых данных.

Между этими методами существует ещё одно различие, которое заключается в интенсивности использования компьютеров. Контент-анализ возник ещё задолго до изобретения первых ЭВМ и до сих пор предполагает ограниченное использования компьютера: если подсчёт слов легко можно автоматизировать, то процедура кодирования требует непосредственного участия исследователя. Text mining же с самого начала развивался вместе с развитием электронно-вычислительной техники. К тому же его связь с наукой искусственного интеллекта отражается в том, что после программирования модели вмешательство человека часто почти не требуется. Впрочем, это различие постепенно сходит на нет, поскольку сейчас появляются компьютерные системы автоматического контент-анализа, а для успешного интеллектуального анализа текста необходимо пристальное внимание исследователя на всех его этапах.

[4]Например, во время второй мировой войны произошёл один из самых известных случаев применения контент-анализа. На основе анализа нацистской пропаганды на радио, сотрудники BBC предсказали ракетную атаку на Британские острова.

Однако, следует заметить, что изложенная нами позиция по определению терминов контент-анализа и text mining не единственная. Как говорилось ранее, существует большая путаница по разграничению объёмов этих понятий.

Глава 2

Практическая часть. Исследование тематического профиля Интернет-СМИ Омской области

2.1. Определение целей исследования

В данной части работы мы разработаем и проведём небольшое исследование, цель которого состоит в построении тематического профиля Интернет-СМИ Омской области. На примере данного исследования будут показаны возможности метода интеллектуального анализа текста в социологии и дано представление о том, как конкретно и с использованием каких инструментов пройти все ранее выделенные этапы интеллектуального анализа текста. Исследование тематического профиля будет включать в себя следующие задачи:

1. Оценка доступности и характера данных. Сбор данных.
2. Выявление распределения статей и комментариев в времени.
3. Предварительная обработка данных.
4. Тематическое моделирование
   (a) Определение оптимального количества тем.
   (b) Создание модели LDA
   (c) Выявление распределения статей по темам
5. Анализ комментариев
   (a) Выявление распределения комментариев в времени
   (b) Определение тональности комментариев
   (c) Определение тональности по темам

Задачи надо формулировать более крупно? Суть работы в том, чтобы выделить темы и определить отношение комментаторов к этим темам. Две задачи

2.2. Оценка доступности и характера данных. Сбор данных.

Прежде чем начать сбор данных, нам придётся поставить перед собой несколько вопросов, представляющих особенную трудность в данного типа исследованиях. А именно, необходимо определить, что будет являться носителем знаний по исследуемой проблеме (т. е. эмпирическим объектом исследования), каковы границы генеральной совокупности, какой метод будет являться адекватным для построения выборочной совокупности, как определить качественные и количественные характеристики выборки, каковы критерии репрезентативности выборки [34].

Определение эмпирического объекта.

В исследованиях подобного вида эмпирическим объектом могут быть посты, комментарии, отдельные высказывания и многое другое. В нашем случае можно сказать, что источником знаний о проблемах, затронутых в данном исследовании являются новостные статьи в Интернет-СМИ Омской области. Углубляясь дальше, мы должны решить какие аспекты статей нас интересуют. Статья в Интернет-СМИ – не просто текст. Это документ, который имеет свою структуру. В этой структуре нас буду интересовать такие элементы как собственно текст, название, дата публикации, комментарии к статье, принадлежность к тому или иному СМИ. Первая причина, по которой они были выбраны состоит в представленности этих элементов в статьях каждого из рассматриваемых нами Интернет-СМИ. Количество просмотров и ключевые слова (тэги), например, на некоторых ресурсах бывают не указаны. Другая причина – достаточность данных элементов для решения исследовательских задач.

Определение генеральной и выборочной совокупностей.

Определение эмпирического объекта позволяет перейти к установлению генеральной и выборочной совокупностей. На сегодняшний день не существует единой позиции как определять эти совокупности при исследование текстов в сети Интернет. Каждый исследователь придумывает сам, каким способом наиболее полно реализовать принципы выборки.

Зизи Папачарисси [43], например, при исследовании блогосферы в качестве генеральной совокупности определяла все блоги, расположенные на платформе blogger.com. Она объясняет свой выбор тем, что это наиболее популярный и большой по числу блоггеров англоязычный сайт, который предоставляет возможности для персональных публикаций в стиле любительской журналистики. Любой блог, по мнению Папачарисси, размещённый это этом сайте, представлял собой единицу анализа, отвечающую по своим характеристикам признакам принадлежности к генеральной совокупности. Однако нельзя согласиться, что блоги с blogger.com репрезентативны относительно всей блогосферы. Выборочную совокупность блогов Папачарисии составляла использую случайную отправную точку и случайный выборочный интервал. Однако исследователем не было оговорено, какие именно данные вводились в поисковую систему для поиска релевантных блогов

и какие именно блоги считались релевантными, сколько блогов входило в генеральную совокупность и почему было отобрано именно 260. К тому же использование поисковых систем для поиска блогов выглядит сомнительно: алгоритмы данных систем неизвестны исследователю и нельзя сказать, почему были отобраны эти блоги, а не иные.

Генеральную совокупность в данном исследовании составляют новостные статьи Интернет-СМИ г. Омска. Омским Интернет-СМИ считается веб-сайт, ставящий своей задачей выполнение функции средства массовой информации (СМИ) в сети Интернет и ориентированный на аудиторию, живущую в Омской области. По данным Агентства Региональных Исследований за июнь 2014 года в Омске работает около 18 Интернет-СМИ с месячным количеством уникальных посетителе в месяц более 10000 [44].

Использование данных со всех возможных ресурсов – очень трудоёмкая задача, поскольку добавление нового источника требует практически полного переписывания соответствующей части компьютерной программы, ответственной за собственной сбор данных, и частичной переработки модуля предварительной обработки. Вполне привычным для социолога решением будет конструирование выборки. Однако как рассчитать выборку, если не известны объёмы генеральной совокупности? А даже если бы мы знали количество статей каждого из рассматриваемых Интернет-СМИ за любой промежуток времени, разве было бы корректно использовать традиционные методы определения выборочной совокупности в такого типа исследованиях? Эта аналогия видится некорректной по причине кардинального различия эмпирических объектов – человека и текста. При определении людей в качестве эмпирических объектов исследования социолог как правило предполагает, что они в равной степени могут служить источником информации о проблеме. Исключение из этого правила встречается, когда исследователь дополнительно изучает мнение экспертов. Но такие опросы – это отдельная часть исследования, в которой как правило используются другие методы сбора и анализа информации.

Тексты в Интернете не равнозначны по своему значению. По нашему мнению, новостная статья заслуживает тем больше внимания, чем больше количество её просмотров. Статья, которую никто не прочитал, – не существует в медийной сфере.

В нашем случае необходимо определить несколько Интернет-СМИ, все статьи которых будут отобраны для исследования. Мы решили, что при определении значимости статьи определяющей характеристикой является количество просмотров. Хотя Интернет-СМИ в Омске и немало, не все из них одинаково популярны. Судя по тем же данным АРИ [44], в Омске существует всего четыре новостных ресурса, страницы которых просматривают более одного миллиона раз в месяц. На их долю приходится 65% всех просмотров. Представляется, что анализ статей, получивших более половины всех просмотров является достаточным основанием для выделения их в качестве выборочной совокупности, по результатом анализа которой можно будет делать выводы об омских Интернет-СМИ в целом. Таким образом в исследовании будут проанализированы все новостные

статьи с сайтов «Город 55»[1], «БК55»[2], «НГС Омск»[3], «Омск-Информ»[4] за период с 1 сентября 2013 по 1 сентября 2014. Новостными статьями будут считаться те, которые публикуются на данном ресурсе в разделе «Новости». Статьи из категорий «Работа», «Объявления», «Блоги» и др. в анализе не участвуют.

Определившись с данными, которые необходимо собрать, нужно решить, каким способом это сделать, т. е. с использованием каких инструментов и технологий будет производится сбор данных. Для этого мы будем использовать язык программирования Python. Основанием для такого выбора является его простота, поддержка многопоточности, что полезно для более быстрого сбора данных, наличие сторонних библиотек, что позволяет избежать написание рутинного кода, а также то, что обработка и анализ данных также будет производится на этом языке – это обеспечивает некоторую консистентность исследования.

Результаты сбора данных следующие:

– С сайта gorod55.ru было собрано 6302 статьи

– Больше всего новостных статей за указанный промежуток времени было опубликовано на bk55.ru – 14078 статей на bk55.ru

– Наименьшее количество статей – 4780 – было найдено на ngs55.ru

– 8727 статей по указанным параметрам было собрано с сайта omskinform.ru

Всего, таким образом, в анализе участвовало 33887 статей.

На этом этапе крайне важно контролировать корректность и полноту собираемых данных. Сложнее всего было с сайтом bk55.ru, поскольку в нём использовались несколько различных шаблонов для отображения информации, каждый из которых необходимо было отследить и создать под него набор правил для извлечения данных.

Исследуем распределение статей во времени, построив график (рисунок 2.1), на оси x которого отложены дни, а на оси y – количество статей, опубликованных в каждый из дней. Дополнительно выделим выходные дни красным цветом.

Анализ графика позволяет сделать несколько выводов. Во-первых, заметна неравномерность распределения статей по дням недели. В будние дни в среднем публикуется статей 116, в то время как в выходные – только 33. Во-вторых, наблюдается сильное снижение количества публикаций на новогодние каникулы. Наблюдаемые колебания очевидно зависят от рабочего графика сотрудников СМИ.

[1] http://gorod55.ru
[2] http://bk55.ru
[3] http://ngs55.ru
[4] http://omskinform.ru

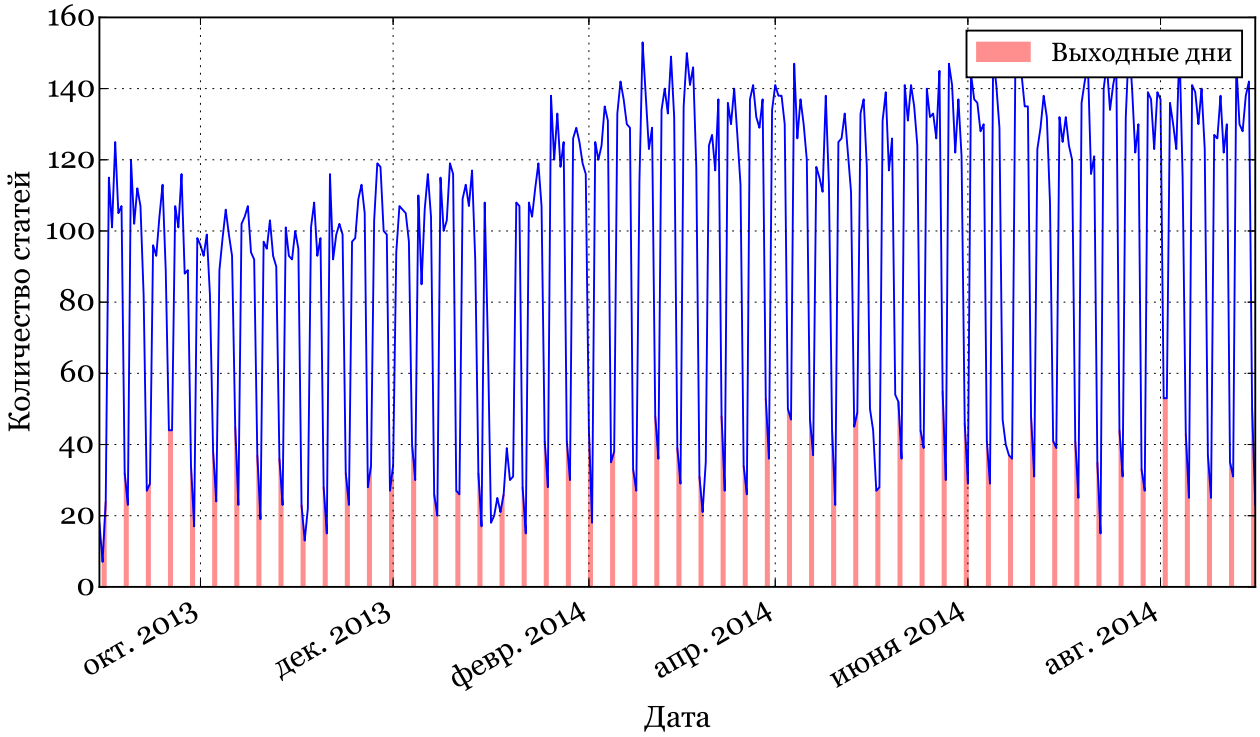

Рисунок 2.1: Количество статей по дням

2.3. Предварительная обработка данных

Предварительная обработка данных – один из важнейших этапов в анализе текста. Наша цель на этом этапе – удаление несущественных и вносящих помехи данных и преобразование данных к удобному для анализа виду.

На самом деле удалять несущественные данные мы начали ещё на этапе сбора данных, поскольку перед записью в базу данных весь текст, если это было необходимо, очищался от HTML-разметки. Преобразование же данных на том этапе заключалось в конвертации текста, содержащего информацию о дате публикации, в специальный тип данных, позволяющий обращаться к этим данным как к дате, например, производить выборку статей за определённый период.

Дальнейшая обработка данный заключалась в следующем:

1. Удаление лишней информации
2. Перевод текста в нижний регистр
3. Токенизация
4. Удаление пунктуации
5. Лемматизация
6. Удаление стоп-слов

Поясним некоторые из этих этапов.

Удаление специфических признаков

Данный этап предварительной обработки данных заключается в удалении из каждой статьи признаков, свидетельствующих о её принадлежности к какому-либо источнику. Если посмотреть на полученные тексты, то можно увидеть, что редакция каждого СМИ устанавливает собственные правила оформления документов, касающиеся оформления ссылок на источники данных, фотографий, указание имён авторов. В случаем если эти отличительные черты не будут устранены, алгоритмы тематического моделирования, которые мы в дальнейшем собираемся применить к собранному корпусу текстов, будут стремиться образовать темы вокруг источников. Процедура унификации статей из различный источников достаточно трудоёмка и требует ручного анализа множества статей с каждого из них, с тем чтобы выявить в них специфические черты для каждого сайта. Такими чертам могут быть имена журналистов данного издания или правила оформления фото и видео материалов (например, около каждой фотографии может указываться копирайт).

Например, чтобы удалить имена журналистов из текстов статей на сайте bk55.ru, необходимо было во-первых, составить их список. Для составления списка, была написана небольшая программа, выводящая два последних слова каждого документа, если они начинались с заглавной буквы (как правило имена авторов указывались в конце документа, хоть и не всегда). Из полученного списка примерно в пятьсот пар были вручную отсеяны пары, не являющиеся именем и фамилией. Те пары из этого списка, которые встречались больше двух раз, считались нами именем и фамилией журналистов сайта bk55.ru. На последнем этапе фамилии журналистов удалялись из каждого документа. К тому же, так как после имён журналистов часто указывалась другая мета-информация (главным образом ссылки источники информации), то также удалялся весь текст после имён, если по размеру этот текст не превышал определённое количество символов (чтобы предотвратить удаление не мета-информации).

После устранения специфической информации данные из различных источников объединялись в единый корпус и подвергались дальнейшей обработке.

Токенизация

Следующим этапом предварительной обработки текста является токенизация. Именно с неё начинается обработка естественного языка как наука и как конкретная деятельность [45]. Под токенизацией понимают процесс сегментации текста на отдельные части, называемые токенами. Именно токены являются теми первичными элементами, которые непосредственно участвуют в процессе анализа.

Выделяют два основных признака токена – лингвистическая значимость и методологическая полезность [45, стр. 1106]. В языках с иероглифической письменностью токенизация является серьёзной проблемой, поскольку один иероглиф может обозначать как морфемы (в таком случае он не удовлетворяет требованиям для того, чтобы считаться токеном), так и целые слова. В английском и русском языках проблема токенизации не стоит так остро и чаще всего токены опре-

деляются через пробелы между словами и знаки препинания. Тем не менее, даже в этих языках существуют определённые нюансы.

Нами было протестировано несколько алгоритмов токенизации (токенайзеры TreebankWordTokenizer, WordPunctTokenizer, PunctWordTokenizer и WhitespaceTokenizer из программы NLTK[5] и токенайзер из Pattern[6]). Корректнее всех выделял токены изначально не презназначенный для работы с русски языком токенайзер из программы Pattern. Например, он единственный интерпретировал URL'ы как цельные токены, не выделяя в них отдельные сегменты на основе знаков препинания.

Стемминг и лемматизация

После токенизации и удаления токенов, являющихся знаками препинания, мы перешли от представления документов как набора символов к документам как списку слов. Дальнейшие наши шаги будут направлены на уменьшение длинны этого списка, т. е. на снижение как общего количества токенов, так и количества их уникальных единиц. Необходимость этих шагов обусловлена желанием снизить вычислительную сложность анализа данных.

Первый шаг направлен на снижение количества уникальных токенов. Для компьютера различные формы одного и того же слова являются совершенно разными словами. Существует два способа для приведения словоформ к одной лексеме. Первый, самый простой, называется стемминг. Он состоит в отсечении слово- и формообразующих частей – префиксов, суффиксов, окончаний, в результате чего остаётся основа слова – неизменная часть, выражающая его лексическое значение.

Более сложным подходом к решению проблемы унификации словоформ является лемматизация. Лемматизация – это процесс приведения словоформы к лемме — её нормальной (словарной) форме. В русском языке нормальная форма имени существительного имеет именительный падеж и единственной число, для прилагательных добавляется требование мужского рода, а глаголы, деепричастия и причастия в нормальной форме должны стоять в инфинитиве.

Для постановки слова в нормальную форму необходимо иметь словарь, где для каждого слова определены его характеристики, т. е. часть речи, падеж, число, род, форма глагола (если это глагол). Создание такого словаря требует колоссальных трудов. В отличии от этого, стемминг предполагает наличие лишь списка приставок, суффиксов и окончаний, количество которых исчисляется несколькими десятками. К счастью, для русского языка существует так необходимый для лемматизации словарь, созданный в рамках проекта OpenCorpora[7]. Использующая этот словарь программа pymorphy2[8] позволяет приводить слова к нормальной форме.

[5] http://www.nltk.org
[6] http://www.clips.ua.ac.be/pattern
[7] http://opencorpora.org
[8] https://pymorphy2.readthedocs.org

Между вышеозначенными способами мы выбрали лемматизацию, поскольку получаемые в результате этого процесса леммы удобнее интерпретировать, по сравнению с усечёнными основами слов, значение которых не всегда легко восстановить.

Удаление стоп-слов

Дальнейшие усилия по уменьшению количества токенов связаны с удалением так называемых стоп-слов. Эти слова, сами по себе почти не неся полезного смысла, тем не менее, необходимы для нормального восприятия текста. Чаще всего к разряду стоп-слов относятся служебные части речи – предлоги, союзы, частицы. Будучи широко распространёнными в тексте, они мало могут сказать о его теме.

В качестве базы для списка стоп-слов был использован список русских стоп-слов из программы NLTK. Однако его нельзя считать достаточно полным. Включая в себя 151 слово, данный список покрывает лишь самые основные случаи. Для его пополнения необходимо обратиться к собранным ранее данным. На их основе был составлен список наиболее часто встречающихся в корпусе токенов. Среди них были выбраны несколько десятков слов, наиболее точно подходящие под описание стоп-слов (это, который, такой, некоторый, другой, тот и др.), которые затем были добавлены в соответствующий список. Представляется, что такой список, дополненный словами, выбранными из числа наиболее распространённых, является достаточно полным, поскольку стоп-слова по своему характеру всегда относятся к наиболее часто встречающимся в тексте. Редкие слова как правило свидетельствуют о принадлежности текста к какой-либо теме, а потому не могут относиться к разряду стоп-слов.

Выводы

Как видно, в общих чертах данный набор процедур повторяет составляющие предварительной обработки данных из методологии CRISP-DM.

Необходимо отметить, что после каждой операции с данными на этапе предварительной обработки следует контролировать последствия производимых изменений. Такой контроль поможет выявить проблемы на раннем этапе, что убережёт от лишней работы в будущем [9].

2.4. Тематическое моделирование

2.4.1. Обзор методов тематического моделирования

Одна из главных задач данного исследования – выявление тем собранных ранее статей. Данная задача известна как тематическое моделирование (topic modeling).

[9] Например, одной из таких проблем, выявленных на раннем этапе, было наличие в текстах некоторых СМИ неразрывных пробелов. Они мешали токенизации, поскольку сегментация производилась по обычным пробелам. Так как визуально неразрывные пробелы почти ничем не отличаясь от обычных, их наличие было установлено только благодаря ручному контролю результатов токенизации. Решением стала замена всех неразрывных пробелов на обычные.

Построение тематическое модели может рассматриваться как задача одновременной кластеризации документов и слов по одному и тому же множеству кластеров, называемых темами. В терминах кластерного анализа тема – это результат би-кластеризации, то есть одновременно кластеризации и слов и документов по их семантической близости. Обычно выполняется нечёткая кластеризация, то есть документ может принадлежать нескольким темам в различной степени. Таким образом, сжатое семантическое описание слова или документа представляет собой вероятностное распределение на множестве тем. Процесс нахождения этих распределений и называется тематическим моделированием [46].

Тематическое моделирование активно развивается последние двенадцать лет и находит своё применение в широком спектре приложений. Оно применяется для выявления трендов в научных публикациях, для классификации и кластеризации документов, изображений и видеопотоков, для информационного поиска, в том числе многоязычного, для тегирования веб-страниц, для обнаружения текстового спама, для рекомендательных систем и других приложений [47, стр. 4].

Тематическое моделирование постепенно находит признание и среди социологов. Помимо уже упоминавшегося исследования египетских СМИ [33], можно упомянуть проект, цель которого заключалась в отслеживании того, как менялось освещение СМИ культурной политики США [48].

В российской социологии подобного вида исследования проводились исследовательским коллективом Лаборатории Интернет-исследований Санкт-Петербургского филиала ВШЭ [49]. Материалом для тематического моделирования послужили записи 2000 самых популярных блогеров по рейтингу популярности Живого Журнала. Для тематического моделирования в данном исследования была использована созданная в лаборатории программа TopicMiner, которая сменила использовавшийся ранее Stanford Topic Modeling Toolbox. Обе этих программы реализовывают алгоритм латентного размещения Дирихле с сэмплированием Гиббса.

Что касается конкретных методов тематического моделирования, то одним из первых был предложен вероятностный латентный семантический анализ (probabilistic latent semantic analysis, PLSA), основанный на принципе максимума правдподобия, как альтернатива классическим методам кластеризации, основанным на вычислении функций расстояния. Вслед за PLSA в 2003 году был предложен метод латентного размещения Дирихле (latent Dirichlet allocation, LDA) [50] и его многочисленные обобщения [51], [52]. В том числе благодаря этим обобщениям LDA безусловно лидирует среди вероятностных тематических моделей.

Эти обобщения учитывают специфические переменные, что улучшает работу алгоритма в приложении к конкретным задачам. Например, когда исследуемые документы имеют дату публикации, можно применить модель Topics over Time LDA, которая более корректно показывает изменение присутствия тем во времени [53]. Другие модификации могут учитывать такую переменную как авторство текста, ведь тексты одного автора имеют большую вероятность относиться к определённому набору тем (Author LDA) [54].

Параллельно множеству обобщений, существует две основных разновидности методов LDA, отличающихся методами оценивания, т. е. нахождения значения параметров модели, при кото-

рых наблюдаемая обучающая выборка максимально правдоподобна [55], [56, стр. 1]. Первая разновидность – вариационная модель LDA, чья численная схема основана на принципе максимизации функции правдоподобия. В рамках данной модели реализовано предположение о том, что одна функция Дирихле описывает лишь одно распределение (одного слова по темам или одного документа по темам); соответственно поиск распределение каждого слова и каждого документа по темам приводит к работе с огромными матрицами. Таким образом размерность матриц существенно зависит от размера словаря, поэтому качественный препроцессинг документов играет важную роль в тематическом моделировании. Кроме того, наличие произведение большого числа функции приводит к множеству локальных максимумов в функции правдоподобия. Таким образом, метод максимального правдоподобия может приводит не к оптимальным результатам, так как этот метод лишь даёт гарантия попадания в один из локальных максимумов, но не позволяет находить наибольший максимум среди множества локальных экстремальных точек.

Второй разновидностью метода LDA является метод сэмплирования Гиббса – статистический алгоритм на основе методов Монте-Карло, в котором строится марковская цепь, сходящаяся в апостериорному распределению тем, по которым далее строятся оценки параметров. Сэмплирование Гиббса позволяет эффективно находить скрытые темы в больших корпусах текстов. Сложно сказать, какой из двух подходов лучше. Многое зависит от особенностей конкретной реализации.

В данном исследовании используется подход, разработанный Мэтью Хоффманом [56] и реализованный в программе Gensim[10]. Он относится к первой группе алгоритмов – вариационной модели LDA. Данный выбор обусловлен тем, что в рамках выбранных инструментов эта программа является самым популярной и хорошо документированным вариантом.

Написать более простым языком, как конкретно работает алгоритм. Связать с байесом из первой части

2.4.2. Подготовка данных

Прежде, чем приступать к тематическому моделированию, необходимо произвести предварительную обработку данных, специфичную для данного этапа, а именно удаление редко встречающихся токенов. До обработки мы имеем 118718 уникальных токенов, что может быть причиной долго работы алгоритма. Однако токены, встречающиеся в корпусе всего лишь один раз не влияют на построение тематической модели, так что мы легко можем от них избавиться, сократив количество уникальных токенов до 69447. Удалённые токены в представляли собой слова с ошибками, цифры, гиперссылки, английские слова (в том числе написанные транслитом), имена собственные и просто редкие слова.

[10] https://radimrehurek.com/gensim/models/ldamulticore.html

2.4.3. Определение оптимального количества тем и их идентификация

Определение оптимального числа тем – важная подзадача в тематическом моделировании, поскольку её решение существенно влияет на осмысленность получаемого набора тем. Занижение числа тем приводит к чрезмерно общим результатам. Завышение же чревато сложностями интерпретации. Оптимальное число тем зависит от числа документов в анализируемом корпусе: в малых корпусах оптимальным является, как правило меньшее число тем. Согласно оригинальному исследованию [50], оптимальное число тем для корпуса из 16333 новостных статей составило 100, тогда как для корпуса из 5225 аннотаций научных статей – 50. Однако не существует однозначного метода определения оптимального количества тем, и часто это количество определяется «на глазок», исходя из личного мнения исследователя.

В данном исследовании первым был опробован метод определения оптимального количества тем на основе перплексии – это стандартный способ оценки качества модели. Перплексия равняется экспоненте от минус усреднённого логарифма правдоподобия и показывает, насколько хорошо модель приближает наблюдаемые частоты появления слов в документах. Качество модели тем выше, чем меньше перплексия.

Для измерения перплексии необходимо разделить выборку на две части – тренировочную – которая будет использоваться при построении модели, и текстовую, на которой будет проверяться точность предсказаний модели. В данном исследовании контрольную выборку составляли 10% случайно выбранных документов, остальные использовались для тренировки модели. Модели рассчитаны для количества тем от 5 до 100 с шагом в 5.

Используя стандартные методы расчёта перплексии из программы Gensim мы получили результаты, показанные на рисунке 2.2.

Как видно из графика, в нашем случаем по мере увеличения количества тем перплексия также увеличивается, в то время как должен происходить обратный процесс – большее количество тем лучше описывает распределение. Скорее всего это недостатки реализации расчёта перплексии в Gensim, поскольку сам автор программы признаёт наличие проблемы у некоторых пользователей[11].

Попробуем рассчитать перплексию с помощью другого инструмента и используем для этого популярную программу для тематического моделирования Mallet[12]. Как упоминалось ранее, данная программа использует совершенно другой подход к тематическому моделированию, поэтому график 2.3, полученный в ней, сильно отличается от предыдущего. Как видно из графика, мы получили несколько локальных минимумов перплексии при 45, 60 и 85 темах.

Какие могут быль альтернативы расчёту перплексии? Во-первых, можно использовать алгоритмы тематического моделирования, которые автоматически подбирают оптимальное количество тем. Таким алгоритмом является, например, иерархический процесс Дирихле (hierarchical Dirichlet process, HDP), который напоминает LDA с то разницей, что данный подход относится

[11] https://groups.google.com/d/msg/gensim/TpuYRxhyIOc/JbTjqCcC6uYJ
[12] http://mallet.cs.umass.edu

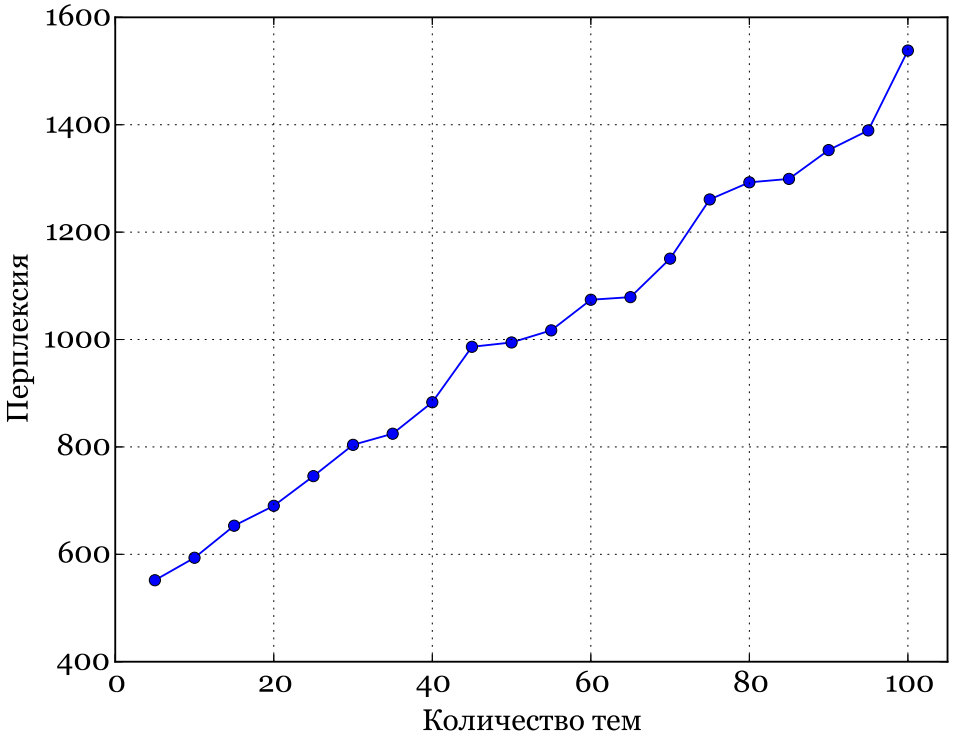

Рисунок 2.2: Изменение перплексии в зависимости от количества тем в программе Gensim

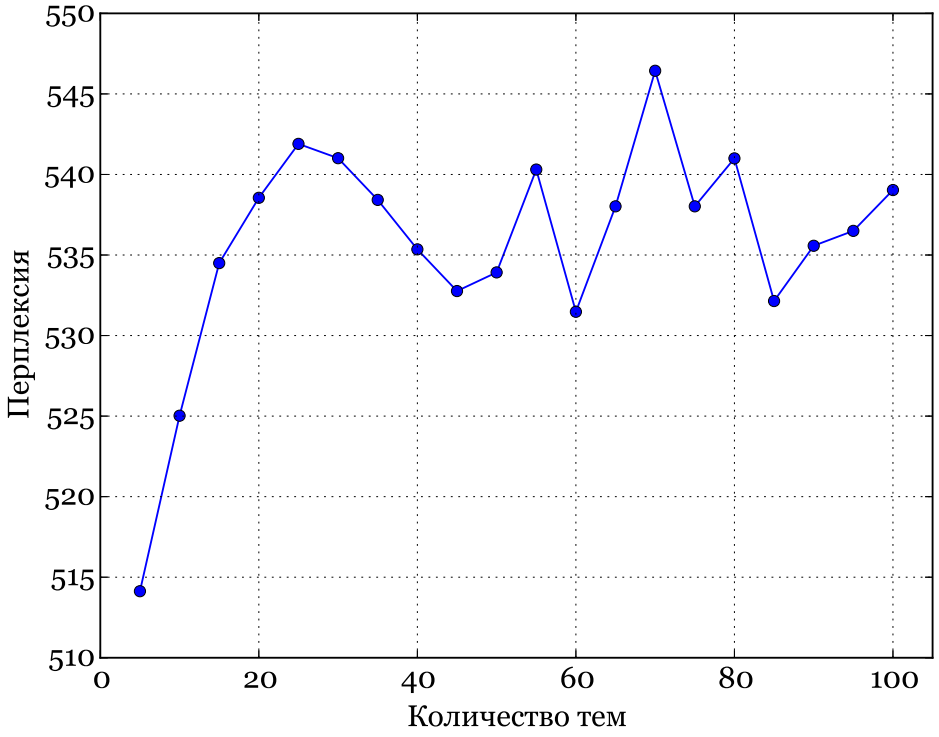

Рисунок 2.3: Изменение перплексии в зависимости от количества тем в программе Mallet

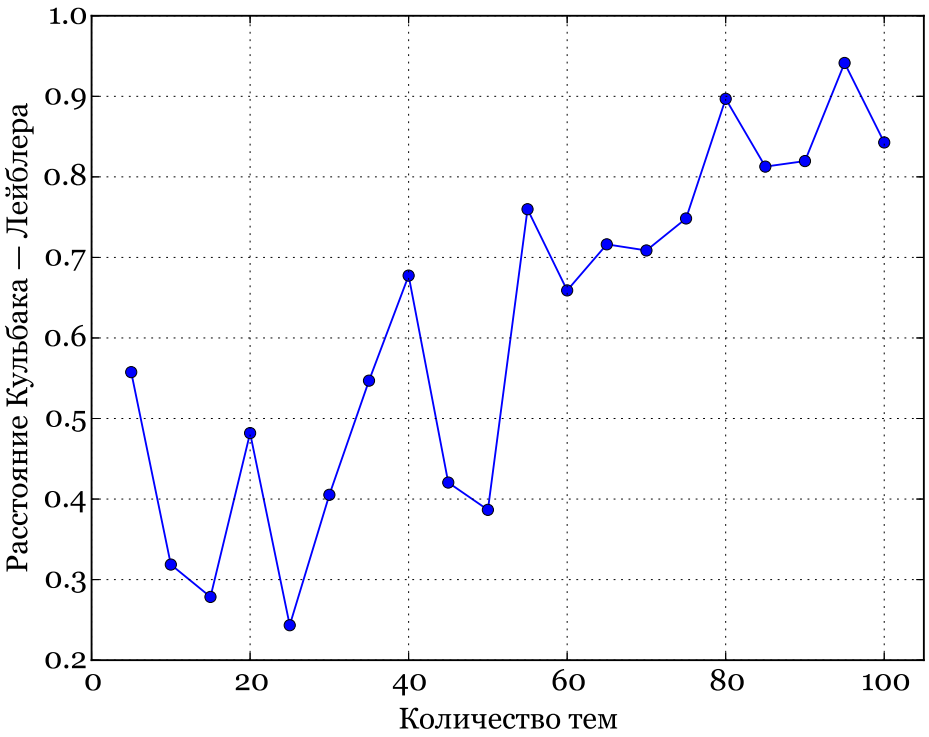

Рисунок 2.4: Изменение расстояния Кульбака — Лейблера в зависимости от количества тем в программе

к непараметрическим, а модель сама определяет оптимальное количество тем. Так как в Gensim присутствует реализация данного алгоритма, не составит труда применить его на нашей выборке.

В результате иерархического процесса Дирихле мы получили более 500 тем. Однако данное количество тем довольно сложно интерпретировать, ведь нам необходимо проанализировать каждую тему и дать ей название.

Ещё один опробованный нами способ решения данной задачи описан в статье под названием «О нахождении естественного числа тем в LDA: некоторые наблюдения» [57]. В ней автор предлагает использовать расстояние Кульбака — Лейблера как способ оценки качества модели. Чем меньше указанное расстояние, тем лучше модель. В результате расчёта этого расстояния на моделях с разным количеством тем, построенных в Gensim, мы получили график, показанный на рисунке 2.4.

Из него видно, что оптимальное количество тем равняется 15, 25, 50.

В конечном итоге после сравнения моделей с разным количеством тем, мы выбрали модель с 50-ю темами, поскольку сгенерированные ей темы легче всего подвергались интерпретации и сильнее всего отличались друг от друга.

Полученные темы и их интерпретация представлены в приложении А. Как видно, в омских Интернет-СМИ представлен широкий спектр тем: от достаточно частных, касающихся ареста бывшего вице-мэра Юрия Гамбурга или убийства боксёра Ивана Климова, до взаимоотношений с Украиной и США.

Название темы определялись после анализа слов, которые эта тема генерирует с наибольшей вероятностью. Ниже показано, как программа описывает одну из тем. Рядом к каждым словом указана вероятность, с которой оно генерируется данной темой. Из этого распределения становится понятно, что данная тема имеет отношение к прогнозу погоды в городе Омске. Однако, как мы увидим позже, не все темы можно так легко интерпретировать.

`0.030*омск + 0.018*температура + 0.017*день + 0.015*снег + 0.014*погода + 0.014*воздух + 0.012*градус + 0.011*ветер + 0.010*область + 0.010*днём + 0.009*ожидаться + 0.009*дождь + 0.008*ночью + 0.007*выходной + 0.006*составить + 0.006*неделя + 0.006*управление + 0.006*м/с + 0.005*атмосферный + 0.005*тёплый`

Также мы можем рассчитать вероятностное тематическое распределение для каждого отдельного документа, выявив наиболее связанные с ним темы. Так как в LDA используется нечёткая кластеризация, каждый документ с определённой вероятностью можно отнести к любой теме. В связи с этим необходимо определить порог, который будет служить ориентиром для отнесения документа к каким-либо темам.

Итак, о чём пишут в омских Интернет-СМИ? Для ответа на этот вопрос, надо определить критерий, на основании которого среди множества тем, будет выбрана та, которая будет считаться основной для данного документа или множества документов. Самый просто способ – отнесение документа к теме, в которую он попадает с наибольшей вероятностью. Таким образом мы получим, что самая популярная тема, которая наиболее вероятная для 1855 документов – это ДТП (тема № 19), вторая по популярности, к которой относятся 1845 документов – преступления (тема №43), третья (1609 документов) – взаимоотношения с Украиной (тема №39).

Этот способ хорошо подходит для определения темы отдельного документа, но если мы хотим таким способом оценить тематическое распределение на некотором множестве документов, то мы упустим важное преимущество LDA – нечёткую кластеризацию, а именно возможность отнесения документа сразу к нескольким темам. Поэтому для выявления наиболее популярной темы разумно рассчитать среднюю вероятность для каждой темы путём сложения её вероятностей для всех документов и деления получившейся суммы на количество документов, как показано в формуле 2.1.

$$X_b = \frac{\sum_{a=0}^{A} prob_{ab}}{A} \tag{2.1}$$

где A – общее количество документов, a – номер документа, b – номер темы, X_b – значение популярности для темы b, $prob_{ab}$ – вероятность присутствия темы b в документе a.

В таком случае самыми популярными у нас будут темы под номерами 19 (ДТП), 43 (преступления) и 4 (пожары). Этот способ мы считаем предпочтительным. Полные результаты представлены в таблице C.1 приложения B

Ещё один способ решения задачи поиска наиболее популярной темы во множестве документов – объединение текстов данных документов и поиск вероятностного распределения для нового

большого текста. При таком подходе на первый план вышли темы 2 (сложно интерпретировать), 39 (Украина), 1 (региональная власть).

Здесь мы встречаемся с такой проблемой, как сложность интерпретации некоторых выделенных тем. В нашем случае таких тем написать количество, а одна из них – та самая тема номер 2 – к тому же очень распространена. Проанализировав слова, которые она генерирует мы видим, что в них сложно найти что-то общее:

`0.009*человек + 0.007*большой + 0.006*нужно + 0.005*город + 0.005*омск + 0.005*время + 0.005*деньги + 0.005*сделать + 0.004*хороший + 0.004*вопрос + 0.004*делать + 0.004*знать + 0.004*проблема + 0.004*журналист + 0.004*работа + 0.004*работать + 0.004*должный + 0.003*проект + 0.003*метро + 0.003*думать`

К тому же вероятности, которыми тема генерирует данные слова чрезвычайно низки. Самая большая вероятность находится на уровне 0.009, в то время как в других темах примерно от 0.2 до 0.6

Одна из причин этому – большое количество слов, которые ничего не могут сказать нам об особенностях темы. В основном это прилагательные и глаголы, которые обозначают признак предмета или его действие, но не называют сам предмет (большой, нужно, сделать, хороший и др.). Возможно, часть этих слов стоило занести в список стоп-слов.

Анализ документов, в которых проявление этой темы наиболее вероятно, также показывает сложность её интерпретации. Вот примеры заголовков некоторых из этих документов: «Обзор блогов. Блоги – это маленькая жизнь», «Сколько ещё простоят хрущевки в России?», «Обзор СМИ: Страшно далеки они от народа», «Кустурица стоя аплодировал омским рокерам».

Наличие таких «мусорных» тем – нормальное явление в тематическом моделировании, которого тем не менее надо старательно избегать, проводя качественный препроцессинг документов и выбирая оптимальное количество тем для модели. В нашем случае сложности возникли с двумя темами, что можно считать неплохим результатом.

2.5. Анализ комментариев

2.5.1. Общая характеристика

Переходя к комментариям, мы вначале дадим общую их характеристику в данном корпусе. К 26783 статьям из 33877 пользователи оставили 258121 комментариев – в среднем этот составляет 7.6 комментария на статью.

Относительно распределения комментариев во времени, можно сказать, что оно ожидаемо практически полностью повторяет распределение статей. Подробнее на рисунке C.1 приложения C.

2.5.2. Комментируемость тем

Определим самые резонансные темы, подсчитав, статьи какой тематики комментируют чаще всего, а какой – реже. Для этого отнесём каждый документ к одной из тем, на основании того, к какой теме он принадлежит с наибольшей вероятностью. Затем рассчитаем отношение общего числа комментариев к статьям данной тематики к количеству этих статей. Полученное число и будет являться индикатором резонансности темы.

Как и в случае с расчётом наиболее популярной темы, помня что каждый документ можно отнести ко многим темам, мы можем внести улучшения в эту формулу. Более валидные результаты можно получить, рассчитав показатель комментируемости темы путём сложения рассчитанных для каждого документа произведений вероятности присутствия темы в документе на количество комментариев в данном документе. Для того, чтобы нивелировать влияние размера темы, разделим получившееся таким образом для каждой темы значение на рассчитанный ранее показатель популярности темы. Модель расчёта показателя комментируемости представлена в формуле 2.2:

$$Y_b = \frac{A \sum_{a=0}^{A} prob_{ab} * qcomments_a}{\sum_{a=0}^{A} prob_{ab}} \tag{2.2}$$

где A – общее количество документов, B – общее количество тем, a – номер документа, b – номер темы, Y_b – значение комментируемости для темы b, $prob_{ab}$ – вероятность присутствия темы b в документе a, $qcomments_a$ – количество комментариев в документе a.

Полученное таким образом значение само по себе почти ничего не значит, важно лишь то, как оно различается от темы к теме. Поэтому для наглядности мы можем без последствий принять наибольшее значение комментируемости за 100%, а для остальных тем рассчитать долю, которую они составляют от этих 100%.

Таким образом было выявлено, что самыми комментируемыми темам являются темы, связанные с Украиной, домашними животными и арестом первого вице-мэра Омска Юрия Гамбурга. Безразличнее всего читатели отнеслись к статьям, посвящённым продаже автомобилей, банкам и кредитам и чрезвычайным происшествиям. Полные данные представлены в таблице **??** приложения C.

2.5.3. Анализ тональности комментариев

Общая теория

Анализ тональности – ещё одна сфера интеллектуального анализа текста. Да данном этапе мы собираемся оценить тональность комментариев к различным темам. Гипотеза заключается в том, что если существует какая-то социальная напряжённость в обществе по отношению к какой-либо теме, то это находит выражение в комментариях к статьям соответствующей тематики. Наша задача, таким образом, состоит в том, чтобы найти темы, которые вызывают социальную напряжённость.

Существует несколько подходов к классификации тональности. Первый подход основан на наборе правил, применяя которые система делает заключение о тональности текста. Например, для предложения «Я люблю кофе», можно применить следующее правило: если сказуемое («люблю») входит в положительный набор глаголов («люблю», «обожаю», «одобряю» ...) и в предложении не имеется отрицаний, то классифицировать тональность как «положительная». Многие коммерческие системы используют данный подход, несмотря на то что он требует больших затрат, т.к. для хорошей работы системы необходимо составить большое количество правил. Зачастую правила привязаны к текстам определённой тематики (например, «ресторанная тематика») и при смене тематики («обзор фотоаппаратов») требуется заново составлять правила. Тем не менее, этот подход является наиболее точным при наличии хорошей базы правил.

Следующий подход основан на машинном обучении, чаще всего с учителем. В этом случае необходимо разметить некоторое количество текстов, на которых обучается подстроенная с помощью каких-либо алгоритмов модель. Часто для этого используется обычный байесовский классификатор. В дальнейшем эта модель распределяет тексты по заданным категориям. Это достаточно простой метод, а потому он широко распространён. Недостатками метода является невысокая точность ($\approx 70\%$) и необходимость ручной разметки обучающей выборки.

Подходы, основанные на словарях, используют так называемые тональные словари (affective lexicons) для анализа текста. В простом виде тональный словарь представляет из себя список слов со значением тональности для каждого слова. Каждому слову из словаря, встречающемуся в тексте присваивается соответствующее значение, а затем вычисляется общая тональность текста.

Подготовка программы и словарей

Существует не так много бесплатных программ, предназначенных для анализа тональности текста. Ещё меньше из них – а в действительности ни одна из них – умеют адекватно определять тональность текстов на русском языке. В таких сложных условиях наилучшим вариантом нам виделась программа SentiStrength[13] за авторством Майкла Фелвола[14]. Будучи бесплатной для некоммерческого использования, данная программа может работать почти с любым языком и, по крайней мере со стандартными словарями для английского языка, показывает хорошие результаты [58]. Для работы с другими языками, необходимо загрузить в неё тональные словари на нужных языках. Основания часть этих словарей представляют собой простой текстовый файл со списком слов, к каждому из которых в поставлена в соответствие оценка позитивной (по шкале от 1 до 5) или негативной составляющей (по шкале от -1 до -5). Большее значение соответствует большей выраженности эмоциональной составляющей. Другие части словаря, также представляющей собой текстовые файлы, содержат список слов-усилителей, которые усиливают значение тональности для слова, на которое они действуют («очень плохой» будет иметь более негативную оценку, чем просто «плохой»), идиоматические выражения, слова-отрицания, смайлы, вопросительные

[13] http://sentistrength.wlv.ac.uk/

[14] Глава Statistical Cybermetrics Research Group университета Вулверэмптона, ассоциированный научный сотрудник Oxford Internet Institute, Великобритания.

слова, сленговые слова и слова обозначающие иронию. Все эти части учитываются алгоритмом и помогают достичь более точного результата. Результат выдаётся в виде двух оценок – оценка позитивной составляющей текста (по шкале от +1 до +5) и оценка негативной составляющей (по шкале от -1 до -5) или в виде бинарной оценки (позитивный/негативный текст).

Но в этих словарях и кроются самые большие сложности использования программы. Первая версия словарей для русского языка[15], созданная факультетом прикладной лингвистики Санкт-Петербургского государственного университета аэрокосмического приборостроения, не отличается хорошим качеством. Из-за излишне общих правил данные словари часто оценивали нейтральные слова как эмоционально окрашенные[16]

Вероятно, по этой причине для своего исследования тональности комментариев к постам в Живом Журнале коллективом Лаборатории Интернет-Исследований был создан новый тональный словарь[17] [59]. Процесс адаптации включал в себя перевод англоязычного словаря, на основе которого работает ПО, на русский язык, подбор подходящих русских эквивалентов полученным словам, составление частотного словаря на основе комментариев к постам ЖЖ, включение частотных слов в словарь и кодирование словаря по шкале эмоциональности от -5 до 5. После сопоставления результатов работы программы с результатами ручного кодирования был сделан вывод, что количество совпадений между автоматическим кодированием с данным словарём и ручным кодированием с помощью экспертов значительно уступает результатам аналогичных экспериментов на английском языке.

По сравнению с первым вариантом словаря, второй вариант обладал прямо противоположной проблемой – тенденцией оценивать негативно эмоционально окрашенные тексты как нейтральные[18] [59, стр. 3].

---

[15] http://sentistrength.wlv.ac.uk/SentStrength_Data/russian

[16] Дело в том, что в словарях к SentiStrength для создания простых правил можно использовать оператор *, который обозначает любое количество любых символов кроме пробела. Под шаблон «*плом*», например, подходят слова «**плом**ба», «ди**плом**ированный» и др. В своих словарях авторы переусердствовали с применением данного оператора. Это привело, например, к тому, что словари дают три балла негативной эмоциональной составляющей любому слову, начинающемуся на «ад» (приравнивая **ад**министраторов к исчадиям **ад**а) и два балла позитивной словам на «мил» (а это далеко не только слово «**мил**ый», как вероятно задумывали авторы словарей, но вызывающее не самые позитивные эмоции слово «**мил**иция»).

[17] http://sentistrength.wlv.ac.uk/SentStrength_Data/russian2

[18] Нами было выделено несколько вероятных причин проблем у второго варианта словаря. Главная из них заключается в том, что словарь состоит только из слов в нормальной форме и оператор * в них не используется. Поэтому для оценки текстов с использованием этого словаря необходима предварительная лемматизация. Сотрудники Лаборатории Интернет-исследований, естественно, знали об особенностях своих словарей и не обошли вниманием этот этап.

Здесь, однако, стоит сказать, что лемматизация достаточно плохо работает на словах, которые ярче всего свидетельствуют о негативных эмоциях – обсценной лексике. Для лемматизации необходим словарь со всеми словоформами, но мат в эти словари включают редко.

К тому же, многие программы, производящие лемматизацию (тот же используемый нами pymorphy2), при приведении слова к нормальной форме не отбрасывают при этом приставки («набрал» → «набрать», а не «брать»), которые играют огромную роль в обсценной лексике. В такой ситуации необходимо включать в словарь слова обсценной лексики со всеми вариантами приставок, что сложно при наличии проблемы, указанной в предыдущем абзаце.

Ещё одна возможна причина неудовлетворительного результата работы второго варианта словаря может заключаться в том, что для перед лемматизацией необходима токенизация. Но SentiStrength предназначена для анализа «сырого» текста сама производит токенизацию по своему специфическому алгоритму. Так конструкция «какое-либо-слово))))» благодаря тому, что смайл и слово являются одним целым, будет присвоено два балла положительного эмо-

Также общим недостатком рассматриваемых русскоязычных словарей является практически полное отсутствие в них идиоматических выражений, сленговых и ироничных слов.

В условиях невысокого качества словарей, было принято решение о создании нового тонального словаря. В целях экономии ресурсов было решено построить его на основе предыдущих версий, учтя их достоинства и недостатки, с добавлением специфичных для исследуемых текстов эмоционально окрашенных слов.

Моделью для построения нового словаря стал первый словарь, основанный на правилах. Из него были удалены правила, касающиеся слов меньше, чем из четырёх символов и допускающие некорректные соответствия словам (правило ад* допускало соответствие слову администратор). Затем правила были преобразованы в обычные слова (ад). Затем все слова (только слова, не правила) с помощью программы pymorphy2, были развёрнуты в свои всевозможные словоформы (ад, ада, адом и др.) Таким образом явное указание подходящих слов позволило уйти от слишком общих правил.

Второй словарь, который, напомним, состоял только из нормальных форм и требовал лемматизации поступающих текстов, было решено преобразовать к формату первого. Для этого длинные слова данного словаря были пропущены через процедуру стемминга с последующей заменой усечённых частей на оператор *. Для более коротких слов, как и в предыдущем случае, были найдены все словоформы.

Затем получившиеся словари были объединены и к ним добавились 113 тональных слов, выделенных на основании изучения более двухсот случайных комментариев из исследуемой выборки и отсутствующих в исходных словарях. Некоторые их этих слов специфичны для текстов, написанных в исследуемый промежуток времени (например, слово «укроп»). Баллы данным словам выставлялись на основе произвольного мнения автора, что, конечно, нельзя признать методологически корректным.

Конструирование нового словаря закончилось удалением повторяющихся записей.

Новый словарь обладает следующими преимуществами:

1. Он объединяет словарные базы двух разных словарей – а значит полнее каждого из них по отдельности.

2. В отличие от второго словаря, для пользования данным словарём не надо модифицировать входящие данные.

3. Правила нового словаря корректнее, точнее, чем правила первого словаря, а значит ложных срабатываний будет меньше.

4. Данный словарь включает слова, которых не было ни в одном из предыдущих.

ционального заряда. Но если отделить смайл от слова «какое-либо-слово))))», то программа просто не увидит здесь никакой эмоции.

Возможно, лучший результат дала бы замена лемматизации на стемминг. Так как для стемминга не нужна база основ слов, а лишь правила и набор морфем, проблема обсценных слов была бы решена.

Далее необходимо оценить эффективность определения тональности с использованием нового словаря по сравнению с предшественниками. Для этого была использована коллекция цитат из новостного потока с разметкой по оценочной тональности[19], предоставленная РОМИП. Данная коллекция состоит из 4260 оценок. Оценка может принимать одно из 4-х значений: положительная, отрицательная, смешанная и нет оценки. Для простоты тестирования тексты со смешанными оценками были исключены.

Вариант словаря созданный факультетом прикладной лингвистики на тестовой коллекции показал точность оценивания в 47%. Использование созданного нами словаря позволило увеличить точность оценивания до 50%. Много это или мало?

Для начала следует сказать, что тестовую выборку составляли отрывки из новостных статей. Сравнивая эти рафинированные тексты, написанные профессиональными журналистами с реальными комментариями, следует заметить, что эмоциональная тональность последних, как правило, выражена гораздо чётче. Это замечание позволяет предположить, что в приложении к реальным комментариям эффективность оценивания несколько повысится.

Далее отметим, что по результатам тестов, проведённых автором программы, на англоязычных текстах с использованием соответствующих словарей SentiStrength показывается точность в 60% (что лучше, чем большинство других программ) [58]. Исследователям из ВШЭ удалось добиться точности работы на русскоязычных текстах от 40% до 48% [60, стр. 49].

Исходя из сказанного, нельзя считать полученный нами результат в 50%, особенно с учётом характера тестовой коллекции, абсолютно провальным. Необходимо понимать, что даже для английского языка точность определения тональности выше 70% считается практически идеальной. Однако и хорошим признать полученный результат тоже нельзя. Алгоритмы, основанные на машинном обучении, по оценке того же Майкла Фелвола, позволяют добиться точности до 58,5%. И это при том, что эффективность данных методов не зависит от языка. Их минусом, как уже говорилось ранее, является необходимость в обучающей выборке и более высокая сложность применения.

Анализ тональности

Изменить заголовок

Настало время задействовать полученные словари и посмотреть различия в настроениях комментирующих от темы к теме.

[19] http://romip.ru/ru/collections/sentiment-news-collection-2012.html

Заключение

Основные результаты работы заключаются в следующем.

1. На основе анализа …
2. Численные исследования показали, что …
3. Математическое моделирование показало …
4. Для выполнения поставленных задач был создан …

И какая-нибудь заключающая фраза.

Список литературы

1. *Дюк В. А., Флегонтов А. В., Фомина И. К.* Применение технологий интеллектуального анализа данных в естественнонаучных, технических и гуманитарных областях // *Известия Российского государственного педагогического университета им. А.И. Герцена*. — 2011. — № 138. — С. 77–84.

2. Прогноз выборов в Венесуэле. — Доступ: 2014-08-25. Режим доступа: http://vox-populi.ru/venezuala.phtml.

3. *Asur Sitaram, Huberman Bernardo A.* Predicting the Future With Social Media. — Available: http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf.

4. *Давыдов А. А.* Knowledge Discovery and Data Mining в системной социологии. — 2013. — Режим доступа: http://www.isras.ru/Davydov_Knowledge.html.

5. *Давыдов А. А.* Фатальная ошибка социологии. — 2010. — 04. — Режим доступа: http://ecsocman.hse.ru/text/28973359/.

6. *Орлов А. И.* Черная дыра отечественной социологии. — Режим доступа: http://www.ssa-rss.ru/index.php?page_id=19&id=456.

7. *Schrodt Philip A.* Seven Deadly Sins of Contemporary Quantitative Political Analysis // APSA 2010 Annual Meeting Paper. — 2010. — Available: http://eventdata.psu.edu/7DS/Schrodt.7Sins.APSA10.pdf.

8. *Silver Nate*. The Signal and the Noise: Why So Many Predictions Fail – but Some Don't. — Barnes & Noble, 2012.

9. *Nisbet Robert, Elder John, Miner Gary*. Handbook of statistical analysis and data mining applications. — Academic Press, 2009. — P. 864.

10. Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians / R. Christensen, W. Johnson, A. Branscum, T. Hanson. — 1 edition. — CRC Press, 2010. — P. 516.

11. *Зельнер А*. Байесовские методы в эконометрии. — Москва: Статистика, 1980.

12. *Гнеденко Б. В.* Курс теории вероятностей. — 8 изд. — Москва: Едиториал УРСС, 2005.

13. *Efron Bradley*. Modern Science and the Bayesian-Frequentist Controversy. — 2005. — Available: http://www-stat.stanford.edu/~ckirby/brad/papers/2005NEWModernScience.pdf.

14. *Talbott William*. Bayesian Epistemology // The Stanford Encyclopedia of Philosophy / Ed. by Edward N. Zalta. — 2013.

15. *Айвазян С. А., Мхитарян В. С.* Прикладная статистика. Основы эконометрики. — 2-е, исправленное изд. — Москва: Юнити-Дана, 2001. — Т. 1. — С. 656. — Режим доступа: http://ecsocman.hse.ru/text/33442857.

16. *Айвазян С. А.* Байесовский подход в эконометрическом анализе // *Прикладная эконометрика*. — 2008. — № 1(9). — С. 93–130. — Режим доступа: http://pe.cemi.rssi.ru/pe_2008_1_93-130.pdf.

17. *Jeffreys Harold*. Theory of Probability. — 3 edition. — Oxford: Clarendon Press, 1983.

18. *Бастиан Хильда*. Роль статистической значимости в неудачах науки // *Scientific American*. — 2013. — Режим доступа: http://inosmi.ru/world/20131114/214743342.html.

19. *Wilhelm Adalbert*. Handbook of Computational Statistics: Concepts and Methods / Ed. by J. E. Gentle, W. Härdle, Y. Mori. — Springer, 2004. — Pp. 789–803.

20. *Han Jiawei, Kamber Micheline*. Data Mining: Concepts and Techniques / Ed. by Jim Gray. — 2 edition. — Elsevier, 2006.

21. Анализ данных и процессов / А. А. Барсегян, М. С. Куприянов, И. И. Холод и др. — 3 изд. — БХВ-Петербург, 2009.

22. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications / G. Miner, D. Delen, J. Elder et al. — Elsevier, 2012.

23. *C. Tetlock Paul*. Giving content to investor sentiment: The role of media in the stock market // *The Journal of Finance*. — 2007. — June. — Vol. 62, no. 3. — Pp. 1139–1168. — Available: http://www0.gsb.columbia.edu/faculty/ptetlock/papers/Tetlock_JF_07_Giving_Content_to_Investor_Sentiment.pdf.

24. *Archak Nikolay, Ghose Anindya, Ipeirotis Panagiotis*. Deriving the pricing power of product features by mining consumer reviews // *Management Science*. — 2011. — August. — Vol. 57, no. 8. — Pp. 1485–1509. — Available: http://pages.stern.nyu.edu/~aghose/pricingpower_print.pdf.

25. *Askitas Nikolaos, Zimmermann Klaus F.* Google econometrics and unemployment forecasting // *Applied Economics Quarterly*. — 2009. — April. — Vol. 55, no. 2. — Pp. 107–120. — Available: http://ftp.iza.org/dp4201.pdf.

26. *Tausczik Yla R., Pennebaker James W.* The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods // *Journal of Language and Social Psychology*. — 2010. — Vol. 29, no. 1. — Pp. 24–54. — Available: http://homepage.psy.utexas.edu/HomePage/Faculty/Pennebaker/Reprints/Tausczik&Pennebaker2010.pdf.

27. *Golder Scott A., Macy Michael W.* Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures // *Science*. — 2011. — September. — Vol. 333. — Pp. 1878–1881. — Available: http://www3.ntu.edu.sg/home/linqiu/teaching/psychoinformatics/DiurnalandSeasonalMoodVaryAcrossDiverseCultures.pdf.

28. *Mosteller F., Wallace D.L., Nerbonne J.* Inference and Disputed Authorship: The Federalist. The David Hume Series. — Center for the Study of Language and Information, 2008. — Available: http://books.google.ru/books?id=g7wbAQAAMAAJ.

29. Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopedie / Russell Horton, Robert Morrissey, Mark Olsen et al. // *Digital Humanities Quarterly*. — 2009. — Vol. 3, no. 2. — Available: http://www.digitalhumanities.org/dhq/vol/3/2/000044/000044.html.

30. A latent variable model for geographic lexical variation / Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, Eric P. Xing // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. — 2010. — Pp. 1277–1287. — Available: http://www.cs.cmu.edu/~nasmith/papers/eisenstein+oconnor+smith+xing.emnlp10.pdf.

31. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series / Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, Noah A Smith // International AAAI Conference on Weblogs and Social Media. — Washington: 2010. — Available: http://www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf.

32. Media coverage in times of political crisis: a text mining approach / E. Junqué de Fortuny, T. De Smedt, D. Martens, W. Daelemans // *Expert Systems with Applications*. — 2012. — Октябрь. — Vol. 39.

33. *Causey Charles*. The Battle for Bystanders: Information, Meaning Contests, and Collective Action in the Egyptian Uprising of 2011. — 2012.

34. *Кольцова О. Ю., Павлова Ю.* К методологии сбора Интернет-данных для социологического анализа. — СПб, 2011. — Режим доступа: http://www.hse.ru/

data/2013/06/10/1283698963/JulijaPavlova,OlesjaKolcova,Kmetodol..dljasociologicheskogoanaliza.pdf.

35. *Иудин А.А., Рюмин А.М.* Контент-анализ текстов: компьютерные технологии: Учебное пособие. — Н.Новгород: Нижегородский государственный университет им. Н.И.Лобачевского, 2010. — Режим доступа: http://window.edu.ru/resource/004/74004/files/Ctt_3.pdf.

36. *Ландэ Д.В.* Основы интеграции информационных потоков. — Киев: Инжиниринг, 2006.

37. *Smith C. P.* Content analysis and narrative analysis // Handbook of research methods in social and personality psychology / Ed. by H. T. Reis, C. M. Judd. — Cambridge, UK: Cambridge University Press, 2000. — Pp. 313–335.

38. *Почепцов Г.Г.* Теория и практика коммуникации (от речей президентов до переговоров с террористами). — Москва: Центр, 1998.

39. *Осипов Г. В.* Рабочая книга социолога.

40. *Ho Yu Chong, Jannasch-Pennell Angel, DiGangi Samuel*. Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability // *The Qualitative Report*. — Vol. 16, no. 3. — Pp. 730–744. — Available: http://www.nova.edu/ssss/QR/QR16-3/yu.pdf.

41. *Аверьянов Л.Я.* Контент-анализ. — Москва: РГИУ, 2007.

42. *Морозова В. Н.* Методы политического анализа: Учебно-методическое пособие. — Воронеж, 2007.

43. *Papacharissi Zizi*. Audiences as Media Producers: Content Analysis of 260 Blogs // *Blogging, citizenship, and the future of media*. — 2007. — Available: http://tigger.uic.edu/~zizi/Site/Research_files/TremayneChapterBlogs.pdf.

44. Рейтинг АРИ омских интернет-СМИ. Сводка. — Доступ: 2014-08-25. Режим доступа: http://omsk-journal.ru/publ/9-1-0-116.

45. *Webster Jonathan J., Kit Chunyu*. Tokenization As the Initial Phase in NLP // Proceedings of the 14th Conference on Computational Linguistics - Volume 4. — COLING '92. — Stroudsburg, PA, USA: Association for Computational Linguistics, 1992. — Pp. 1106–1110. — Available: http://dx.doi.org/10.3115/992424.992434.

46. *Коршунов Антон, Гомзин Андрей*. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. — Т. 23. — ИСП РАН, 2012. — С. 215–244.

47. *Воронцов К. В.* Вероятностное тематическое моделирование. — 2013. — Октябрь.

48. *DiMaggio Paul, Nag Manish, Blei David*. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding // *Poetics*. — 2013. — 12. — Vol. 41, no. 6. — Pp. 570–606. — Available: http://www.theculturelab.umd.edu/uploads/1/4/2/2/14225661/exploitingaffinities_dimaggio.pdf.

49. Моделирование семантических связей в текстах социальных сетей с помощью алгоритмаLDA (на материале русскоязычного сегмента Живого Журнала) / О. А. Митрофанова, А. С. Шиморина, Кольцова О. Ю., Кольцов С. Н. // *Структурная и прикладная лингвистика*. — Т. 11.

50. *Blei David M., Ng Andrew Y., Jordan Michael I.* Latent Dirichlet Allocation // *J. Mach. Learn. Res.* — 2003. — March. — Vol. 3. — Pp. 993–1022. — Available: http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf.

51. Interval Semi-supervised LDA: Classifying Needles in a Haystack / Svetlana Bodrunova, Sergei Koltsov, Olessia Koltsova et al. // Advances in Artificial Intelligence and Its Applications / Ed. by Félix Castro, Alexander Gelbukh, Miguel González. — Vol. 1. — Springer Berlin Heidelberg, 2013. — November. — Pp. 265–274. — Available: http://www.hse.ru/data/2013/10/03/1277898420/micai2013-182-final-easychair.pdf.

52. Knowledge discovery through directed probabilistic topic models: a survey. / Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad // *Frontiers of Computer Science in China*. — 2010. — Vol. 4, no. 2. — Pp. 280–301. — Available: http://www.machinelearning.ru/wiki/images/9/90/Daud2009survey-rus.pdf.

53. *Wang Xuerui, McCallum Andrew*. Topics over Time: A non-Markov Continuous-time Model of Topical Trends // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '06. — New York, NY, USA: ACM, 2006. — Pp. 424–433. — Available: http://doi.acm.org/10.1145/1150402.1150450.

54. Learning Author-topic Models from Text Corpora / Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths et al. // *ACM Trans. Inf. Syst.* — 2010. — January. — Vol. 28, no. 1. — Pp. 4:1–4:38. — Available: http://doi.acm.org/10.1145/1658377.1658381.

55. *Ю. Кольцова О., Н. Кольцов С.* Статистический и тематический профиль «Живого журнала». — СПб.

56. *Hoffman Matthew D., Blei David M., Bach Francis R.* Online Learning for Latent Dirichlet Allocation. // NIPS / Ed. by John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor et al. — Curran Associates, Inc., 2010. — Pp. 856–864. — Available: https://www.cs.princeton.edu/~blei/papers/HoffmanBleiBach2010b.pdf.

57. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations / R. Arun, V. Suresh, C. E. Veni Madhavan, M. N. Narasimha Murthy // Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I. — PAKDD'10. — Berlin, Heidelberg: Springer-Verlag, 2010. — Pp. 391–402. — Available: http://dx.doi.org/10.1007/978-3-642-13657-3_43.

58. Sentiment in Short Strength Detection Informal Text / Mike Thelwall, Kevan Buckley, Georgios Paltoglou et al. // *J. Am. Soc. Inf. Sci. Technol.* — 2010. — 12. — Vol. 61, no. 12. — Pp. 2544–2558. — Available: http://www.scit.wlv.ac.uk/~cm1993/papers/SentiStrengthPreprint.pdf.

59. *Павлова Юлия Валерьевна, Кольцова Олеся Юрьевна*. Метод автоматического анализа тональности текста в применении к социологическим задачам: на примере анализа комментариев к постам Живого Журнала // Избранные тезисы докладов IV Студенческой социологической межвузовской конференции / Под ред. М. Р. Демин. — СПб: НИУ ВШЭ (Санкт-Петербург), 2013.

60. *Кольцов С. Н., Павлова Ю., Кольцова О. Ю.* Метод автоматического анализа тональности текста в применении к социологическим задачам. — Available: http://www.hse.ru/data/2013/09/27/1277458071/Metodicheskoe_posobie.doc.

Приложение А

Результаты тематического моделирования

0. Детская медицина

0.040*ребёнок + 0.020*больница + 0.019*девочка + 0.019*омск + 0.016*мальчик + 0.012*медицинский + 0.012*врач + 0.010*область + 0.009*женщина + 0.009*находиться + 0.009*помощь + 0.008*мать + 0.008*полиция + 0.007*родитель + 0.007*подросток + 0.007*состояние + 0.007*сообщить + 0.006*время + 0.006*проверка + 0.006*дом

1. Местная власть

0.034*омск + 0.028*губернатор + 0.026*назаров + 0.020*виктор + 0.018*область + 0.015*глава + 0.014*регион + 0.010*мэр + 0.008*вячеслав + 0.008*двораковский + 0.008*министр + 0.008*правительство + 0.007*сегодня + 0.007*вопрос + 0.007*россия + 0.006*заявить + 0.006*первый + 0.005*синюгин + 0.005*развитие + 0.005*президент

2. Сложно определить

0.009*человек + 0.007*большой + 0.006*нужно + 0.005*город + 0.005*омск + 0.005*время + 0.005*деньги + 0.005*сделать + 0.004*хороший + 0.004*вопрос + 0.004*делать + 0.004*знать + 0.004*проблема + 0.004*журналист + 0.004*работа + 0.004*работать + 0.004*должный + 0.003*проект + 0.003*метро + 0.003*думать

3. IT

0.013*система + 0.009*новый + 0.009*сайт + 0.009*сеть + 0.008*интернет + 0.008*связь + 0.007*мобильный + 0.006*информация + 0.006*оператор + 0.006*россия + 0.006*пользователь + 0.006*компания + 0.006*электронный + 0.006*tele2 + 0.006*услуга + 0.005*абонент + 0.005*время + 0.005*доступ + 0.004*космический + 0.004*дать

4. Пожары

0.024*пожар + 0.022*омск + 0.016*дом + 0.015*человек + 0.013*мчс + 0.012*пожарный + 0.010*область + 0.009*произойти + 0.009*место + 0.009*улица + 0.008*причина + 0.008*сообщить + 0.007*часы + 0.007*огонь + 0.007*результат + 0.006*сообщение + 0.006*мужчина + 0.006*сегодня + 0.006*происшествие + 0.006*возгорание

5. Сложно определить

0.025*омск + 0.012*область + 0.011*рейтинг + 0.011*компания + 0.009*омич + 0.009*место + 0.008*тариф + 0.007*регион + 0.007*оао + 0.007*город + 0.007*население + 0.006*сибирь

+ 0.006*житель + 0.006*2013 + 0.006*число + 0.006*рэк + 0.005*показатель + 0.005*тысяча + 0.005*россия + 0.005*уровень

6. Деятельность правоохранительных органов

0.030*омск + 0.018*россия + 0.015*полиция + 0.015*сотрудник + 0.012*область + 0.012*полицейский + 0.010*умвд + 0.007*гражданин + 0.007*проверка + 0.007*задержать + 0.006*пресс-служба + 0.006*наркотик + 0.006*алкоголь + 0.006*территория + 0.005*обнаружить + 0.005*изъять + 0.005*омич + 0.005*дело + 0.005*административный + 0.005*сообщить

7. Экономика области

0.030*омск + 0.016*область + 0.012*предприятие + 0.011*регион + 0.010*развитие + 0.009*производство + 0.009*продукция + 0.008*проект + 0.007*завод + 0.007*компания + 0.006*продукт + 0.005*россия + 0.005*рынок + 0.005*бизнес + 0.004*новый + 0.004*цена + 0.004*предприниматель + 0.004*хозяйство + 0.004*комплекс + 0.004*реализация

8. Банковский сектор

0.024*компания + 0.020*банка + 0.019*клиент + 0.019*магазин + 0.017*банк + 0.013*кредит + 0.010*карта + 0.008*новый + 0.007*покупка + 0.007*услуга + 0.007*рубль + 0.007*кредитный + 0.006*салон + 0.006*россия + 0.006*банковский + 0.006*продажа + 0.006*сеть + 0.006*товар + 0.006*плюс + 0.006*покупатель

9. Праздники, свадьбы

0.010*день + 0.007*праздник + 0.006*подарок + 0.005*... + 0.004*хороший + 0.004*большой + 0.004*костюм + 0.004*девушка + 0.004*кольцо + 0.004*друг + 0.004*время + 0.004*пара + 0.004*земля + 0.003*цвета + 0.003*новый + 0.003*сделать + 0.003*женщина + 0.003*гость + 0.003*зоопарк + 0.003*брак

10. Суды

0.026*суд + 0.019*омск + 0.016*рубль + 0.015*прокуратура + 0.014*тысяча + 0.011*нарушение + 0.008*область + 0.007*проверка + 0.007*дело + 0.007*штраф + 0.007*требование + 0.006*закон + 0.006*россия + 0.006*признать + 0.006*решение + 0.006*прокурор + 0.006*размер + 0.006*лицо + 0.006*районный + 0.005*срок

11. Организация движения и общественный транспорт

0.069*улица + 0.024*транспорт + 0.020*движение + 0.019*маршрут + 0.019*автобус + 0.016*омск + 0.010*№ + 0.010*проспект + 0.009*пробка + 0.009*маркс + 0.008*участок + 0.008*часы + 0.007*город + 0.007*департамент + 0.007*остановка + 0.006*транспортный + 0.006*ленин + 0.005*путь + 0.005*работа + 0.005*поворот

12. Погода

0.030*омск + 0.018*температура + 0.017*день + 0.015*снег + 0.014*погода + 0.014*воздух + 0.012*градус + 0.011*ветер + 0.010*область + 0.010*днём + 0.009*ожидаться + 0.009*дождь + 0.008*ночью + 0.007*выходной + 0.006*составить + 0.006*неделя + 0.006*управление + 0.006*м/с + 0.005*атмосферный + 0.005*тёплый

13. Дело Юрия Гамбурга[1]

[1] Бывший вице-мэр города Омска. Арестован по обвинению в коррупции.

0.028*гамбург + 0.014*юрий + 0.014*военный + 0.010*суд + 0.009*пенсионный + 0.007*адвокат + 0.007*пенсия + 0.006*дело + 0.006*часть + 0.006*отношение + 0.006*министр + 0.006*имущественный + 0.006*вице-губернатор + 0.005*следствие + 0.005*первое + 0.005*меренкова + 0.005*решение + 0.005*арест + 0.005*оборона + 0.005*находиться

14. Военные учения

0.016*учение + 0.016*ракета + 0.015*корабль + 0.012*вопрос + 0.011*чёрный + 0.010*мор + 0.009*море + 0.008*ракетный + 0.007*ответ + 0.007*полигон + 0.006*флаг + 0.006*ип + 0.006*турецкий + 0.005*адрес + 0.005*портал + 0.005*боевой + 0.005*вооружение + 0.005*вид + 0.004*военный + 0.004*цель

15. Местная власть: Горсовет

0.036*депутат + 0.016*вопрос + 0.015*омск + 0.015*совет + 0.010*город + 0.010*городской + 0.010*горсовет + 0.009*заседание + 0.009*решение + 0.007*мэр + 0.007*комитет + 0.007*госдума + 0.007*принять + 0.007*закон + 0.006*должный + 0.006*директор + 0.006*общественный + 0.006*председатель + 0.005*муниципальный + 0.005*предложение

16. Деятельность мэрии: строительство и реконструкция

0.029*омск + 0.016*город + 0.013*департамент + 0.013*мэрия + 0.013*городской + 0.012*участок + 0.010*строительство + 0.010*администрация + 0.009*улица + 0.009*территория + 0.008*проект + 0.008*объект + 0.007*работа + 0.006*земельный + 0.006*реконструкция + 0.006*дерево + 0.006*земля + 0.005*мэр + 0.005*директор + 0.005*двораковский

17. Школьные и дошкольные учреждения

0.043*ребёнок + 0.029*школа + 0.024*детский + 0.017*омск + 0.012*сад + 0.011*образование + 0.010*родитель + 0.010*учреждение + 0.009*семья + 0.008*школьник + 0.007*социальный + 0.006*учитель + 0.006*область + 0.005*день + 0.005*учебный + 0.005*человек + 0.005*образовательный + 0.004*школьный + 0.004*педагог + 0.004*место

18. Продажа автомобилей

0.024*автомобиль + 0.019*• + 0.015*тело + 0.012*3812 + 0.009*центр + 0.009*:(+ 0.009*скидка + 0.008*реклама + 0.007*право + 0.006*дом.ru + 0.006*000 + 0.005*официальный + 0.005*дилер + 0.005*улица + 0.005*hyundai + 0.005*цена + 0.005*комплектация + 0.005*система + 0.005*клиника + 0.005*акция

19. ДТП

0.030*водитель + 0.026*дтп + 0.023*омск + 0.021*автомобиль + 0.013*улица + 0.012*происшествие + 0.012*результат + 0.011*место + 0.011*район + 0.010*произойти + 0.009*пассажир + 0.009*авария + 0.009*сбить + 0.009*мужчина + 0.009*область + 0.008*двигаться + 0.008*медицинский + 0.007*установить + 0.007*травма + 0.007*умвд

20. Высшее образование.

0.018*омск + 0.014*россия + 0.012*студент + 0.011*вуз + 0.009*образование + 0.008*университет + 0.007*работа + 0.007*государственный + 0.006*программа + 0.006*молодая + 0.005*ректор + 0.005*академия + 0.005*выпускник + 0.005*учебный + 0.005*высокий + 0.005*экзамен + 0.005*человек + 0.005*наука + 0.004*получить + 0.004*егэ

21. Недвижимость: строительство, продажа

0.033*метр + 0.024*строительство + 0.023*тысяча + 0.017*площадь + 0.016*рубль + 0.015*кв + 0.013*миллион + 0.011*омск + 0.010*дом + 0.010*здание + 0.009*компания + 0.008*построить + 0.008*объект + 0.008*участок + 0.008*комплекс + 0.007*ооо + 0.006*строительный + 0.005*микрорайон + 0.005*километр + 0.005*аукцион

22. Искусство.

0.013*человек + 0.009*книга + 0.007*жизнь + 0.006*слово + 0.006*женщина + 0.006*полежаев + 0.004*время + 0.004*леонид + 0.004*бывший + 0.004*григорьев + 0.004*история + 0.003*отношение + 0.003*автор + 0.003*имя + 0.003*библиотека + 0.003*случай + 0.003*считать + 0.003*закон + 0.003*мужчина + 0.003*дело

23.Ремонт и строительство городской инфраструктуры

0.044*дорога + 0.032*мост + 0.029*дорожный + 0.024*работа + 0.024*переход + 0.019*ремонт + 0.015*омск + 0.011*пешеходный + 0.011*ленинградский + 0.009*подземный + 0.009*улица + 0.009*движение + 0.008*строительство + 0.007*сентябрь + 0.007*покрытие + 0.007*срок + 0.006*автомобильный + 0.006*остановка + 0.006*огонёк + 0.006*часть

24. Жилищный вопрос, социальная сфера

0.049*дом + 0.046*квартира + 0.032*жильё + 0.010*омич + 0.010*жилищный + 0.009*семья + 0.009*инвалид + 0.008*гражданин + 0.008*омск + 0.007*житель + 0.007*человек + 0.007*программа + 0.007*фонд + 0.007*получить + 0.006*жилой + 0.006*недвижимость + 0.006*капремонт + 0.005*вопрос + 0.005*жилец + 0.005*жить

25. Домашние животные: собаки

0.014*человек + 0.007*собака + 0.007*друг + 0.004*слово + 0.004*животное + 0.004*видео + 0.004*сеть + 0.003*несколько + 0.003*животный + 0.003*фото + 0.003*день + 0.003*социальный + 0.003*фотография + 0.003*имя + 0.003*жить + 0.003*место + 0.003*жизнь + 0.003*приют + 0.003*знать + 0.002*оказаться

26. Городские мероприятия

0.032*омск + 0.022*город + 0.019*выставка + 0.016*омич + 0.012*день + 0.010*мероприятие + 0.009*площадка + 0.009*музей + 0.008*центр + 0.008*открытие + 0.008*парк + 0.007*гость + 0.007*смочь + 0.007*культура + 0.007*пройти + 0.006*проект + 0.006*праздник + 0.006*представить + 0.005*работа + 0.005*программа

27. Присоединение крыма

0.048*крым + 0.024*россия + 0.017*время + 0.010*республика + 0.010*навальный + 0.009*севастополь + 0.008*крымский + 0.007*пиво + 0.007*полуостров + 0.007*референдум + 0.006*состав + 0.005*сан + 0.005*москва + 0.005*час + 0.005*зимний + 0.004*инбеть + 0.004*присоединение + 0.004*симферополь + 0.004*житель + 0.004*новый

28. Концерты

0.022*концерт + 0.020*группа + 0.016*омск + 0.010*билет + 0.010*музыка + 0.008*музыкант + 0.007*музыкальный + 0.007*песня + 0.007*песнь + 0.007*выступление + 0.006*шоу + 0.006*рос-

сия + 0.006*выступить + 0.006*программа + 0.006*певица + 0.005*город + 0.005*известный + 0.005*артист + 0.005*сцена + 0.004*альбом

29. Хоккей: «Авангард»; «Омичка»[2]

0.022*авангард + 0.022*команда + 0.015*омск + 0.013*матч + 0.012*клуб + 0.008*тренер + 0.008*сезон + 0.007*игрок + 0.007*игра + 0.007*чемпионат + 0.006*россия + 0.006*омичка + 0.006*болельщик + 0.005*хоккеист + 0.005*главный + 0.005*хоккейный + 0.005*сборный + 0.005*победа + 0.004*время + 0.004*хороший

30. Хоккей: «Авангард»

0.024*минута + 0.020*матч + 0.020*счёт + 0.019*авангард + 0.019*шайба + 0.015*период + 0.015*игра + 0.013*омич + 0.012*ворот + 0.012*второй + 0.012*ястреб + 0.008*омск + 0.008*забросить + 0.008*гость + 0.008*хозяин + 0.008*первый + 0.008*нападать + 0.007*третий + 0.007*денис + 0.007*гол

31. Регулирование и надзор на предприятиях

0.027*омск + 0.013*предприятие + 0.012*завод + 0.010*проверка + 0.009*роспотребнадзор + 0.007*управление + 0.007*результат + 0.007*теплоход + 0.006*иртыш + 0.006*производство + 0.006*сыр + 0.006*нарушение + 0.006*человек + 0.006*речной + 0.005*безопасность + 0.005*оао + 0.005*лошадь + 0.005*вещество + 0.005*область + 0.005*специалист

32. Бюджет Омской области0.065*рубль + 0.030*миллион + 0.019*омск + 0.018*бюджет + 0.015*тысяча + 0.012*стоимость + 0.012*миллиард + 0.009*область + 0.008*цена + 0.008*сумма + 0.008*1 + 0.007*средство + 0.007*доход + 0.007*составить + 0.007*2014 + 0.006*проезд + 0.006*2013 + 0.006*составлять + 0.006*деньги + 0.006*зарплата

33. Объявления о поиске пропавших

0.020*полиция + 0.018*омск + 0.014*телефон + 0.014*реклама + 0.014*пропасть + 0.012*цвет + 0.011*чёрный + 0.009*информация + 0.008*поиск + 0.008*сантиметр + 0.008*искать + 0.007*примета + 0.007*уйти + 0.007*волос + 0.007*02 + 0.007*дом + 0.007*рост + 0.006*сообщить + 0.006*найти + 0.006*одетый

34. Театры

0.032*омск + 0.027*театр + 0.013*спектакль + 0.010*культура + 0.009*фестиваль + 0.008*россия + 0.007*артист + 0.006*актёр + 0.005*коллектив + 0.005*театральный + 0.005*зритель + 0.005*владимир + 0.005*зал + 0.005*сцена + 0.005*имя + 0.005*режиссёр + 0.005*директор + 0.004*искусство + 0.004*творческий + 0.004*сергей

35. Омские СМИ: телевидение и газеты.

0.022*омск + 0.018*канал + 0.013*телеканал + 0.013*газета + 0.011*пляж + 0.011*вода + 0.010*озеро + 0.010*иртыш + 0.009*издание + 0.007*отдых + 0.007*правда + 0.007*директор + 0.006*журналист + 0.006*лодка + 0.006*редактор + 0.006*дождь + 0.005*эфир + 0.005*место + 0.005*тв + 0.005*новый

36. Местная власть

[2]Популярный в Омске женский волейбольный клуб

0.018*омск + 0.010*область + 0.006*человек + 0.006*министр + 0.006*сми + 0.006*информация + 0.006*дело + 0.005*регион + 0.005*чиновник + 0.005*ситуация + 0.005*власть + 0.004*руководитель + 0.004*андрей + 0.004*сергей + 0.004*главный + 0.004*владимир + 0.004*начальник + 0.004*управление + 0.003*глава + 0.003*сайт

37. Экономическая ситуация в связи с событиями на Украине

0.028*россия + 0.014*доллар + 0.011*рынок + 0.011*украина + 0.009*газпром + 0.008*миллиард + 0.007*цена + 0.007*рост + 0.006*компания + 0.006*страна + 0.006*газа + 0.006*экономика + 0.005*евро + 0.005*газ + 0.004*экономический + 0.004*валюта + 0.004*беженец + 0.004*нефть + 0.004*сша + 0.004*поставка

38. Коммунальная сфера: отопление

0.019*дом + 0.011*работа + 0.009*котельная + 0.008*вода + 0.007*сезон + 0.007*человек + 0.007*посёлок + 0.007*житель + 0.006*омск + 0.006*ремонт + 0.006*объект + 0.006*тепло + 0.006*отопительный + 0.006*необходимый + 0.005*ситуация + 0.005*новый + 0.004*время + 0.004*степной + 0.004*безопасность + 0.004*отметить

39. Международные отношение между Россией, Украиной с США

0.044*россия + 0.021*украина + 0.014*президент + 0.013*страна + 0.013*путин + 0.009*сша + 0.007*украинский + 0.006*владимир + 0.006*государство + 0.005*заявить + 0.005*власть + 0.005*военный + 0.004*территория + 0.004*американский + 0.004*сторона + 0.004*глава + 0.004*киев + 0.003*против + 0.003*сила + 0.003*санкция

40. Информация о различных конкурсах. Авиакомпании.

0.035*конкурс + 0.025*омск + 0.022*аэропорт + 0.011*победитель + 0.010*хороший + 0.010*россия + 0.010*акция + 0.009*участник + 0.009*участие + 0.007*рейс + 0.006*получить + 0.006*приз + 0.006*проект + 0.006*москва + 0.006*самолёт + 0.005*номинация + 0.005*пройти + 0.005*место + 0.005*девушка + 0.005*авиакомпания

41. Арбитражные суды, «Мостовик»[3]

0.020*компания + 0.018*суд + 0.015*мостовик + 0.012*долг + 0.012*судебный + 0.011*омск + 0.011*ооо + 0.011*рубль + 0.010*миллион + 0.010*пристав + 0.008*иск + 0.007*предприятие + 0.007*задолженность + 0.006*нпо + 0.006*дело + 0.006*договор + 0.006*директор + 0.005*решение + 0.005*бывший + 0.005*бизнесмен

42. Торжества в честь победы в ВОВ

0.019*день + 0.019*омск + 0.014*победа + 0.012*акция + 0.012*ветеран + 0.011*война + 0.011*мероприятие + 0.011*май + 0.010*площадь + 0.010*праздник + 0.009*пройти + 0.008*омич + 0.008*отечественный + 0.008*великий + 0.007*памятник + 0.007*праздничный + 0.007*парад + 0.006*митинг + 0.006*память + 0.006*человек

43. Уголовные дела

0.021*уголовный + 0.020*дело + 0.018*омск + 0.013*мужчина + 0.012*россия + 0.012*полиция + 0.010*следственный + 0.010*преступление + 0.009*область + 0.009*статья + 0.008*возбудить +

[3]Крупнейшая в Омске строительная компания. В 2014 г. столкнулась с экономическими трудностями.

0.007*ук + 0.006*подозревать + 0.006*задержать + 0.006*час + 0.006*следователь + 0.006*сообщить + 0.006*время + 0.006*расследование + 0.006*отношение

44. Фильмы. Новый год.

0.027*фильм + 0.010*новый + 0.009*новогодний + 0.007*режиссёр + 0.007*картина + 0.006*актёр + 0.006*зритель + 0.006*роль + 0.005*главный + 0.005*герой + 0.005*кино + 0.005*мир + 0.005*лёд + 0.004*съёмка + 0.004*ёлка + 0.004*первый + 0.004*кинотеатр + 0.004*сериал + 0.004*мороз + 0.003*премьера

45. Олимпиада 2014.

0.020*олимпийский + 0.017*омск + 0.015*спорт + 0.015*россия + 0.012*сочи + 0.011*спортсмен + 0.011*олимпиада + 0.010*эстафета + 0.010*игра + 0.009*огонь + 0.009*марафон + 0.009*спортивный + 0.009*соревнование + 0.007*медаль + 0.006*участник + 0.006*турнир + 0.005*чемпион + 0.005*международный + 0.005*пройти + 0.004*мир

46. Сложно однозначно определить. Значительная часть статей посвящена таможенному контролю

0.027*россия + 0.017*область + 0.014*омск + 0.010*помощь + 0.009*вертолёт + 0.009*управление + 0.008*служба + 0.008*паспорт + 0.008*средство + 0.007*право + 0.006*начальник + 0.006*пункт + 0.006*дмитриев + 0.005*удостоверение + 0.005*сотрудник + 0.005*граница + 0.005*территория + 0.005*гражданин + 0.004*таможенный + 0.004*груз

47. Убийство Ивана Климова[4].

0.017*иван + 0.017*омск + 0.014*климов + 0.014*убийство + 0.011*лебедовый + 0.011*боксёр + 0.009*конфликт + 0.009*ян + 0.008*версия + 0.006*полиция + 0.006*дело + 0.006*россия + 0.006*ранение + 0.006*расследование + 0.005*преступление + 0.005*человек + 0.005*стрельба + 0.005*информация + 0.005*область + 0.005*ноябрь

48. Сводки нарушений правил ПДД

0.041*автомобиль + 0.032*водитель + 0.024*машина + 0.017*омск + 0.014*гибдд + 0.010*транспортный + 0.009*очевидец + 0.009*сотрудник + 0.009*видео + 0.008*улица + 0.008*средство + 0.007*движение + 0.007*нарушение + 0.007*дорожный + 0.006*иномарка + 0.006*правило + 0.006*административный + 0.006*фото + 0.006*дорога + 0.005*полицейский

49. Районы области

0.100*район + 0.042*область + 0.025*глава + 0.021*омск + 0.015*сельский + 0.011*поселение + 0.011*калачинский + 0.010*житель + 0.010*местный + 0.009*деревня + 0.008*село + 0.008*областной + 0.007*выбор + 0.007*районный + 0.007*муниципальный + 0.007*кормиловский + 0.007*сель + 0.006*черлакский + 0.005*цыганков + 0.005*тарский

[4]Известный омский боксёр. Был убит в возрасте 23-х лет.

Приложение В

Рейтинг популярности тем

Таблица В.1: Самые популярные темы, рассчитанные через среднюю вероятность

Порядок	Номер темы	Средняя вероятность
1	19	0.0478
2	43	0.0448
3	4	0.0389
4	1	0.0383
5	32	0.0378
6	39	0.0369
7	10	0.0347
8	29	0.0333
9	2	0.0319
10	7	0.0313
11	11	0.0306
12	16	0.0286
13	15	0.0255
14	0	0.0251
15	45	0.024
16	6	0.0237
17	36	0.0233
18	25	0.0228
19	12	0.0227
20	48	0.0215
21	41	0.021
22	5	0.0179
23	20	0.0179
24	3	0.0176

продолжение следует

(продолжение)		
Порядок	Номер темы	Средняя вероятность
25	17	0.0161
26	26	0.0157
27	34	0.0157
28	40	0.0152
29	21	0.0148
30	37	0.0143
31	31	0.0143
32	9	0.014
33	44	0.0129
34	38	0.0128
35	8	0.0111
36	49	0.011
37	22	0.011
38	13	0.0108
39	33	0.0107
40	28	0.0106
41	23	0.0106
42	30	0.0103
43	47	0.0097
44	24	0.0093
45	18	0.0092
46	42	0.0082
47	46	0.0072
48	35	0.0063
49	27	0.0062
50	14	0.0032

Приложение C

Комментарии

C.1. Количество комментариев

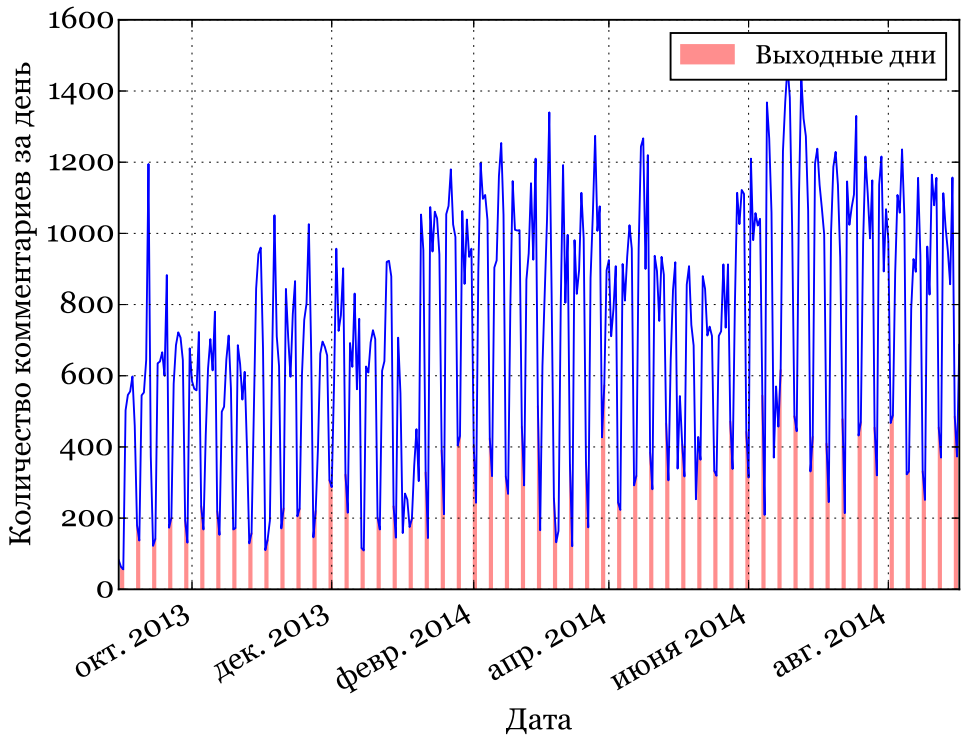

Рисунок C.1: Количество комментариев по дням

Анализ графика на рисунке C.1 позволяет сделать несколько выводов. Во-первых, видно, что в к статьям, которые вышли в выходные дни пользователи оставляют намного меньше комментариев, чем к тем, которые опубликованы в будни. Среднее количество комментариев к статьям в выходного дня составляет 293, в то время как к статьям, написанным в будние дни – 867.

C.2. Комментируемость тем

Таблица С.1: Самые комментируемые темы

Порядок	Номер темы	Процент комментируемости от наиболее комментируемой
1	39	100.0%
2	25	94.0%
3	13	86.8%
4	36	79.9%
5	2	76.2%
6	48	74.9%
7	22	70.3%
8	27	69.7%
9	47	67.3%
10	32	65.7%
11	14	65.1%
12	37	65.0%
13	11	61.8%
14	1	61.6%
15	23	61.0%
16	16	58.8%
17	21	58.7%
18	15	57.9%
19	41	57.6%
20	42	57.1%
21	35	56.0%
22	19	55.9%
23	9	54.6%
24	5	54.3%
25	24	54.1%
26	0	51.6%
27	17	51.0%
28	34	49.1%
29	43	48.2%
30	46	47.9%
31	40	46.4%
32	31	46.2%
33	7	45.9%
34	6	44.8%
35	10	44.7%

продолжение следует

(продолжение)		
Порядок	Номер темы	Процент комментируемости от наиболее комментируемой
36	30	43.3%
37	3	42.6%
38	29	41.7%
39	33	41.3%
40	12	40.2%
41	38	39.8%
42	44	39.6%
43	26	39.2%
44	45	38.7%
45	20	37.0%
46	28	37.0%
47	49	35.4%
48	4	34.6%
49	8	34.2%
50	18	20.7%

С.3. Тональность комментариев по темам