50.3.	Цели и основные задачи Проекта (указать, как в заявке)	Задачей исследования является создание методологии для решения вышеназванной фундаментальной проблемы на основе комплекса математических методов и компьютерных моделей и ресурсов, состоящего из следующих компонент: 1. Систематизированного описания типов систем в социогуманитарной области, динамика которых может описываться в терминах последовательностей событий и значений ключевых параметров (временных рядов). 2. Формализованного описания социогуманитарных систем и трендов их развития в терминах теории временных рядов, нечеткой логики и мягких вычислений. 3. Модели анализа временных рядов на предмет выявления положительных и отрицательных ассоциаций. 4. Системы извлечения сущностей, событий и фактов, характерных для социогуманитарной предметной области из текстов на русском языке (преимущественно). 5. Новых методов выявления тематики текстов (методы тематического моделирования), улучшающие ранее известные методы за счет использования разработанных в рамках проекта методов выделения сущностей. 6. Методов выделения эмотивно-тональных трендов развития заданной тематики в социальных сетях. 7. Методов выявления новых закономерностей путем соотнесения политических, социальных, экономических событий, фактов и трендов, извлеченных из текстовых массивов, с моментами изменений во временных рядах адекватных событиям параметров. Разработанная методология и комплекс средств должны быть апробированы на примерах из областей, изучаемых в социологии, политологии, культурологии, психологии, филологии и экономических науках в очень широком диапазоне временных интервалов - от нескольких часов до нескольких столетий на наборах данных, адекватных этим системам и решаемым задачам.
50.4.	Степень выполнения поставленных в Проекте задач	Поставленные задачи, в целом, выполнены. Некоторая корректировка задач обусловлена логикой исследований и получаемыми результатами.
50.5.	Важнейшие результаты, подученные в ходе реализации Проекта с указанием их междисциплинарности и значимости для развития междисциплинарных направлений; их соответствие задачам Темы исследований	Перечислим основные результаты по проекту, которые можно условно разбить на три направления: анализ социогуманитарных трендов с использованием Google Books Ngram (долговременные тренды - на временных интервалах 1-2 века), анализ информации из социальных сетей и новостных потоков (кратковременные тренды), анализ данных социологических опросов. Для изучения социальной динамики в СССР/России выделены следующие параметры: Доля сельского населения (The population of Russia for 100 years (1897-1997): statistical compilation / Goskomstat of Russia. - M., 1998; Statistical year-book of Russia, 2014: statistical compilation / Goskomstat of Russia. - M., 2014), Ожидаемая продолжительность жизни при рождении, Подушевой валовой национальный доход (в $ 1990 года) (A. Markevich, M. Harrison. Great War, Civil War, and Recovery: Russia’s National Income, 1913 to 1928 // Journal of Economic History. Vol. 71. №3. P. 672—703; N. Dzis-Voynarovskiy. The graph of the century: Russia's economy for 1913-2012 years. M: 2012; E.M. Andreev, L.E. Darsky, T.L. Harkova. Demographic history of Russia: 1927-1959. Moscow, 1998), Суммарный коэффициент рождаемости (Википедия, https://ru.wikipedia.org/wiki/Население_России). Перечень данных, доступных в базе данных МВФ по России, находится по адресу http://www.imf.org/en/data. Исследовались взаимосвязи трех групп параметров: социодемографические – валовой продукт на душу

населения, средняя продолжительность жизни, доли сельского/городского населения, коэффициент рождаемости; личные местоимения; ключевые концепты, характерные для русской культуры. Последние разбиты на несколько групп: A. Патриархальные ценности сельского населения (привязанность, община, уважать), B. Неконтролируемость ситуации, неагентивность (покорный, смиренный, фатальный), C. Ценности, предположительно, характерные для городского населения – строптивый, самостоятельный, жалость, равнодушие, служить, работать. Временные ряды частотности употребления слов берутся по корпусу Google Books Ngram (сокращенно GBN). Проведен анализ развития индивидуализма в российском обществе. Показано, что развитие индивидуализма коррелирует с увеличениям частоты использования местоимения ‘я’ по сравнению с местоимением ‘мы’. Установлено, что во многих аспектах при переходе к рыночной экономике наблюдается увеличением частоты использования слов с индивидуалистическим значением и уменьшение частоты слов, связанных с коллективистскими ценностями. Таким образом, социально-экономические изменения приводят к изменениям в структуре ценностей общества, изменениям психологии людей. В некоторых случаях слова, связанные с коллективистскими ценностями, использовались чаще, чем слова, связанные с индивидуалистическими ценностями. Это можно объяснить переходным состоянием российского общества. Более детально, обнаружено, что уменьшение частоты слов из групп A и B, а также слов ‘жалость’ и ‘служить’ коррелирует с уменьшением доли сельского населения. Для слов группы A, связанных с патриархальным укладом жизни, аналогичный результат имеет место для английского и китайского языков. Слова группы B, а также слова ‘жалость’ и ‘служить’, вообще говоря, не связаны с сельским образом жизни, а являются просто характерными для образа жизни и мышления населения России (согласно И. Левонтиной, А. Шмелеву, Анна Зализняк, 2005) Уменьшение их частоты указывает на одновременный распад и этих особенностей психологии россиян. Результаты частично подтверждают теорию Р.М.Гринфелд (P. M. Greenfield. Linking social change and developmental change: Shifting pathways of human development. Developmental Psychology, 45, 2009. Pp. 401–418), согласно которой ценности общества, поведение, психология адаптируются к экологическим (в широком смысле слова) изменениям. Однако обнаружены и принципиально новые моменты, не выявленные в аналогичных исследованиях для США и Китая. На некоторые ключевые концепты российской ментальности большее влияние оказали не экологические изменения, включая процессы урбанизации и повышения уровня жизни, а государственная идеология.

Изучалась динамика употребления ключевых концептов, таких, как ‘счастье’, ‘богатство’, ‘война’, ‘здоровье’, ‘еда’ ‘работа’, ‘развитие’. Исследование проводилось в сопоставительном аспекте для основных европейских языков (английский, русский, немецкий, французский, итальянский, испанский). Описаны тренды изменения отношения в обществе к указанным концептам, проведен сопоставительный анализ для России, США, Франции, Великобритании, Германии, Испании, Италии. В качестве примера полученных результатов приведем анализ концепта ‘счастье’. Впервые проведено исследование корреляции уровня счастья и таких факторов, как удача, уровень благосостояния и уровень достижений в обществе. Оказалось, что Россия заметно выделяется направленностью трендов, например, Россия оказалась единственной из рассмотренных стран, в которой уровень счастья рос в 19 веке. Интересным наблюдением является то, что во всех языках кривые для слов богатство и счастье антикоррелируют. Получены данные по динамике благополучия (удовлетворенности жизнью) для 6 стран (впервые на материале лингвистических данных

сверхбольшого корпуса). Оказалось, что ощущение благополучия существенно различается для разных языков, причем эти 6 языков можно разделить на две группы со схожей картиной: русский, французский, итальянский и английский, немецкий, испанский. В первой из групп наблюдается тенденция к повышению уровня удовлетворенности жизнью, во второй – к снижению. Испанский язык демонстрирует смешанную тенденцию: в 19 веке наблюдался рост удовлетворенности жизнью, как для языков первой группы, но в 20 веке имеет место общая тенденция к снижению, как для языков второй группы. Для всех анализируемых слов получены нетривиальные результаты. Так для слова war в английском, кроме ожидаемого увеличения частотности в период войн, наблюдается повышение его частоты с конца 20-го века, что заслуживает самого пристального внимания.

Для анализа текстов и извлечения информации из соцсетей и новостных потоков разработан ряд методов.

1. Метод RTE (Recognizing textual entailment – распознавание текстовых взаимосвязей) для коротких текстов на ограниченном подмножестве русского языка на основе созданного нами словаря глагольного управления Russian FrameNet (аналог англоязычного FrameNet). Метод применялся для установления причинно-следственных связей между событиями, извлеченными из новостных текстов на ограниченном подмножестве русского языка: события, связанные с безопасностью и массовыми волнениями.
2. Группа методов для автоматического выделения из текстов естественного языка признаков характеризующих эмоциональное состояние автора (эмотивность текста) в момент написания текста; определения тональности текста по отношению к определённым объектам; выделения вложенных тем на основе семантических, вероятностных, энтропийных характеристик коллекции текстов; кластеризации документов, использующий информацию полученную при построении вложенных тем.
3. Адаптирован и реализован семантический алгоритм Гинзбурга для обработки больших объёмов данных в среде BigData на платформе Hadoop; разработана методика ранжирования документов в тематической коллекции относительно заданных ключевых слов опирающаяся на контекстно-семантический граф темы; развит подход к представлению динамики развития темы для наглядной визуализации.
4. Предложена методика извлечения именованных сущностей из текстов русского языка в контексте построения сложных современных нейросетевых моделей глубокого обучения. Предложено улучшение метода морфологического анализа на основе нейронных сетей глубокого обучения. Данный метод включает в себя два уровня анализа входного текста: уровень отдельных предложений и уровень слова. Структура сети для обоих уровней схожая и включает в себя конволюционные CNN слои в сочетании с полносвязными слоями. Сравнение с другими морфологическими анализаторами проводилось на корпусе SynTagRus в оригинальном формате морфологических признаков и его модификаций в Universal Dependencies версий 1.3 и 1.4. Достигнута точность определения частей речи: 98,34%, 98.49%, 97,60% (соответственно для варианта разметки). Результаты немного выше, чем аналогичных полученных системой Syntaxnet Google. Полученные результаты демонстрируют большой потенциал сложных нейронных сетей глубокого обучения по сравнению с традиционными подходами, такими как SVM.
5. Разработана методика по выявлению информационных вбросов в сети Интернет.
6. Разработана нелинейная регрессионная модель появления событий и роста числа словоупотреблений в новостных потоках. Реализация модели (на языке java) позволяет в качестве функций, представляющих отдельные отрезки времени, задавать экспоненциальные функции, логистическую функцию, функцию

Гомперца, либо определять произвольную функцию.
На основе разработанной регрессионной модели проведены следующие эксперименты:
- использование модели для обнаружения событий;
- использование модели для определения различных типов слов-индикаторов событий и построения классификации событий.
Эксперименты для выявления событий проводились на новостном корпусе, содержащем более 187 тысяч словоформ в новостях, появившихся в период с декабря 2012 года по апрель 2016. Данный корпус создан коллективом авторов КФУ. Корпус был предварительно обработан следующим образом. Все сообщения, появившиеся в один и тот же день, были объединены в один псевдо-документ. Затем были произведены стандартные этапы предобработки текста: токенизация, приведение всех словоформ к нижнему регистру и подсчет числа каждой из словоформ во всех псевдо-документах. На основе этих данных был построен временной ряд. Для каждого слова выполнялось нормирование числа упоминаний так, что значения Zi изменяются от 0 до 1. Предлагаемая модель может применяться для классификации событий. Исходный код и набор данных выложены на сайте проекта.
На корпусе новостных текстов построены временные ряды для n-грамм (n = 2, 3). Обеспечен доступ по ссылке: https://drive.google.com/open?id=1hYxvj-wqWsZL4UJ0Q6isaSARSXq4pZX9. На выходе получен набор данных, содержащий 2 867 852 записей о 246677 биграммах и 1304681 о 229812 триграммах (число упоминаний биграмм и триграмм не менее 5). Упоминаемость каждой отдельно 2-граммы (3-граммы) в тексте меньше, чем упоминаемость ее составляющих, поэтому для эффективного применения регрессионной модели требуется выполнять агрегирование не по дням (как это было для отдельных слов), а по неделям или месяцам. В этом случае временной ряд для отдельной 2-граммы будет иметь меньший размер (в текущей версии датасета представлены 72 месяца). Альтернативным подходом является объединение нескольких биграмм в одну группу. После агрегации возможно применение регрессионной модели для выделения событий, характеризующихся именно выбранной 2-граммой или группой (что может обеспечить большую точность). Создан набор данных, содержащий вектора слов, построенные по модели word2vec. Вектора слов строились при помощи программного обеспечения FastText (https://fasttext.cc), которое позволяет строить вектора с учетом n-грамм символов, входящих в слово. Этот аспект важен при работе с русским языком, поскольку позволяет частично учесть в модели флективность языка. При построении учитывались n-граммы из 2, 3, 4 и 5 символов. Общее число векторов составило 1 470 087. Размер вектора для одного слова равен 100. Размер модели на диске: 1,3 Гб. Обеспечен доступ по ссылке: https://drive.google.com/open?id=1ae9DLQEDYgbsQu4gweG11w93Ze-j9x-T.
В ходе работ проводились эксперименты по апробации разработанных методов применительно к задаче выделения эмотивно-тональных трендов. Для экспериментов использовался корпус русскоязычных новостных статей опубликованных в различных интернет изданиях за период с 01.01.2013 по 14.03.2015 года, общим количеством 2,9 млн. текстов. Анализировались целевые выборки текстов, содержащие мнение по таким объектам интереса как «Ангеля Меркель», «Франсуа Олланд», «Танк армата», «Арктика». Результаты анализа продемонстрировали эффективность разработанных алгоритмов для задач мониторирования социогуманитарных трендов.
Разработан метод функциональной разметки для разметки речевого материала в корпусах, включающих

		мультимодальную информацию (видео, аудио). Проведена разметка корпуса REC (Russian Emotional Corpus) в объеме более 6000 аннотаций. Предлагаемый подход позволяет объединить мультимодальную информацию, содержащуюся в коммуникации, а также речевую информацию, традиционно являющуюся объектом лингвистических исследований. Разметка корпуса позволяет применять компьютерные модели и математические методы к анализу материала мультимодальной коммуникации. Результаты изложены в монографии. Проведены масштабные социологические опросы, на основе которых выявлены основные факторы, определяющие здоровье российских студентов в совокупности ее составляющих, тренды изменения показателей социального и физического самочувствия у различных категорий студентов, характеристики профиля психосоматического здоровья студентов, показатели выраженности НПН (нервно-психическое напряжения), зависимость ощущения счастья и психологической устойчивости от физиологического состояния. Таким образом, в рамках работ по гранту на основе синтеза вероятностных, нейросетевых подходов, других интеллектуальных средств, а также методов корпусной и компьютерной лингвистики, кластеризации документов был разработан базовый комплекс методов и программ для анализа социогуманитарных трендов. Полученные результаты покрывают все заявленные темы исследования и очевидным образом носят междисциплинарный характер на стыке социологии, психологии, лингвистики и основаны на применении компьютерных технологий. Полученные результаты свидетельствуют о плодотворности предлагаемого подхода и дают импульс развития нового междисциплинарного направления “вычислительная социология”.
50.6.	Сопоставление результатов, полученных в ходе выполнения Проекта, с мировым уровнем (сравнение результатов, полученных в ходе выполнения проекта, с результатами российских и зарубежных коллег-привести ссылки на их работы, с указанием их новизны)	В направлении исследований – анализ социогуманитарных трендов по общей постановке задач и полученным результатам наиболее близки работы для Китая (R. Zeng, P.M. Greenfield. Cultural evolution over the last 40 years in China: Using the Google Ngram Viewer to study implications of social and political change for cultural values. International Journal of Psychology, Vol. 50, No. 1, 2015. pp. 47–55) и для США (I. Grossmann. Social Structure, Infectious Diseases, Disasters, Secularism, and Cultural Change in America. Psychological Science, February, 2015; P. M. Greenfield. The Changing Psychology of Culture From 1800 Through 2000. Psychological Science, 2013). Общие корреляции социодемографических параметров и ценностей общества, выражаемых во взятых нами ключевых словах лексикона, в значительной степени похожи на аналогичные результаты для США и Китая, приведенные в вышеуказанных работах. Представление о счастье и динамика употребления соответствующих слов изучалась в ряде работ (Acerbi et al. The Expression of Emotions in 20th Century Books. PLoS ONE, 8(3): e59030. 2013, Oishi et al. Concepts of Happiness Across Time and Cultures. Personality and Social Psychology Bulletin, vol. 39, pp. 559-577, 2013 и др.). Методы исследований в этих работах аналогичны, используемым нами, хотя эти работы носят не кросскультурный характер, а посвящены только английскому языку. В частности, для американского и английского общества выявлены “счастливые” и “несчастливые” периоды. Полученные результаты в этих работах и в нашей работе сопоставимы, но различаются в деталях. С созданием корпуса Google Books Ngram в 2010 г. на основополагающую работу Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books

Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*. Quantitative Analysis of Culture Using Millions of Digitized Books. Science (Published online ahead of print: 12/16/2010) в мировой литературе имеется уже около 1,5 тыс. ссылок (по https://scholar.google.ru/), что указывает на важность данного компьютерного ресурса и актуальность исследований на его основе. Из самых последних работ можно указать: Roberta Amato, Lucas Lacasa, Albert Díaz-Guilera, Andrea Baronchelli. The dynamics of norm change in the cultural evolution of language. Proceedings of the National Academy of Sciences. Aug. 2018, 201721059; DOI:10.1073/pnas.1721059115; Roberta Amato, Lucas Lacasa, Albert Díaz-Guilera, Andrea Baronchelli. The dynamics of norm change in the cultural evolution of language. Proceedings of the National Academy of Sciences. Aug. 2018, 201721059; DOI:10.1073/pnas.1721059115; AchimEdelmann, John W. Mohr. Formal studies of culture: Issues, challenges, and current trends. Poetics. Vol. 68, 2018, Pages 1-9.

В России лишь один научный коллектив работает в подобном русле исследований – на филологическом факультете Высшей школы экономики под руководством А. Бонч–Осмоловской. Однако вместо GBN они используют Национальный корпус русского языка. Ими реализован совместно с «Рамблер-Афиша» проект «Имена времени», в котором анализируются частоты употребления слов, характеризующие различные периоды истории России. Изучалась также динамика употребления следующих “ключевых” для России слов: дураки, дороги, воровство, коррупция, мужик, мужчина, парень, женщина, баба, девушка (http://www.hse.ru/news/science/124378966.html). А. Бонч–Осмоловской подготовлен обзор “Предсказания, большие данные и новые измерители: о возможности технологий компьютерной лингвистики в теоретических лингвистических исследованиях” // Вопросы языкознания. 2016, №2, с. 100-120, затрагивающий применение подобных технологий исследования в лингвистике.

Влияние здоровья населения и роли спортивного участия и спортивной инфраструктуры в ее поддержании являются объектом исследования во многих странах мира (Bohm F. Sport participation in Europe: The extent of contextual effects. Tilburg, 2013. URL: http://arno.uvt.nl/show.cgi?fid=130585; Wicker P., Hallmann K., and Breuer C. Analyzing the impact of sport infrastructure on sport participation using geo-coded data: Evidence from multi-level models. Sport Management Review, Volume 16, Issue 1, February 2013, Pages 54-67). В этих работах авторы высказывают соображения о необходимости количественных исследований, близких к нашим.

В направлении исследований – методы анализа текстов – при сравнении с зарубежными аналогами стоит отметить тот факт, что подавляющее большинство академических работ посвящено методам машинного обучения. В нашем проекте предлагается совмещать методы машинного обучения с учителем (для этого формируется корпус с разметкой событий) с методами, основанными на правилах.

Задачам анализа мнений в зарубежных работах уделяется большее внимание. В силу строгого порядка слов английский язык лучше поддается средствам анализа. Большая часть работ проводится на корпусах текстов на английском языке. Одно из наиболее известных соревнований по анализу мнений является SemEval (https://en.wikipedia.org/wiki/SemEval) с большим количеством пополняемых корпусов.

Все большую популярность набирают работы, демонстрирующие различные архитектуры нейронных сетей, так называемые сети глубокого обучения (Deep Learning). Среди которых выделяются подходы обучения без учителя или с частичным обучением (German Rigau and etc., V3: Unsupervised Aspect Based Sentiment Analysis for SemEval-2015 Task 12, Proceedings of the 9th International Workshop on Semantic Evaluation

(SemEval 2015), pages 714–718, Denver, Colorado, 2015). Авторы в работе строят граф предложения на основе пар, которые выделяются с помощью меры близости векторов слов, полученных с помощью word2vec и частотность совместного употребления слов. В работе (Aliaksei Severyn and Alessandro Moschitti, UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 464–469, Denver, Colorado, 2015) применяются глубокие конволюционные нейронные сети. Наибольшую точность для английского языка при этом демонстрируют комплексные подходы с использованием нейросетевых методов, более 80%.
Особенностью предлагаемого нами алгоритма определения тональности является использование обезличивания текстов на этапе предобработки. Таким образом, мы частично избавляемся от зависимости работающей модели от предметной области текстов обучающего корпуса. Для точного выделения признаков текста используется разработанный модуль морфологического разбора слов, который снимает неоднозначность разбора полученного с применением Mystem. Разработанный алгоритм комплексный и объединяет в себе как методы машинного обучения, так и использования тональных словарей. Апробация алгоритма на размеченном объектно-ориентированном корпусе отзывов на русском языке: SentiRuEval, Proceedings of International Conference Dialog продемонстрировала точность правильного определения тональности 74%.
В проекте развит подход, позволяющий определить эмоциональные тексты, которые отражают эмоционально-возбужденное состояние автора. Ретроспективно задача определения состояния автора текста активно изучалась специалистами в области психолингвистики. В статье (Dmitry N. Chernov, Yuri Y. Ignatov. Expression of psychological features in speech quantity indicators. Journal of Psycholinguistics, 1 (15) Moscow, 2012) продемонстрирована корреляция между количественными характеристиками текста и эмоциональным состоянием его автора. Полученные количественные характеристики включают в себя ряд морфологических свойств слов и синтаксические особенности предложений.
При разработке методики использовался опыт как международных, так и российских исследователей психолингвистики, которыми были детально исследованы вопросы формирования языка с психологическими особенностями автора, таких как Чарльз Осгуд, Н. Хомский, Стивен Пинкер (Doris Aaronson, Robert W. Rieber. Psycholinguistic Research: Implications and Applications. Psychology Press, November 20, 2013).
На его основе в нашем проекте предложен контекстно-независимый алгоритм с использованием психолингвистических маркеров для определения эмоциональной реакции общества на события, обсуждаемые в большом наборе текстов, а так же алгоритм снятия морфологической неоднозначности на основе Mystem.
Таким образом, в рамках данного проекта разработан алгоритм, позволяющий комплексно оценить высказываемое мнение по отношению к объекту интереса в потоках информации социальных сетей и других интернет-ресурсов, включая метод определения тональности текста и метод оценки эмотивности текста. Что касается работ российских исследователей, развивающих подходы близкие к предлагаемому в данной работе, но они направлены в большей степени на анализ психотипа авторов текстов (Литвинова Т.А. Идентификация и диагностирование личности автора письменного текста / Т. А. Литвинова, О.А. Литвинова. – Воронеж: Изд-во Воронеж. гос. пед. ун-та, 2015. – 332 с.), а не на анализ их эмоционального

состояния в момент написания. Сказанное определяет новизну предлагаемого в этой работе подхода.
На сегодняшний день существует ряд систем, частично позволяющих решить задачу поиска тематически схожих документов. В частности, в поисковых системах, основанных на Apache Lucene, поиск подобных текстов производится, используя заранее определенный документ, и реализуется по методу «мешка слов». Несмотря на такие достоинства данного метода, как производительность и универсальность, его использование для анализа тем «эволюционирующих» с изменением с течением времени состава ключевых слов вызывает трудности.
Другой подход заключается в представлении документов в виде вектора заданной размерности и использовании различных метрик близости двух векторов. В рамках этого подхода используется набор методов, как статистических: LDA, PLSA (Blei, David M и Ng, Andrew Y и Jordan, Michael I, «Latent dirichlet allocation», the Journal of machine Learning research №3, стр. 993 – 1022, 2003) так и основанных на нейронных сетях – Doc2vec (Le, Quoc и Mikolov, Tomas, «Distributed Representations of Sentences and Documents», Proceedings of The 31st International Conference on Machine Learning, стр. 1188 – 1196, 2014). Российским разработчиками предложен подход для тематического моделирования на основе аддитивной регуляризации (Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M., Yanina A. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications, October 19, 2015 - pp. 29-37). Преимущества данного подхода по сравнению с другими моделями LSA:
* многие байесовские тематические модели (или заложенные в них идеи) удаётся переформулировать через регуляризаторы;
* в BigARTM регуляризаторы не обязаны иметь вероятностный смысл;
* суммируя регуляризаторы, взятые из разных моделей, можно строить многоцелевые комбинированные модели;
* BigARTM проще, чем байесовский подход. Тематические модели в АРТМ легче понимать, легче выводить и легче комбинировать;
* снижается порог вхождения в область тематического моделирования для исследователей из смежных областей.
Данный подход хорошо работает для поиска высокоуровневых (общих) тем (политика, информационные технологии, медицина и т.п.), но плохо работает с узкими темами (поиск информации по конкретному объекту внутри большой предметной области, например, Меркель в политике), для которых очень важны конкретные ключевые слова. Недостатками такого подхода также являются: требование к наличию большого корпуса для обучения модели, невысокая точность и сложность определения необходимого уровня близости.
Наш алгоритм основан на выделении набора ключевых слов и словосочетаний из представленной пользователем подборки текстов по теме и дальнейшем поиске на основе выделенных слов с ранжирование результатов. Отличие нашего метода состоит в способе выделения ключевых слов и словосочетаний (комбинация 2-х или 3-х слов в рамках одного предложении, без учёта порядка), использования комбинации вероятностно-энтропийных характеристик текстов, семантического алгоритма Гинзбурга и дополнительных источников информации, таких как Национальный Корпус Русского языка.

В настоящее время активно развиваются методы выявления ложных новостей (в т.ч. слухов, информационных вбросов и т.д.) на базе алгоритмов с учителем. При этом строится бинарный классификатор для 2 классов: ложная новость или нет. При этом задача определения класса ложных новостей решается как типичная задача классификации текстов с поиском признаков для векторизации текстов и его фрагментов, а также поиска наиболее эффективного алгоритма классификации, оцениваемого по критериям f1-score, precision, recall. Так, в работах (Yang, Fan, et al. "Automatic detection of rumor on Sina Weibo." Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. ACM, 2012; Zhang, Q., Zhang, S., Dong, J., Xiong, J., & Cheng, X. Automatic detection of rumor on social network. In Natural Language Processing and Chinese Computing, pp. 113-122. Springer, Cham, 2015; E. Seo, P. Mohapatra, and T. Abdelzaher. Identifying rumors and their sources in social networks, Proc. SPIE 8389, Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR III – 2012. - C. 83891; Hamidian, S., & Diab, M. T. Rumor detection and classification for twitter data. In Proceedings of the Fifth International Conference on Social Media Technologies, Communication, and Informatics (SOTICS), pp. 71-77, 2015) в качестве признаков были предложены различные наборы характеристик текста (n-грамы, сентимент анализ, частеречный анализ и т.д.), автора текста (пол автора, активность автора, число публикаций и т.д.), источника данных (перепост, использование хештега, верификация автора и т.д.) и др., а в основными классификаторами стали Машина Опорных Векторов (SVM), логистическая регрессия и деревья решений. В данном проекте предложен и реализован метод без учителя на основе автоматического поиска похожих документов на базе их векторной модели с весами рассчитанными с помощью TF-IDF (реализовано в функции MoreLikeThis) с нахождением пороговых значений для поиска дубликатов, а также их визуализации. MoreLikeThis - это хорошо-зарекомендованный алгоритм из поискового движка Apache Lucene (Gospodnetic, O. and Hatcher, E. Lucene in Action. Manning Publications, Greenwich, 2005). Так в работах (Schwarzer, M., Schubotz, M., Meuschke, N., Breitinger, C., Markl, V., Gipp, B. Evaluating link-based recommendations for Wikipedia. In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, pp. 191-200. ACM, 2016; Schuemie, M. J., and Kors, J. A. Jane: suggesting journals, finding experts. Bioinformatics, 24(5), 727-728, 2008; Jakubina, Laurent, and Phillippe Langlais. WMT 2016: a Bilingual Document Alignment Platform Based on Lucene. Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. Vol. 2. 2016) на задаче оценки близости ссылок и содержания веб-страниц алгоритм MLT в случаях небольших статей не уступает алгоритмам на базе прямого анализа ссылок с расстановкой приоритетов об их удаленности по тексту. Кроме этого, в исследованиях по выявлению информационных вбросов (или слухов) зачастую рассматривается лишь один источник данных, например Twitter (Takahashi, Tetsuro, and Nobuyuki Igata. "Rumor detection on twitter", Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on. IEEE, 2012). Так в работе (Jakubina, Laurent, and Phillippe Langlais. WMT 2016: a Bilingual Document Alignment Platform Based on Lucene. Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. Vol. 2. 2016) рассматривается обнаружение слухов и их источников в социальной сети Twitter, на основе похожей идеи о том, что слухи генерируются только из небольшого числа источников, в то время как достоверная информация представлена большим числом независимых источников одновременно. Разработанный авторами метод продемонстрировал хороший потенциал, но ограничен в использовании только одной

		социальной сетью “Twitter” и неудобен постоянным отслеживанием всех выбранных источников. Наш метод позволяет выявлять информационные вбросы в различных источниках, не ограничиваясь одной социальной сетью, причём, не отбирая конкретных ресурсов для наблюдения. В России систематическое создание цифровых лингвистических ресурсов, аналогичных Russian FrameNet, ведется в Институте русского языка им. В. В. Виноградова РАН и Национальном исследовательском университете Высшая школа экономики. Одной из значительных работ в этом направлении стала статья: Ляшевская О. Н., Кашкин Е. В. Автоматическая классификация русских глаголов с использованием информации о морфо-синтаксическом оформлении и семантических ролях участников фреймов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 14 (21). — М.: Изд-во РГГУ, 2015. — Т. 1., С. 427-440. При сравнении с подходом в нашем проекте, следует отметить, что предлагаемый нами подход на основе Google Books в первую очередь ориентирован на применение FrameNet для прикладных задач, а не на получение новых знаний о русском языке (что является предметом исследования в упомянутой публикации). Извлечению событий из текста в российской компьютерной лингвистике уделяется крайне мало внимания. Однако стоит заметить, что в последние годы наметились положительные изменения. Примером является работа Danilova, V. and Popova, S., Extraction of Events from the Unstructured Text for the Tasks of Internet Sociology. (June 17, 2015). Available at SSRN: https://ssrn.com/abstract=2625905 or http://dx.doi.org/10.2139/ssrn.2625905. Созданный в рамках проекта корпус REC является одним из крупнейших в мире корпусов реального мультимодального взаимодействия. Функциональная разметка является новаторской для корпусов подобного вида. Демонстрация генеративной грамматики на основе данных корпуса является практическим подтверждением идей ряда иностранных учёных о возможности структурирования и генеративного описания коммуникативного поведения. Предлагаемая компьютерная модель использует эксплицитные лингвистические данные в области грамматики и семантики, записанные в виде XML-грамматики. В перспективе данная модель может дополняться статистическими методами для повышения её точности и полноты. Вместе с тем, иностранные научные группы и IT-компании исходно используют статистические методы, что ограничивает возможность уровневого представления лингвистических данных, гибкость в использовании лингвистического парсера. Все результаты проекта определяют мировой уровень в соответствующих областях.
50.7.	Методы и подходы, использованные в ходе реализации Проекта (описать, уделив особое внимание степени оригинальности и новизны, при необходимости-сравнить с работами зарубежных и российских коллег)	Для исследований социогуманитарных трендов ключевой идеей является использование корпуса Google Books Ngram. Корпус содержит для английского языка более 500 миллиардов слов, для русского – 67 миллиардов (в 200 раз больше, чем в НКРЯ). Он охватывает период с начала 16 века до 2008 г., но считается, что статистически надежные данные только с 1800 г. Корпус включает также данные по 9 языкам, в том числе, по основным европейским языкам. В базовой публикации: Michel J. et al. Quantitative Analysis of Culture Using Millions of Digitized Books. Science. 14 January 2011. Vol. 331. № 6014. pp. 176-182 сформулирован основной тезис: “the most robust historical trends are associated with frequent n-grams”, который лежит в основе наших исследований. Важным преимуществом этого ресурса является графическое

представление частот n-грамм (https://books.google.com/ngrams/) и доступность исходных данных, что позволяет работать с временными рядами частот словоупотреблений. Выбор слов в наших исследованиях обусловлен, с одной стороны, отражает наиболее универсальные концепты человеческой цивилизации, с другой стороны, определяется необходимостью сопоставления с аналогичными зарубежными работами, использующими определенные слова других языков. Подобные по методологии социолингвистические исследования проводятся только в последние годы после появления GBN в 2010 г., и их имеет смысл проводить только на основе GBN, т.к. других сравнимых по величине корпусов с диахронической разметкой данных просто нет (так, корпус современного Американского языка COHA имеет объем в 100 раз меньший). Нами выполнена предобработка корпуса GBN (удаление несловарных токенов), что обеспечивает повышение точности результатов. Все расчеты проводились в с помощью пакета MathLab.
Для работы с данными GBN требуется механизм численной оценки степени корреляций временных рядов. Обычно используемые в литературе коэффициенты корреляции Пирсона и Спирмена не вполне подходят для временных рядов. Трудностью является отсутствие универсальной меры схожести временных рядов. В частности, в работе (Koplenig A. Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions. Digital Scholarship in the Humanities. Oxford University Press. 2015) отмечено, что в общем случае не удается ограничиться чисто формальными методами, обойтись без визуальной интроспекции временных рядов. В этом направлении предложены новые методы поиска таких корреляций, основанные на анализе паттернов (локальных форм или образов) кривых изменения частот слов. Значимыми паттернами временных рядов частот слов являются следующие: изменение динамики кривой с возрастающей на убывающую и наоборот, рост кривой, падение кривой, максимум кривой, паттерн «холм», минимум кривой, паттерн «яма», интервал «больших» значений, интервал «малых» значений, интервал «колеблющихся» значений, локальные положительные ассоциации графиков частот разных слов, локальные отрицательные ассоциации графиков частот разных слов.
Разработаны меры ассоциации на множестве подинтервалов интервала [0.1], используемые при построении моделей анализа данных и принятия решений с нечеткими интервальными значениями истинности. Два временных ряда имеют локальную положительную ассоциацию, если существует временной интервал, на котором локальные углы наклона соответствующих кривых имеют одинаковый знак (синхронно возрастают или синхронно убывают). Они имеют локальную отрицательную ассоциацию, если существует временной интервал, на котором локальные углы наклона соответствующих кривых имеют противоположный знак (одна кривая возрастает, когда другая убывает). Предложенный метод назван “преобразование скользящих аппроксимаций”. Для его сопоставления с другими мерами использован стандартный набор временных рядов (E. Keogh, X. Xi, L. Wei, and C. A. Ratanamahatana. (2006) The UCR time series classification/clustering homepage. Доступно по ссылке: http://www. cs. ucr. edu/? eamonn/time series data. 17.05.2017), применяемый в мировой практике для сопоставления мер близости временных рядов. Показано, что на ряде наборах данных из стандартного тестового множества предложенный метод дает лучшие результаты, чем все ранее рассматривавшиеся.
Более подробно опишем методы, развитые для извлечения и обработки информации из текстов.
Метод RTE для коротких текстов на ограниченном подмножестве русского языка на основе словаря Russian FrameNet. Метод применяется для установления причинно-следственных связей между событиями,

извлеченными из новостных текстов на ограниченном подмножестве русского языка.
Метод состоит из следующих шагов:
1. Построение словаря индикаторов причинно-следственных связей (примеры индикаторов: “который повлек”, “что вызвало”, “стало причиной” и т.п.).
2. Построение элементарных грамматических шаблонов (на основе Russian FrameNet), для извлечения пар триггеров событий, состоящих в связи “причина-следствие” (например, “нападение (повлекло) массовые протесты”, “бомбардировка - гибель” и т.п.).
3. Построение направленного графа возможных зависимостей между всеми парами типов событий-следствий и событий-причин.
4. Поиск примеров связей между событиями, извлеченными из новостей на основе построенного графа возможных зависимостей.
Разработан комплекс математических методов для задачи выявления эмотивно-тональных трендов развития социальных процессов на основе текстовой информации из различных интернет-источников, включающий:
• математические методы проведения морфологического анализа текстов естественного языка (русского) для анализа динамики развития социальных процессов в интернет-ресурсах.
• математические методы определения тональности текстов и выделения эмотивных маркеров текстовых сообщений в потоках информации в социальных сетях и других интернет-ресурсах.
• математические методы для определения характерных тем социальных процессов в потоках неструктурированной или слабоструктурированной социальной информации из интернет-ресурсов, а также для визуализации результатов анализа выделенных тем в виде графа взаимосвязей значимых признаков.
Таким образом, представленный набор взаимоувязанных математических методов, позволяет разбирать текстовую информацию, выделять и анализировать заданные темы, для выделения трендов развития социальных процессов с характеристиками их тональности и эмотивности. Отличительными особенностями комплекса методов являются:
• удобство описания временного изменения тематики событий,
• наличие средств оценки эмоциональной насыщенности документов,
• оригинальная визуализация динамики изменения тем.
Кратко остановимся на некоторых наиболее важных для данного проекта алгоритмах.
Общий алгоритм определения тональности сводится к следующему:
1) Предобработка текста содержащего объект мониторинга, включая:
a. Нормализацию с использованием Mystem
b. Выделение предложений содержащих объект
c. Разбиение предложения на n-граммы
2) Классификацию предложения обученной модели (SVM с линейным ядром)
3) Интегральная оценка полученных результатов
Проведённые исследования основывались на корпусе сообщений из различных типов источников (социальных сет, новости, блоги, отзывы, микроблоги) размером более 500 тысяч текстов. В качестве базы знаний для обучения системы определения тональности использовался корпус размеченных сообщений размером более 2000, дополненный размеченным объектно-ориентированным корпусом текстов:

SentiRuEval, Proceedings of International Conference Dialog.
Методика определения эмотивности текста на основе выделения в тексте психолингвистических маркеров. На основании результатов исследования применения психолингвистических маркеров (Гудовских Д.В. и др., «Анализ эмотивности текстов на основе психолингвистических маркеров с определением морфологических свойств», Вестник Воронежского Государственного Университета 2015. № 3, С. 117-123), которые показали, что некоторые коррелирующие маркеры позволяют отделить среди набора текстов объективные тексты с фактами и новостями от субъективных текстов, выражающих мнение автора. Наиболее сильно отражают эмоциональную возбужденность и соответственно используются в данной работе следующие маркеры: Коэффициент Трейгера (КТ), Коэффициент определенности действия (КОД), Коэффициент агрессивности (КА). Выбранные маркеры имеют общие черты, описанные в психолингвистической диагностике: высокие значения демонстрируют наличие эмоционального волнения. Это характерно для лиц, склонных к немедленным действиям; низкие значения указывают на такие личностные характеристики, как неуверенность, тревожность.
Для использования представленных маркеров к анализу текстов различной длины вводится комплексный показатель, на основе основной группы маркеров (КТ, КОД, КА). Показатель основан на нормированных значений маркеров с учетом задаваемых коэффициентов значимости. При расчете суммарного ранга, коэффициенту Трэйгера и агрессивности назначаются веса равные 2. Далее значение суммарного ранга сообщения используется для оценки его эмотивности документа. В результате экспериментов были выделены границы значений эмотивности: крайне эмоциональные сообщения – значение суммарного ранга выше 5; средне эмотивные сообщения - от 3,5 до 5; без эмотивные тексты – менее 3,5. Большинство текстов носят рекламный характер, так же это ленты новостей, дайджест и информационный мусор, который не содержит мнения.
Алгоритм формирования ключевых словосочетаний состоит из нескольких шагов:
1. Выбрать N наиболее частотных слов в эталонной коллекции в кандидаты на ключевые слова (N=1000).
2. Рассчитать для кандидатов все описанные ранее индикаторы, за исключением индикатора связанности по Гинзбургу (приложение 3, рис. 2.)
3. Нормализовать полученные значения индикаторов. Получается ранг отражающий принадлежность данного слова к теме. Слова с наивысшим значением ранга являются ключевыми словами темы.
4. На основе выделенных ключевых слов формируются биграммы и триграммы без учёта последовательности и их положения в предложении.
5. Рассчитывается ранг словосочетаний с использованием ранее описанных индикаторов, включая индикатор связанности по Гинзбургу. Биграммы и триграммы с наивысшим значением ранга являются ключевыми словосочетаниями темы.
Результатом работы данного алгоритма является взвешенный относительно темы, заданной эталонной коллекцией, набор ключевых слов и словосочетаний.
Поиск тематически схожих документов. После того, как получена эталонная коллекция и выделены ключевые слова и словосочетания темы, можно переходить к этапу поиска тематически-схожих документов в хранилище. Он основан на описанных ранее алгоритмах и трёх источниках информации. Первый – это заданная пользователем эталонная коллекция документов, подсказывающая, что необходимо искать, второй

– это Национальный Корпус Русского Языка, использующийся для вычисления частоты общеупотребительных слов, и третий – набор документов, полученный в результате поиска по основным ключевым словам. Используя эти входные данные, модуль формирует ключевые слова и словосочетания для ранжирования документов, на релевантность заданной эталоном теме. При этом рассчитывается минимальная граница релевантности теме документов эталонной коллекции. Анализируя ключевые слова и словосочетания вместе с результатами поиска по основным ключевым словам, модуль формирует «минус» слова темы. Минус слова темы могут встречаться с основными ключевыми словами только в других предметных областях. Например для темы «Автомобили Форд» система выдаст следующие минус минус слова: Марк, Том, Харрисон. Марк Форд – поэт, Том Форд – дизайнер и кинорежиссёр, Харрисон Форд – знаменитый киноактёр. Полученные из хранилища документы взвешиваются и фильтруются на основе ключевых слов, словосочетаний и минус слов темы.

Методы построения контекстно-семантического графа. Для наглядной визуализации вложенных тем в большом количестве документов используется контекстно-семантический граф. Узлами данного графа являются ключевые слова, а ребрами ключевые биграммы, полученные в ходе анализа результатов поиска, представленными ранее методами. Размер узлов, расстояние между ними и толщина связывающих линий характеризуют связанность подтем в коллекции документов. Коллекция документов обрабатывается аналитическим модулем, с получением списка ключевых слов (узлов) и биграмм с рангами для построения графа. Из ключевых биграмм строится матрица смежности, описывающая связи будущего графа. К этой матрице применяется метод Affinity Propagation кластеризации на основе близости узлов через соседей, используя библиотеку scikit-learn (Pedregosa, F. and Varoquaux и др., «Scikit-learn: Machine Learning in Python», Journal of Machine Learning Research №12, стр. 2825 – 2830, 201), результатом работы алгоритма является набор вложенных тем, представленных кластерами из ключевых слов. Для визуализации графов использовался алгоритм укладки «Force Atlas 2», реализованный в программном продукте Gephi (Bastian M., Heymann S., Jacomy M., «Gephi: an open source software for exploring and manipulating networks», International AAAI Conference on Weblogs and Social Media, 2009). Для расчёта связанности тематических кластеров и дальнейшей визуализации, веса связей для узлов из разных кластеров объединяются и отражаются в виде ребра между центрами кластеров.

К решению задачи распознавания именованных сущностей применен метод на базе подхода CharSCNN , основанного на свёрточной нейронной сети. Этот подход впервые был предложен в 2014 году для анализа тональности (C. Santos, M. Gatti . Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts . Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 69–78). Идея CharSCNN заключается в анализе предложения, начиная с уровня символов слов и заканчивая анализом предложения в целом.

Использовалась свёрточная нейронная сеть, объединяющая две последовательно обучаемые модели: модель уровня слов и модель уровня предложения. Топология модели уровня слов: One-hot кодировщик, Свёрточный слой (1024 нейрона, окно 5), GlobalMaxPooling1D, BatchNormalization, 2 Dense слоя (256 нейронов, функция активации relu), BatchNormalization, выходной Dense слой с активацией softmax. После обучения этой модели, у неё убирались последние 2 слоя, и она добавлялась в модель уровня предложения: модель уровня слов, Dropout(0.5), 2 свёрточных слоя (256 нейронов, окно 3, функция активации relu),

		выходной свёрточный слой с активацией softmax. Слова текста были представлены в виде символьных n-грамм фиксированной длины. Предложения тоже имели фиксированную длину и в случае необходимости дополнялись нулевыми n-граммами. Для выявления информационных вбросов был исследован характер распространения новостей в Интернете. Отличительной чертой многих информационных вбросов является отношения большого количества дубликатов новости к маленькому количеству её оригиналов. На основе этого характерного признака была разработана методика по выявлению информационных вбросов в сети Интернет, которая включает в себя: - выделение новостей по интересующей тематике на основе ключевых слов и словосочетаний; - определение дубликатов (перепечаток) и оригиналов этих новостей; - визуализацию графиков отношение количества оригиналов новостей к количеству дубликатов новостей за определённый период. Для определения не только точных дубликатов выделенных новостей по определенной тематике, но и их перепечаток, в которые включены некоторые изменения, использовался метод векторного представления текстовых документов (векторная модель) с методом взвешивания TF-IDF. В качестве меры сходства реализовывалась функция, основанная на словах с наибольшим весом в сравниваемых документах. Для документов различной длины были определены свои минимальные степени похожести. Из всех выявленных дубликатов, оригиналом выбирается новость, у которого указана наиболее ранняя дата создания новости. После визуализации отношения количества оригинальных новостей к количеству их неявных дубликатов за определённый период появляется возможность выявить аномальное поведение новости, характерное для информационных вбросов. При проведении социологических исследований основной метод – стандартизированное интервью с использованием формализованной, электронной анкеты. В анкетировании приняли участие 6032 студентов очной формы обучения всех курсов бакалавриата всех институтов и филиалов КФУ. Источником информации послужили также статистические данные, размещенные на официальных сайтах Федеральной службы государственной статистики Российской Федерации, Территориального органа Федеральной службы государственной статистики по Республике Татарстан, в том числе «Окончательные итоги Всероссийской переписи населения 2010 г.». Исследование реализуется на основе сочетания качественной (гуманистической) и количественной (статистической) стратегий. На разных этапах исследования выявляются ключевые внутренние (обусловленных состоянием физического здоровья и изменениями в установках) и внешние (обусловленных средой вуза) факторы, оказывающие влияние на состояние здоровья. Междисциплинарный характер исследования обусловливает необходимость сочетания данных подходов. Все использованные в проекте методы являются либо оригинальными, либо модификацией известных методов, либо комбинацией оригинальных методов и наиболее современных методов и технологий.
50.8.3.	Полученные в ходе выполнения Проекта результаты-объекты интеллектуальной собственности (номера патентных заявок и	Объектами интеллектуальной собственности являются научные результаты, опубликованные в статьях, перечисленные в списке публикаций. Заявки на патенты не подавались.

	т.п.)	
50.9.	Участие в научных мероприятиях по тематике Проекта за период, на который предоставлен грант (каждое мероприятие с новой строки, указать названия мероприятий и тип доклада)	1. Четвертая международная конференция "Cognitive modeling" (Испания), пленарный доклад. 2. Международная конференция по компьютерной и когнитивной лингвистике (Казань), секционный доклад. 3. Седьмая международная конференция по когнитивной науке (Светлогорск), секционный устный доклад. 4. International Conference "Supercomputer Simulations in Science and Engineering" (Москва), секционный устный доклад. 5. Всероссийская научная конференция "Междисциплинарность в современном социально-гуманитарном знании" (Ростов-на-Дону), стендовый. 6. Международная научная конференция «ПРОБЛЕМЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ», (Воронеж), секционный 7. 8th ESE International Conference on Sports, Health and Management (Париж), секционный 8. VII-я международная социологическая Грушинская конференция. Навстречу будущему. Прогнозирование в социологических исследованиях. (Москва), секционный 9. Всероссийская научно-практическая конференция с международным участием «Социальная справедливость – основа общественного здоровья». (СПб), секционный 10. «Здоровье – основа человеческого потенциала: проблемы и пути их решения» xii всероссийская научно-практической конференции с международным участием. (СПб), секционный 11. Пятый международный форум по когнитивному моделированию (Лиссабон), стендовый 12. DIGITAL TRANSFORMATION & GLOBAL SOCIETY (СПб), стендовый. 13. 3rd International Conference on Social Science, (Шанхай), секционный. 14. Международный конгресс по когнитивной лингвистике (Белгород), пленарный 15. 18-ая международная конференция CML-2017 (Лиссабон), пленарный
50.10.	Участие в экспедициях по тематике Проекта за период, на который предоставлен грант	В экспедициях не участвовали.
50.12.	Адреса (полностью) ресурсов в Интернете, подготовленных авторами по данному проекту, например, http://www.somewhere.ru/mypub.html	Сайт проекта https://kpfu.ru/kompjuternye-modeli-i-matematicheskie-metody-dlya.html Исходный код модуля извлечения событий: https://www.dropbox.com/sh/rnxtdutr0dxyyhv/AADY7GhAXhZWH070lQmRPekUa?dl=0 Данные словаря Russian FrameNet: http://framenet.s3-website-us-east-1.amazonaws.com/ Набор 2- и 3-грамм корпуса новостных текстов: https://drive.google.com/open?id=1hYxvj-wqWsZL4UJ0Q6isaSARSXq4pZX9

		Набор данных для новостного корпуса, содержащий вектора слов, построенные по модели word2vec: https://drive.google.com/open?id=1ae9DLQEDYgbsQu4gweG11w93Ze-j9x-T.
50.13.	Библиографический список всех публикаций по проекту, опубликованных за период, на который предоставлен грант, в порядке значимости: монографии, статьи в научных изданиях, тезисы докладов и материалы съездов, конференций и т.д.	1. Valery Solovyev, Ildar Batyrshin. What does happiness depend on? Quantitative comparative analysis of various cultures. Advances in Social and Behavioral Sciences. v.10, p.3-9, 2015. 2. Dmitry Gudovskikh, Sboev Alexander, Moloshnikov Ivan and Roman Rybka. A quantitative method of text emotiveness evaluation on base of the psycholinguistic markers founded on morphological features // Procedia Computer Science 2015, Volume 66, 2015, pp. 307–316 3. I.A. Moloshnikov, A.G. Sboev, R.B. Rybka, D.V. Gydovskikh, An algorithm of finding thematically similar documents with creating context-semantic graph based on probabilistic-entropy approach, Procedia Computer Science, Volume 66, 2015, Pages 297–306, 2015 г. 4. Котов А. А., Зинина А. А. Функциональный анализ невербального коммуникативного поведения // Компьютерная лингвистика и интеллектуальные технологии. Вып. 14 (21). — М.: Изд-во РГГУ, 2015.– Т. 1.– С. 308-320. 5. S.I. Nikolenko. SVD-LDA: Topic Modeling for Full-Text Recommender Systems. Lecture Notes in Computer Science, vol. 9414, Springer, 2015, pp. 67–79. 6. S.I. Nikolenko. A Probabilistic Rating System for Team Competitions with Individual Contributions. Communications in Computer and Information Science, vol. 542, Springer, 2015, pp. 3–13. 7. A. Sboev, I. Moloshnikov, D. Gudovskikh, R. Rybka. Visualization of Subtopics of the Thematic Document Collection Using the Context-Semantic Graph. Proceedings 2015 International Conference on Computational Science and Computational Intelligence. IEEE Computer Society. 2015. pp. 47—52. 8. Soloviev V.D., Bochkarev V.V., Kaveeva A.D. Variations of social psychology of Russian society in last 100 years // IEEE Conference Proceedings: The 8th IEEE International Conference on Social Computing and Networking, IEEE CS Press, 2015, pp.519-524. 9. Velichkovsky B., Solovyev V., Bochkarev V., Ishkineeva F. Transition to market economy promotes individualistic values: Analysing changes in frequencies of Russian words from 1980 to 2008. International Journal of Psychology. V.52, 2017. DOI: 10.1002/ijop.12411. 10. Соловьев В.Д., Бочкарев В.В. Лексический подход к оценке динамики уровня благополучия в обществе. Интеллект. Язык. Компьютер. Вып. 17, Казань: Издательство Казанского университета. 2016. С. 146-149. 11. Batyrshin I.Z., Kubysheva N., Solovyev V., Villa-Vargas L.A. Visualization of similarity measures for binary data and 2 x 2 tables. Computacion y Sistemas, Vol. 20, No. 3, 2016, pp. 345–353. 12. Ishkineeva F., Kaveeva A., Ozerova K., Gaifullina R., Minzaripov R. Medico-Social Markers of A Federal University Students' Health. Research Journal of Pharmaceutical, Biological and Chemical Sciences. v.7, № 6, 2016. pp. 2828-2831. 13. Ishkineeva F., Zaznaev O., Kaveeva A. Political Views of Russian Youth as the Marker of Social Well-Being. International journal of humanities and cultural studies. V.3, 2016. pp. 370-375. 14. Соловьев В. Д., Бочкарев В. В., Байрашева В. Р. Dynamics of emotions in European languages. Седьмая международная конференция по когнитивной науке. М.: Институт психологии РАН. 2016. с. 71-72.

15. Соловьев В. Д., Бочкарев В. В., Шевлякова А. В., Байрашева В. Р. Размерность пространства эмоций: диахронический подход. Cognitive Modeling in Linguistics. The IV-th International forum on cognitive modeling (IFCM-2016). Р-на-Д: Фонд науки и образования, 2016. с. 432-440.
16. Solovyev V., Bayrasheva V. How Attitudes toward Health Have Changed in Developed Countries (An Answer Based on Data from Google Books Corpus). 2017 7TH ESE INTERNATIONAL CONFERENCE ON SPORTS, HEALTH AND MANAGEMENT. Singapore Management and Sports Science Institute, 2017. p.11-16.
17. Solovyev V., Bayrasheva V., Akhtiamov R. Change of Attitudes to Work and Leisure in the Society. 2ND SSR INTERNATIONAL CONFERENCE ON SOCIAL SCIENCES AND INFORMATION. 2017. p. 96-100.
18. K. GURIN, A. KAVEEVA, V. SOLOVYEV. Discursive Structure of Media Content Discussion by Users of Online Social Networks. 3rd International Conference on Social Science. DEStech Publications, Inc. 2017. p.1327-1332.
19. В. Байрашева, А. Кавеева. ИСТОРИКО-КОГНИТИВНО-ЛИНГВИСТИЧЕСКИЙ ПОДХОД В FOOD STUDIES. Пятый международный форум по когнитивному моделированию. Труды. Р-на-Д: Фонд науки и образования, 2017. с. 222-227.
20. F. Ishkineeva, K. Ozerova, A. Kaveeva, R. Akhtiamov. Availability of sport infrastructure in complex of sport motivations of Russians. Lecture Notes in Management Science. Singapore Management and Sports Science Institute. 2018.
21. А. Г. Сбоев, Д. В. Гудовских, И. А. Молошников, А. В. Наумов. МЕТОД ВЫЯВЛЕНИЯ ИНФОРМАЦИОННЫХ ВБРОСОВ В ИНТЕРНЕТ ИСТОЧНИКАХ. ПРОБЛЕМЫ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ И ТИПОЛОГИИ. т.6., 2017. с. 152-158.
22. Ишкинеева Ф. Ф., Кавеева А.Д., Озерова К.А., Минзарипов Р.Г. Мониторинг социального и физического самочувствия студенчества: опыт эмпирического исследования. Социология медицины. т.6, №2, 2017.
23. A. Kaveeva, F. Ishkineeva, K. Ozerova. IMPACT OF SPORTS INFRASTRUCTURE ON PUBLIC HEALTH: QUANTITATIVE ANALYSIS. QUID-INVESTIGACION CIENCIA Y TECNOLOGIA. vol.28, pp. 853-858. 2017.
24. A. KAVEEVA, V. BAYRASHEVA. Cultural Shifts In Developed Countries In The Last Two Centuries: Attitude To Nutrition. Astra Salvensis. 2017. pp.157-161.
25. A. Kaveeva, V. Solovyev, F. Ishkineeva. Expert Assessment of Sports Policy in Russia. Lecture Notes in Management Science. vol. 91, pp. 3-7. 2018.
26. Kotov A. Zinina A., Filatov A. Semantic Parser for Sentiment Analysis and the Emotional Computer Agents. Proceedings of the AINL-ISMW FRUCT. 2015. с. 167-170.
27. Batyrshin I.Z. Villa-Vargas L.A., Solovyev V. D. Association measures on the set of subintervals of [0,1]. Proceedings NAFIPS. 2015. p.1-3.
28. Соловьев В. Д. Google Books Ngram и социогуманитарные тренды развития культуры и общества. The III International forum on cognitive modeling. Proceedings. 2015. с. 417-420.
29. Gudovskikh D.V., Moloshnikov I.A., Naumov A.V., Rybka R.B., Sboev A.G., Selivanov A.A. A probabilistically entropic mechanism of topical clusterization along with thematic annotation for evolution analysis of meaningful social information of internet sources. Lobachevskii Journal of Mathematics. vol.38, is. 5, p. 910–913.

		30. Valery D. Solovyev, Vladimir V. Bochkarev, Anna V. Shevlyakova, Raouf B. Akhtiamov. Computational approach to the study of global social trends. Advances in Education Sciences. 2018.
50.14.	Приоритетное направление развития науки, технологий и техники РФ, которому, по мнению исполнителей, соответствуют результаты данного проекта	Информационно-телекоммуникационные системы
50.15.	Критическая технология РФ, которой, по мнению исполнителей, соответствуют результаты данного проекта	Нано-, био-, информационные, когнитивные технологии
50.16.	Основное направление технологической модернизации экономики России, которому, по мнению исполнителей, соответствуют результаты данного проекта	не очевидно
50.17.	Направление из Стратегии научно-технологического развития Российской Федерации	Возможность эффективного ответа российского общества на большие вызовы с учетом взаимодействия человека и природы, человека и технологий, социальных институтов на современном этапе глобального развития, в том числе применяя методы гуманитарных и социальных наук

Материал, в научно-популярной форме иллюстрирующий основные результаты проекта

В обществе постоянно происходят изменения как локальные, так и глобальные. Они изучаются социологами в содружестве с представителями многих других наук. В последнее время, благодаря развитию информационных технологий, появились новые возможности и новые направления исследований социальной эволюции. Это связано, в первую очередь, с появлением больших данных, затрагивающих общественную жизнь. Всем хорошо известны социальные сети, в которых уже накоплены гигантские объемы информации. Анализ этой информации очень важен во многих аспектах (упомянем только проблему терроризма и экстремизма). Однако использование стандартных средств компьютерной лингвистики для анализа сообщений в соцсетях затруднено многими обстоятельствами – использованием несловарных форм, краткостью текстов в твиттере и т.д. А главное, сверхбольшим объемом неструктурированных данных, замусоренным фейковыми аккаунтами, информационными вбросами и т.д., из которого трудно вычленят нужную информацию. Это же относится и к новостным агрегаторам.

В рамках проекта разработана группа методов анализа подобных массивов текстов с выделением ключевых тем. Результаты анализа могут быть использованы, в частности, для прогнозирования развития новостных трендов и даже глобальных трендов общественного развития. Предложен метод анализа развития новостных тем и подтем с наглядным представлением результатов в виде карты эволюции аннотированных вложенных подтем. Методы протестированы на созданном нами корпусе, содержащем 3 млн. новостных русскоязычных текстов. Разработана методика по выявлению информационных вбросов в сети Интернет. Методика продемонстрировала возможность выявления информационных вбросов на базе анализа неявных дубликатов документов. Проанализировано влияние фейковых аккаунтов на структуру соцсетей.

Данное направление исследований ограничено временными рамками – соцсети появились совсем недавно. В то же время информация о социуме накапливалась веками в книгах. С 2010 г. благодаря революционному проекту Google Books появилась коллекция беспрецедентно большого объема оцифрованных книг на 9 языках, охватывающая полтысячи лет. Сервис Google Books Ngram обеспечивает удобный графический интерфейс с этой коллекцией, предоставляя данные о частоте встречаемости слов и словосочетаний в каждый год. Этот ресурс позволяет выявлять долговременные тренды развития общества и в последние годы на его основе выполнено много интересных исследований во всем мире.

Приведем лишь два из наиболее интересных результатов, полученных в проекте. Для СССР/России описана взаимосвязь социодемографических параметров (доля городского населения и т. д.), с развитием индивидуализма, характеризующимся изменением частоты употребления определенных слов, включая личные местоимения. Получены количественные и качественные результаты по эволюции ключевых идей языковых картин мира (счастье, развитие, работа, отдых, война и др.) у разных народов, а также эволюции отношения к питанию, здоровью и удовлетворенностью жизнью в целом. Например, оказалось, что в русском, французском и итальянском языках наблюдается увеличение доли слов, обозначающих положительные эмоции, по сравнению с отрицательными, что можно интерпретировать, как рост удовлетворенности жизнью, а в английском, немецком – снижение. Испанский занимает промежуточное положение. Эти примеры указывают на широкие возможности исследований с применением Google Books Ngram.

Итогом выполнения проекта является компьютерная платформа и спектр математических методов и программных средств для изучения социогуманитарных трендов и их взаимосвязей на основе анализа больших данных.