

РЕФЕРАТ

Отчет 109 с., 7 рис., 2 табл., 114 источн., 7 прил.

ОТКРЫТЫЕ ИНФОРМАЦИОННЫЕ ИСТОЧНИКИ, СМИ, ИНФОРМАЦИОННАЯ СИСТЕМА, ИНФОРМАТИВНЫЕ ПРИЗНАКИ, ТОНАЛЬНОСТЬ, ГЕНЕРАЛИЗАЦИЯ, ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, КЛАСТЕРИЗАЦИЯ, ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, МАШИННОЕ ОБУЧЕНИЕ, БОЛЬШИЕ ДАННЫЕ, ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТА, ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ ФРАЗ, ИНФОРМАЦИОННЫЙ ПОВОД

Объектом исследования являются открытые электронные текстовые информационные источники новостного характера и их контент.

Целью данной работы являются исследование и разработка различных методик для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации и их алгоритмическая реализация в составе соответствующей информационно-аналитической системы.

Для достижения цели и решения задач исследования на каждом этапе работ применялись различные методы и подходы, такие как: социологический опрос, экспертный опрос, метод теоретического анализа и обобщения научной и аналитической литературы, аддитивный метод построения индексов, общенаучные методы наблюдения, систематизации, обобщения, статистический метод, лингвистические методы контекстуального, дискурсивного, интерпретативного, прагматического и компонентного анализа, интроспекция, применялись технологии больших данных и машинного обучения и др.

Основные показатели: Созданы методики получения оценки влияния на социум открытых информационных источников на базе анализа текстовых публикаций (в численном виде): на базе тематической модели корпуса СМИ и на базе комплексной оценки различных информативных признаков. Начата разработка методики расчета отдельных критериев для оценки влияния открытых текстовых информационных источников на социум: разработаны и апробированы новый алгоритм тематического моделирования на основе кластерного подхода, метод автоматического установления тональности текстов методом их концептуального анализа, метод для автоматической численной оценки степени информативности публикаций, подход для автоматической классификации публикаций по типам, жанрам и стилям; проведен анализ методов, обеспечивающих выявление деструктивных сообщений в интернет-среде; проведена оценка выраженности информативных критериев на основе словарного подхода. Начата разработка информационной системы оценки влияния открытых текстовых информационных источников на социум: определены назначение, цели задачи, разработана архитектура и определены необходимость и достаточность функциональности информационной системы. Созданы необходимые технические и экспертно-аналитические условия для разработки информационной системы: разработаны новые методы кластерного анализа, метод плотной тематической векторизации текстов, содержащихся в большом в текстовом корпусе (BigData), алгоритм группировки новостных публикаций в соответствии с информационными поводами, метод кластеризации новостных сообщений СМИ на основе их концептуального анализа, алгоритм кластеризации больших данных высокой размерности на основе метода декомпозиции, методы, алгоритмы и инструменты предобработки текстовых публикаций, метод автоматической суммаризации текстовых документов и методы оптимизации.

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 10](#_Toc22749658)

[1 РАЗРАБОТКА МЕТОДИКИ ДЛЯ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ НА ОСНОВЕ АНАЛИЗА ПУБЛИКУЕМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ 12](#_Toc22749659)

[1.1 Методика оценки влияния открытых информационных источников на социум на основе мультимодальной оценки СМИ на базе тематической модели корпуса публикаций 12](#_Toc22749660)

[1.1.1 NLP и MCDA 13](#_Toc22749661)

[1.1.2 Процесс оценки СМИ 15](#_Toc22749662)

[1.1.3 Алгоритм мультимодального оценивания СМИ 18](#_Toc22749663)

[1.1.4 Анализ социальной значимости тематических категорий 20](#_Toc22749664)

[1.1.5 Методика формирования словарей социально-значимых тематик 22](#_Toc22749665)

[1.2 Дорабока методики оценки влияния открытых информационных источников на социум на основе комплексной оценки различных информативных признаков 22](#_Toc22749666)

[2 РАЗРАБОТКА МЕТОДИКИ РАСЧЕТА ОТДЕЛЬНЫХ КРИТЕРИЕВ ДЛЯ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ 25](#_Toc22749667)

[2.1 Определение состава, классов и значимости информативных признаков 25](#_Toc22749668)

[2.2 Тематическое моделирование 26](#_Toc22749669)

[2.2.1 Тематическое моделирование на основе библиотек BigARTM и GenSim 26](#_Toc22749670)

[2.2.2 Алгоритм тематического моделирования на основе кластерного подхода 27](#_Toc22749671)

[2.3 Автоматическое установление тональности текстов методом их концептуального анализа 33](#_Toc22749672)

[2.4 Генерализация 33](#_Toc22749673)

[2.5 Степень информативности публикаций 34](#_Toc22749674)

[2.6 Классификация публикаций по типам, жанрам и стилям 34](#_Toc22749675)

[2.7 Вероятная деструктивность 36](#_Toc22749676)

[2.8 Оценка выраженности в анализируемом тексте информативных критериев на основе словарного подхода 36](#_Toc22749677)

[3 ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ 40](#_Toc22749678)

[3.1 Назначение, цели и задачи разрабатываемой системы 40](#_Toc22749679)

[3.2 Архитектура информационной системы 41](#_Toc22749680)

[3.3 Определение необходимой и достаточной функциональности разрабатываемой информационной системы для качественного решения поставленных задач проекта 45](#_Toc22749681)

[4 МОНИТОРИНГ И КОРРЕКТИРОВКА ОТКЛОНЕНИЙ В РАБОТЕ РАЗРАБОТАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ 47](#_Toc22749682)

[5 СОЗДАНИЕ НЕОБХОДИМЫХ ТЕХНИЧЕСКИХ И ЭКСПЕРТНО-АНАЛИТИЧЕСКИХ УСЛОВИЙ ДЛЯ РАЗРАБОТКИ ИНФОРМАЦИОННОЙ СИСТЕМЫ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ 48](#_Toc22749683)

[5.1 Методы кластерного анализа и их применение при анализе корпуса публикаций СМИ 48](#_Toc22749684)

[5.1.1 Плотное векторное представление текстов на основе анализа взаимной встречаемости слов и тематического моделирования в условиях BigData 50](#_Toc22749685)

[5.1.2 Кластеризация новостных публикаций по информационным поводам 54](#_Toc22749686)

[5.1.3 Метод кластеризации новостных сообщений СМИ на основе их концептуального анализа 55](#_Toc22749687)

[5.1.4 Алгоритм кластеризации больших данных высокой размерности на основе метода декомпозиции 56](#_Toc22749688)

[5.2 Методы, алгоритмы и инструменты предобработки текстовых публикаций 62](#_Toc22749689)

[5.2.1 Автоматическое реферирование текстовых документов с помощью word mover’s distance и извлеченных ключевых слов документа 63](#_Toc22749690)

[5.3 Разработка методов оптимизации 65](#_Toc22749691)

[5.4 Создания необходимых технических условий 66](#_Toc22749692)

[5.5 Создание необходимых экспертно-аналитических условий 68](#_Toc22749693)

[ЗАКЛЮЧЕНИЕ 72](#_Toc22749694)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 73](#_Toc22749695)

[ПРИЛОЖЕНИРЕ А](#_Toc22749696) [Календарный план работы 84](#_Toc22749697)

[ПРИЛОЖЕНИЕ Б](#_Toc22749698) [Список публикаций 89](#_Toc22749699)

[ПРИЛОЖЕНИЕ В](#_Toc22749700) [Список использованных зарубежных информационных ресурсов 95](#_Toc22749701)

[ПРИЛОЖЕНИЕ Г](#_Toc22749702) [Основные параметры текстов СМИ. 96](#_Toc22749703)

[ПРИЛОЖЕНИЕ Д](#_Toc22749704) [Эксперименты по применимости алгоритма мультимодального оценивания СМИ 98](#_Toc22749705)

[ПРИЛОЖЕНИЕ Е](#_Toc22749706) [Результаты ранжирования тематик по методу 1.3 104](#_Toc22749707)

[ПРИЛОЖЕНИЕ Ж 108](#_Toc22749708)

# ВВЕДЕНИЕ

На современном этапе технологического развития общества всё большее социальное значение приобретает цифровая информация, получаемая из различных сетевых информационных источников. Происходит непрерывное увеличение объёмов общедоступной информации. Отдельные индивидуумы и их группы являются как источниками размещаемой в сети информации, так и её потребителями.

Сама по себе информация, представленная в сетевых источниках, может иметь различную степень социальной значимости, может оказывать различные влияния, как на отдельные социальные группы, так и на весь социум в целом. В виду этого возникает важная научно-практическая задача, по комплексной оценке, потенциального воздействия различной информации на различные социальные группы или на социум в целом.

Целью данного проекта является исследование и разработка различных методик для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации и их алгоритмическая реализация в составе соответствующей информационно-аналитической системы.

Разработка данных методик ведётся с учетом того, чтобы в последующем их можно было использовать в составе более обобщённой системы социального доверия. В данном случае под системой социального доверия (ССД) подразумевается способ организации общества на основе численных оценок доверительных отношений между составляющими его субъектами (индивидуумами, группами и организациями). Построение полноценных систем социального доверия требует применения комплексного анализа множества различных информативных признаков, выраженных в численном виде. Каждый из этих признаков может относиться к различным социально-экономическим сферам, отражающим взаимоотношения индивидуумов и их групп с обществом. Взаимодействие в информационном пространстве отдельных индивидуумов и социальных групп также является важным аспектом при построении ССД. Многокритериальная оценка воздействия информационных источников на социум может использоваться в качестве одного из значимых факторов в рамках ССД. В качестве информационных источников могут выступать как отдельные индивидуумы, их группы так и целые организации.

В современное время общество всё больше приобретает черты глобального информационного общества. При этом одним из важнейших современных глобальных вызовов является развитие индивидуальной культуры информационного потребления. Разрабатываемые в рамках данного проекта технологии многокритеритериальной оценки влияния открытых информационных источников на социум способны выступить в качестве одного из инструментов для развития индивидуальной информационно-потребительской культуры. В данной парадигме индивидуум выступает не просто как потребитель информации, но он получает доступ к информации о потребляемой им информации в виде многокритериальной численной оценки. Также, как и общество получает информацию о социально значимых аспектах его информационного пространства, имеющихся в нём информационных источниках и их многокритериальной численной оценки [1-4]. Например, наличие информации об альтернативных точках зрения по исследуемому вопросу при информационном поиске в сети способно оказать существенный положительный эффект на информационно потребительскую культуру отдельной личности и тем самым способствовать её устойчивому развитию.

В соответствии с календарным планом в рамках данного проекта на 2019 отчётный год запланирована реализация следующих задач:

1. Разработка методики для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации.
2. Разработка методики расчета отдельных критериев для оценки влияния открытых текстовых информационных источников на социум
3. Проектирование и разработка информационной системы оценки влияния открытых текстовых информационных источников на социум
4. Мониторинг и корректировка отклонений в работе разработанной информационной системы оценки влияния открытых текстовых информационных источников на социум
5. Создание необходимых технических и экспертно-аналитических условий для разработки информационной системы оценки влияния открытых текстовых информационных источников на социум.

Все поставленные задачи выполнены в полном объёме и их текущее состояние отражено в данном промежуточном отчёте.

# РАЗРАБОТКА МЕТОДИКИ ДЛЯ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ НА ОСНОВЕ АНАЛИЗА ПУБЛИКУЕМОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

## Методика оценки влияния открытых информационных источников на социум на основе мультимодальной оценки СМИ на базе тематической модели корпуса публикаций

По результатам исследования Edelman Trust Barometer 2019, проведенного в 27 странах доверие к информации правительства и медиаканалам остается низким. Растет разрыв между информированной публикой и основной массой населения (2018 – 13 пунктов, 2019 – 16) [5].

В тех случаях, когда аудитория не обладает достоверными знаниями или опытом в отношении происходящего, она особенно зависят от информации, предоставленной СМИ [6]. Согласно исследованиям [7], [8], – для формирования, определенного мнения или фокусирования аудитории на конкретных темах, СМИ используют различные техники и механизмы манипулирования. Дополнительным фактором, влияющим на наше восприятие, является увеличение доступности различных новостей в сети Интернет, что может создавать путаницу, вызванную личными, часто субъективными, источниками информации, такими как личное телевидение, блоги и непроверенные новости [9]. Вместе с тем СМИ является некоторой системой обратной связи между государственными органами и обществом.

В связи с этим, важно понимать, каким образом СМИ могут использовать свое влияние, чтобы можно было смягчить отрицательные воздействия СМИ и поощрить положительные эффекты [10].

Оценка контента средств массовой информации находится в фокусе исследователей ввиду ее практической значимости для новостных агентств, рекламных компаний и государственного сектора. На основе автоматического анализа контента СМИ можно прогнозировать вероятную популярность новостных статей, предоставлять подходящие и качественные статьи для пользователей, планировать PR-стратегии по продвижению товаров или услуг (Alexandru Tatar et al.[[1]](#footnote-1) , Roja Bandari et al.[[2]](#footnote-2) ). Государственный сектор получает инструмент для продвижения и разъяснительной работы по нововведениям, выстраивания медиа-плана и определения негативного контента, запрещенного законодательством. Отдельные пользователи могут оперативно и эффективно отфильтровать большое количество информации.

Каждая публикуемая СМИ статья обладает некоторой совокупностью параметров, часто разнородных. Например, это могут быть как объективные показатели (число ссылок и постов), так и субъективные (эмоциональная окраска или тональность). По существу, нам необходимо, отнести публикуемый текст к определенному классу, что позволяет оценить, например, степень воздействия его на читателей или необходимость более детального анализа. Для этого следует выявить указанные параметры (критерии), оценить их взаимную важность и величину присутствия данного параметра в анализируемом тексте. Затем требуется агрегировать значения параметров для того, чтобы получить оценку принадлежности текста к определенному классу (рисунок 1). Получив указанные оценки по совокупности статей одного СМИ, и, объединив некоторым образом эти оценки, мы можем классифицировать СМИ.

Отметим, что указанные оценки для одной статьи и для одного СМИ могут быть не нулевыми сразу для нескольких параметров и классов. Кроме этого, желательно получать оценки СМИ, имея только часть корпуса публикуемых ими текстов, поскольку получение всех опубликованных статей может быть затруднительным.

## NLP и MCDA

Учитывая, разнообразие цифровых источников информации и соответственно значительные объемы новостных сообщений, существует растущая потребность в их автоматическом анализе. Оценка параметров текстов СМИ может строиться на основе методов обработки естественного языка (natural language processing), где в настоящее время достигнут значительный прогресс в разработке моделей и методов анализа текстов. В свою очередь, агрегирование значений разнородных параметров часто выполняется с использованием систем многокритериального анализа и поддержки принятия решений (multiple-criteria decision making (MCDM)).

Одним из методов, продуктивно применяемых в области NLP, является тематический анализ или тематическое моделирование. Тематическое моделирование – метод, основанный на статистических характеристиках коллекций документов, который применяется в задачах автоматического реферирования, извлечения информации, информационного поиска и классификации [11]. Смысл данного подхода заключается в интуитивном понимании того, что документы в коллекции образуют группы, в которых частота встречаемости слов или сочетаний слов различается.

Тематическое моделирование (topic modeling) отличается хорошей проработанностью алгоритмов и методов, основанных на статистической модели языка, а использование кластеров документов, относящихся к совокупности тем, позволяет решать задачи синонимии и полисемии терминов [12]. Вероятностные тематические модели описывают документы (М) дискретным распределением на множестве тем (Т), а темы дискретным распределением на множестве терминов [13]. Другими словами, тематическая модель определяет к каким темам относится каждый документ и какие слова образуют каждую тему. Для построения тематической модели корпуса документов применяют: Вероятностный латентно-семантический анализ (PLSA), ARTM (Additive regularization of topic models) [14] и, весьма популярное, латентное размещение Дирихле (LDA) [15]. LDA может быть выражен следующим равенством:

(1)

представляющим сумму смешанных условных распределений по всем темам множества T, где p(w | t) условное распределение слов в темах, p(t | m) условное распределение тем по новостям. Переход с условного распределения p(w | t,m) на p(w | t) осуществляется за счет гипотезы условной независимости, согласно которой появление слов в новостях m по теме t зависит от темы, но не зависит от новости m, и есть общее для всех новостей. Данное соотношение справедливо, исходя из допущений об отсутствии необходимости сохранения порядка документов (новостей) в корпусе и порядка слов в новости, помимо этого, метод LDA предполагает, что компоненты и порождены непрерывным многомерным вероятностным распределением Дирихле. Целью алгоритма, является поиск параметров и, путем максимизации функции правдоподобия с соответствующей регуляризацией

(2)

nmw – число вхождений слова w, в новость m, R(φ,θ) – логарифмический регуляризатор.

Каждую тему можно рассматривать как отдельный параметр текста или СМИ в целом, значение которого определяется упомянутыми условными вероятностями. В тех случаях, когда необходимо вынести решение на основании множества разнородных параметров и альтернатив применяется ряд методов, обеспечивающих учет взаимной важности параметров и агрегирование их значений в виде одной или небольшого числа оценок. По своему характеру такие задачи относятся к области многокритериальной поддержки принятия решений, широко используемых в системах поддержки принятия решений (decision support system – DSS) [16,17,18]. В том числе в NLP [19]. MCDM использует ряд методов, обеспечивающих выработку решения на основе разнородных критериев, к которым относят [20]: Weighted Linear Combination (WLC) and Ordered Weighted Averaging (OWA) [21], PAPRIKA (Potentially all pairwise rankings of all possible alternatives/метод попарного ранжирования всех возможных альтернатив) [22], ELECTRE [23], TOPSIS (technique for order performance by similarity to ideal solution) [24], MAUT, PROMETHEE (Preference Ranking Organization METHod for Enrichment of Evaluations/Метод ранжирования предпочтений для улучшения оценки) [25], VIKOR, AHP (analytical hierarchy process) [26], а также Bayesian networks [27,28], fuzzy logic [29] и их сочетание BaFAHP [45] и т.п.

Важным этапом применения перечисленных методов является получения знаний от экспертов (knowledge extraction) [30], которые включают методы итеративного обсуждения коллегией экспертов (Delphi [31, 32]) и различные способы ранжирования критериев принятия решений (параметров) основе шкал Лайкерта [33], Моккена [34] и др. Часто используется попарное сравнение критериев в силу его психологической обоснованности, например, MaxDiff [35], суть которого заключается в выборе опрашиваемым двух факторов из заданного списка – самого важного и наименее важного, Bradley-Terry [36], который использует обработку большого количества попарных сравнений экспертами типа «более/менее важно» методом максимального правдоподобия для получения функции распределения вероятности, оценивающей любую пару критериев.

Некоторые из перечисленных выше методов, такие как PAPRIKA, PROMETHEE, TOPSIS, ELECTRE и AHP включают механизмы получения экспертных оценок.

В настоящем исследовании тематическая модель, построенная с применением LDA, использована для расчета условных распределений вероятности документов по параметрам, темам и классам. Применение тематической модели обеспечивает на наш взгляд уменьшение ошибок, связанных с расчетом соответствия текстов и словарей параметров.

Для оценки весомости вклада параметров в процесс классификации мы используем AHP, который апробирован на многих практических задачах [37].

Агрегирование полученных значений осуществляется с помощью Байесовского метода, который в итоге позволяет рассчитать вероятности гипотез о характере работы СМИ.

## Процесс оценки СМИ

Мы полагаем, что, что материалы, имеющие существенное влияние на информационную среду общества, должны рассматриваться более тщательно, в отличие от материалов частного, бытового, юмористического и т.п. характера. Для этого, предлагается, во-первых, рассматривать преимущественно статьи резонансные, вызывающие большое число откликов, во-вторых, выделить в них социально значимую группу новостей, которую и анализировать более тщательно, оценивая достоверность и объективность содержащейся в статье информации.

Следует сразу оговориться, что социальная значимость тематики далеко не всегда напрямую связана с ее резонансностью. Так, по данным «Яндекса» [38], в 2018 году пользователи из Казахстана наиболее часто делали запросы на следующие темы: 1) Чемпионат мира по футболу. 2) Зимняя Олимпиада в Пхёнчхане. 3) Бой Нурмагомедова и Макгрегора. 4) Пожар в ТРЦ «Зимняя вишня». 5) Лига наций УЕФА. 6) Убийство Дениса Тена. 7) Повышение пенсионного возраста в России. 8) Лунное затмение 27 июля. 9) Вспышка менингита в Казахстане. 10) Президентские выборы в России. Очевидно, что социально значимы для казахстанцев лишь темы «Вспышка менингита в Казахстане» и «Убийство Дениса Тена» (поскольку с уличной преступностью может столкнуться любой гражданин).

В свете сказанного, при определении оценкам социальной значимости тематик практически нет альтернативы экспертным оценкам, при этом в получении таких оценок заинтересованы и органы власти. Так, Президент Казахстана Н.А.Назарбаев на заседании Совета безопасности РК 7 ноября 2018 года назвал наиболее острые вопросы, которые волнуют казахстанцев [39]: «Итоги социологического исследования показали, что на первом месте из шести проблем, которые выделили, стоит высокая стоимость коммунальных услуг». Среди других, особо волнующих казахстанцев вопросов Президент указал дорогое медицинское обслуживание и обучение, низкое качество образования.

Исследование, проведенное Центром политического анализа и стратегических исследований партии «Нур Отан» по итогам 2017 года, показало, что 10 наиболее острых для населения Казахстана проблем таковы [40]: 1) Рост цен на продукты питания, товары первой необходимости; 2) Низкие доходы, нехватка денег; 3) Высокие тарифы на коммунальные услуги; 4) Выплата кредита; 5) Низкое качество медицинского обслуживания; 6) Коррупция; 7) Цены на ГСМ; 8) Страх потерять работу; 9) Отсутствие собственного жилья; 10) Отсутствие работы.

Обобщая итоги двух названных исследований, можно составить список тем, имеющих социальную значимость для жителей Казахстана: Цены на коммунальные услуги, Цены на продукты питания, Цены на ГСМ, Цены на жилье, Уровень доходов населения, Кредитование населения, Коррупция, Медицинское обслуживание и Образование (последние две темы включают в себя весь комплекс проблем, связанных со медициной и образованием: как стоимость, так и качество соответствующих услуг).

Таким образом, для отделения новостной и аналитической информации, требующей более тщательного анализа (класс «подозрительные») от массы статей, не требующих такого анализа (класс «не подозрительные») предлагается следующая схема, включающая несколько последовательных этапов классификации (Рисунок 1).



Рисунок 1 – Этапы классификации текстов с целью выявления «подозрительных» публикаций [41].

Обобщенно процесс анализа текста СМИ выглядит следующим образом:

* Формируется перечень параметров, определяющих принадлежность текста к одному из перечисленных выше классов;
* Оценивается сравнительная важность параметров;
* Для каждого текста, подлежащего рассмотрению, вычисляется оценка параметра;
* Оценки параметров и их сравнительная важность агрегируется для получения оценки принадлежности текста к конкретному классу;
* Используя полученные оценки текстов, формируется оценки СМИ в целом.

Для реализации описанного процесса нами были определены параметры текстов и их взаимная важность в задаче отнесения текста к каждому из обозначенных выше классов (Приложение Г) и разработан алгоритм вычисления оценок параметров, основанный на тематической модели корпуса текстов.

Предложенный подход примечателен тем, что тематическая модель, созданная путем кластерного анализа (обучение без учителя) затем применена в сочетании с экспертно-определяемыми классами и признаками. Таким образом, семантику требуемого распределения задает пользователь (эксперт), хотя изначальный тематический анализ зависит только от корпуса документов.

Использование байесовского подхода позволяет формировать вероятность соответствующей гипотезы при неполной информации, имея часть корпуса опубликованных текстов. Иными словами, мы можем обработать только часть текстов и получить оценки СМИ в упомянутых модальностях, пусть и с пониженной точностью.

## Алгоритм мультимодального оценивания СМИ

Цель алгоритма заключается в том, чтобы, используя условные распределения вероятностей документов корпуса агрегировать показатели соответствия статей тематикам, тематик свойствам (словарям) и классам для получения оценок соответствия средств массовой информации (СМИ) в трех модальностях: тематик, параметров и классов.

Исходные данные и получаемые матрицы условных вероятностей показаны на рисунке 2, где MMS (mass media sources) – множество источников текстов. СМИ (MMS) являются источником m документов (papers), которые получают путем применения систем сбора данных (процесс 2). Полученный корпус документов M разбивают на тематические кластеры T (процесс 1). Эксперты (experts) формируют классы С (процесс 4) и определяют свойства или параметры классов Q (процесс 3). Свойства описываются словарями слов, выражений (dictionaries) или процедурами их выявления в тексте (features).

Используя множество тематик корпуса, во-первых, получаем дискретное распределение условных вероятностей статей и тематик - *p2(k | m)*, где , .

Во-вторых, получаем условное распределение словарей (параметров) и тематик *p1(k | q)*, где , то есть, определяем, в какой степени параметр описывает конкретную тематику. В-третьих, с помощью аналитического иерархического процесса (AHP) рассчитываем важность параметров для классов (отдельно для каждого класса) и получаем *p3(c | q)*, где . Затем, используя p1 и p3, рассчитываем условное распределение тематик по классам - *p4(k | c)*. Зная распределение вероятностей тематики по классам (p4) и распределение вероятностей статьи по тематикам (p2), можно вычислить распределение статьи по классам *p5(m | c)*. В свою очередь, распределение статьи по признакам или словарям - *p6(m | q)* зависит от p1 и p2.

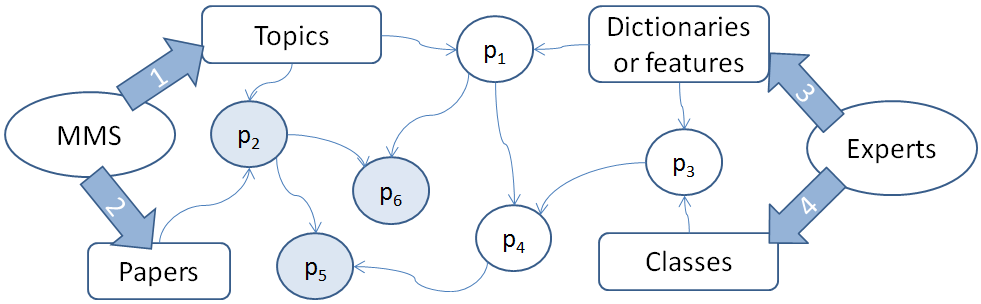
****

Рисунок 2 – Процессы определения условных вероятностей.

Для расчета вероятностей p1 и p2 используется коэффициент Жаккара [42, 43].

Для расчета множества тематик корпуса используется алгоритм LDA.

После получения указанных оценок для каждой статьи, производится оценка СМИ с использованием цепочки байесовских правил. При этом используются матрицы условных вероятностей p2, p5 и p6, выделенные на рисунке 2 цветом.

В соответствии с формулой Байеса мы можем выразить условную вероятность справедливости гипотезы h при наступлении события e в виде

(B1)

где *p(e│h)* – условная вероятность наступления события e при справедливости h, *p(h)* – априорная вероятность гипотезы h, *p(e│∽h)* - условная вероятность e при не справедливости h, *p(∽h)* - вероятность того, что событие h не верно, которое, в соответствии с формулой полной вероятности, может быть вычислено как

(B2)

Таким образом, для вычисления условной вероятности *p(h│e)* достаточно знать вероятности *p(e│h)*, *p(e│∽h)* и априорную вероятность *p(h)*.

В терминах задачи оценки СМИ, условные вероятности *p(e│h)* могут быть интерпретированы как вероятности появления некоторой статьи (где E - множество статей конкретного СМИ) при условии справедливости гипотез трех видов (каждая гипотеза для своей модальности):

1. h1: СМИ работает в одной из тематик из множества тематик T: H1={h1[1],...,h1[k]}
2. h2: СМИ публикует статьи, относящиеся к классу из множества C: H2={h2[1],...,h2[с]}
3. h3: СМИ отличается признаками из Q: H3={h3[1],...,h3[q]}

В общем случае, полагая, что у нас G модальностей и соответственно G гипотез H= {H1, H2, …, HG}, можно записать обобщенный алгоритм расчета их вероятностей для некоторого СМИ

Для всех , ,

В результате выполнения алгоритма получим множество вероятностей

для СМИ . При этом

Где Ng – количество дискретных элементов, описывающих модальность g.

Проведенные эксперименты описаны в Приложение Д и [48].

## Анализ социальной значимости тематических категорий

Тематики, имеющие социальную значимость, определяются, как правило, путем опроса экспертов, обладающих данными о состоянии общественного мнения, которые, в свою очередь, могут быть получены посредством социологических опросов населения [50].

С точки зрения определения социально значимых тематик наиболее результативным методом является экспертное оценивание. Эксперты, обладающие знаниями специфики деятельности СМИ и особенностями социальных взаимоотношений и убеждений социума, смогут внести значительный вклад в оценивание тематик, освещаемых в СМИ. Результаты тематического моделирования не предусматривают наименование топика. Использование экспертного анализа позволит: (1) дать названия полученным топикам; (2) определить топики, являющиеся значимыми для социума.

Для анализа интереса различных социальных групп к различным тематикам наиболее результативным методом является рандомизированное социологическое исследование в форме социологического опроса. Основной целью социологического исследования является определение интереса различных социальных групп к различным тематикам. Выборка должна быть произведена в соответствии с общепринятыми процедурами. Социологический опрос позволит связать тематики, освещаемые СМИ с социальными группами, у которых они вызывают интерес, изучить влияние текстовой информации в разрезе возрастных групп населения, в разрезе стратификации по принципу «город-село», регионов, гендерной стратификации, уровня образования и профессиональной деятельности респондентов.

С точки зрения оценки влияния СМИ на социум наиболее оптимальным подходом к исследованию может стать экспериментальное исследование, поскольку только такой дизайн исследования позволяет однозначно установить и доказать наличие причинно-следственных связей. Основной целью эксперимента будет установление факта влияния информативных признаков (или их части), отобранных в ходе исследования, в качестве признаков, анализируемых в рамках информационной системы. Предположительно, такими признаками будут являться: тональность, манипулятивность, объективность. Данный метод позволит установить факт влияния СМИ на социум, подтвердить необходимость учета в рамках машинного анализа контента электронных СМИ самих отобранных информативных признаков, замерить степень влияния таких информационных признаков на социум. Более детальная стратегия экспериментального исследования будет разрабатываться и обсуждаться в ходе дальнейшего исследования.

В рамках заданного параметра исследовано 2000 текстов для выявления признаков социально значимых публикаций, т.е. публикаций, касающихся социально значимых для общества вопросов, способных вызвать широкий общественный резонанс, затрагивающих интересы общества и/или способных вызвать дестабилизацию общества.

Среди языковых единиц, маркирующих социально значимые публикации, были выделены наименования социальной инфраструктуры – комплекса объектов, предприятий, организаций, которые обеспечивают функциональную жизнедеятельность населения, формирование интеллектуально развитого индивида, общества. Также к данной группе языковых единиц относятся наименования должностей и персоналий, осуществляющих деятельность в социально значимых организациях. В отдельную группу были выделены языковые единицы, маркирующие социально значимые темы.

В результате анализа в текстах зафиксированы слова и словосочетания, сигнализирующие о социальной значимости текста и разработан соответствующий словарь, включающий 2598 языковых единиц. По завершении работы по чистке и коррекции данного словаря, количество языковых единиц составило 2426 слов и словосочетаний. Наибольшее количество слов и словосочетаний выявлено в группе «Темы» – 1196, далее по убывающей в группах «Организации» – 455, «Персоны» – 418, «Должности» – 325, «Объекты» – 32.

С целью увеличения потенциальных возможностей автоматизированного поиска социально значимых публикаций, были составлены 213 синонимических рядов к социально маркированным языковым единицам. Как показали результаты исследования в рамках данного признака, с помощью языковых средств успешно могут быть выявлены новостные тексты, затрагивающие социально значимые темы. В результате работы, языковые единицы были распределены по предложенным классам слов и достаточно четко дифференцированы.

## Методика формирования словарей социально-значимых тематик

В работах [51, 52] приведен метод формирования тематических словарей, построенный на анализе семантической схожести с помощью матрицы co-occurrence и инструмента Word2Vec. В данном исследовании мы пытались показать разные способы автоматического формирования тематических словарей. Для начала мы формировали тематические словари распространенных тем: спорт, политика.

Предлагается следующий алгоритм действий выявления социально-значимых новостей (Рисунок 3):



Рисунок 3 – Алгоритм действий выявления социально-значимых новостей

Следующим этапом перед нами стояла задача выделить социально значимые понятия и сформировать их словари. Для точности понятия социальной значимости мы опирались на определение и результаты экспертных мнений. Для формирования словаря применили 3 способа. Словари будут применены для кластеризации текста с помощью алгоритмов машинного обучения.

## Дорабока методики оценки влияния открытых информационных источников на социум на основе комплексной оценки различных информативных признаков

По результатам анализа предложенной в 2018 году методики оценки влияния открытых информационных источников на социум на основе публикуемой текстовой информации проведен комплекс работ по доработке и усовершенствованию методики [50].

В комплекс работ входит: (1) формализация правил и составление словарей, (2) экспериментальное исследование полноты и корректности сформированных словарей и лингвистических правил, (3) формирование корпуса новостной текстовой информации и (4) его разметка, (5) анализ и выявление закономерностей в корпусе, (6) решение задачи несбалансированности размеченного корпуса, (7) проведение тематического моделирования, (8) разработка алгоритмов распознавания информативных признаков и (9) составление методики их расчета.

Методика представляет собой комплекс взаимосвязанных процедур (1) сбора текстовой информации, (2) тематического моделирования, (3) анализа текстовых данных по распознаванию информативных признаков и (4) выдачи аналитики по результатам анализа.

На рисунке 4 представлена общая концептуальная схема оценки новостных текстов.

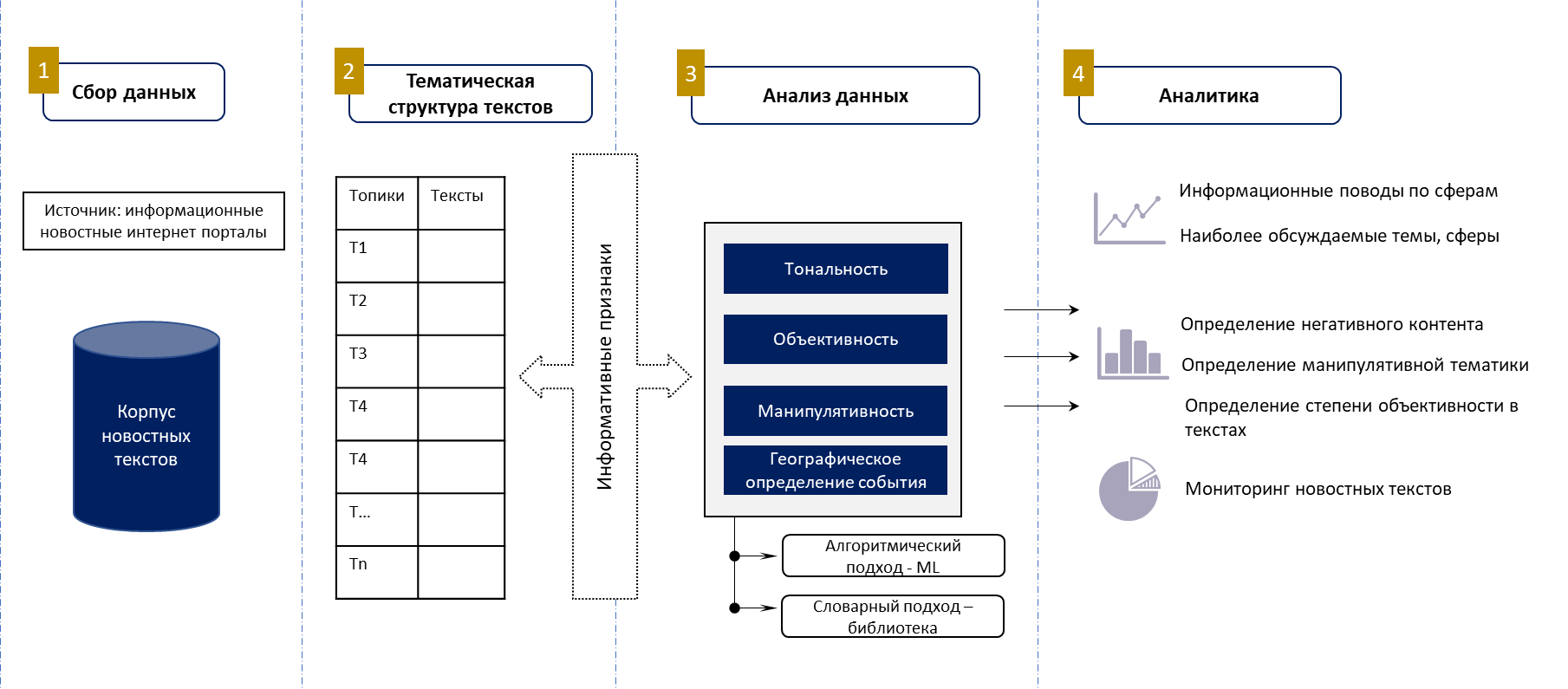


Рисунок 4 – Концептуальная схема оценки новостных текстов

На первом этапе осуществляется сбор новостных текстов из открытых текстовых казахстанских информационных интернет-источников с применением автоматических способов извлечения текстов или путем загрузки готовых коллекций (корпусов) данных.

На следующем этапе осуществляется тематическое моделирование. Большинство моделей разрабатываются на основе латентного размещения Дирихле LDA. Модель латентного размещения Дирихле (Latent Dirichlet Allocation) была предложена Дэвидом Блеем в 2003 году. [53] Это современный активно развивающийся вероятностный инструментарий, который применяется в задачах анализа данных [54]. Для решения данной задачи тематические модели оказались полезными инструментами [55].

В контексте новостных средств массовой информации тематическое моделирование, существуют два основных типичных варианта его применения: анализ повестки дня СМИ и анализ новостных фреймов. Само тематическое моделирование, как процесс, является одним из инструментов автоматизированного анализа корпуса новостных текстов и оценки их влияния на социум [56].

Тематическое моделирование при анализе новостной информации может быть применено для различных задач, как в исследовательском анализе в области социальных наук, так и при решении практических задач в государственном и частном секторах. Результаты тематического моделирования могут быть полезны при определении повестки дня СМИ, информационных поводов, интересующих вопросов, изменения временной значимости тем [55].

На следующем этапе осуществляется распознавание информативных признаков тональности, объективности, манипулятивности в новостных текстах. Распознавание проводится с применением (1) алгоритмического и (2) словарного подходов. В частности, применяются алгоритмы машинного обучения такие как Random Forest, SVM, XGboost и KNN. При втором подходе используются сформированные экспертами-лингвистами словари по информативным признакам.

На завершающем этапе осуществляется визуализация результатов анализа данных с приведением аналитики. По результатам анализа возможно определить информационные поводы по различным тематическим областям, наиболее обсуждаемые тематики, негативное содержание новостей, манипулятивность и степень объективности в текстах.

# РАЗРАБОТКА МЕТОДИКИ РАСЧЕТА ОТДЕЛЬНЫХ КРИТЕРИЕВ ДЛЯ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ

Ранее по данному проекту был проведен комплекс работ по определению состава информативных признаков.

## Определение состава, классов и значимости информативных признаков

В рамках данного проекта был проведён комплекс работ по определению состава, классов и значимости информативных признаков для последующего алгоритмического и/или экспертного анализа с целью построения соответствующих методов классификации и оценивания потенциального влияния анализируемой текстовой информации как на отдельную личность, так и на социум в целом [57, 58, 59, 60].

Обзор научной и аналитической литературы позволил определить косвенные информативные признаки, которые, возможно применить в анализе новостных текстов. Было предложено 16 информативных признаков, объединенных в 4 информативных критерия: (1) достоверность; (2) объективность или предвзятость; (3) тональность; (4) резонансность (популярность) публикации. В частности, (1) наличие ссылки в публикации на компетентный источник; (2) ссылка на первоисточник; (3) соответствие заголовка содержанию публикации; (4) освещение одного и того же события различными изданиями: кросс-чекинг, наличие дискурса с другими статьями; (5) наличие автора в статье; (6) наличие проверяемых фактов в статье; (7) репутация издания, на котором опубликована информация; (8) наличие манипулятивных техник; (9) политизированность публикации; (10) наличие в медиа-тексте выраженного личного мнения автора публикации; (11) генереализация текста; (12) тональность (позитивная/негативная); (13) количество просмотров, (14) шейров, (15) комментариев к публикации; (16) тональность комментариев (позитивная/негативная).

Для определения значимости предложенных информативных признаков и критериев проведена комплексная оценка их значимости. Процедура комплексной оценки значимости включает: (1) формирование выборки корпуса публикаций, (2) построение индексов для измерения информативных признаков, (3) построение статистической модели и (4) модели машинного обучения. Для формирования корпуса новостных текстов была проведена 2-х этапная систематическая кластерная выборка. Для определения значимости информативных признаков была проведена экспертная разметка корпуса новостных публикаций. Разметка осуществлена по 16 информативным признакам. Для построения индексов был использован аддитивный подход. Данный подход был выбран в связи с тем, что структура данных и наличие теоретических концептов позволяют использовать аддитивный подход к построению индексов, цель которых адекватно измерить предложенные критерии оценки влияния публикации на социум. В ходе проведения данной процедуры построены три индекса: индекс объективности, индекс достоверности, индекс тональности. Полученные результаты статистического моделирования показали, что предложенные информативные критерии публикации могут в дальнейшем служить каркасом для создания автоматизированной модели машинного обучения. Полученные результаты были экспериментально подтверждены посредством техники бутстрэпинг, которая показала незначительное отклонение от результатов модели.

Таким образом в предложенной Методике определены информативные признаки и критерии, проведена оценка их значимости с позиции резонансности (популярности), выработана схема оценки текстовой информации и предложен подход распознавания информативных признаков в текстах.

## Тематическое моделирование

## Тематическое моделирование на основе библиотек BigARTM и GenSim

Алгоритмы тематического моделирования позволяют выявить в обрабатываемом корпусе публикаций имеющиеся в нём тематики (топики) [61]. В последствии информация о тематической структуре анализируемых публикаций может быть использована в качестве основных или дополнительных признаков в высокоуровневых процессах классификации, кластеризации, машинного обучения и экспертного анализа.

На сегодняшний день наиболее популярными алгоритмами тематического моделирования являются:

* 1. BigARTM – библиотека алгоритмов эффективной потоковой параллельной реализации вероятностного тематического моделирования на основе аддитивной регуляризации [62].
  2. GenSim – библиотека тематического моделирования на основе алгоритмов латентного семантического анализа (LSA, LSI, SVD) и латентного размещения Дирехле (LDA) [63].

Обе данные библиотеки активно применялись нами для разложения большого анализируемого корпуса публикаций СМИ на тематики. BigARTM в виде опции тематического моделирования был использован в разрабатываемой информационной системе. Каждая из библиотек имеет как свои преимущества, так и свои недостатки. Основным их выявленным недостатком является высокие вычислительные затраты и длительное время при анализе корпусов содержаших более нескольких миллионов документов. Для преодоления вышеперечисленных недостатков нашей группой был разработан собственный алгоритм тематического моделирования на основе кластерного подхода.

## Алгоритм тематического моделирования на основе кластерного подхода

В рамках данного проекта был разработан собственный алгоритм тематического моделирования на основе кластерного подхода. В основе разработанного подхода лежит следующее допущение: контекст определяет тематическую и (или) смысловую близость слов. То есть чем ближе по составу контекстное окружение сравниваемых слов, тем больше их тематическая/семантическая близость. И действительно, очевидно, что слова-синонимы должны иметь одинаковый состав контекстного окружения. И наоборот, чем меньше смысловая связанность слов, тем сильнее различается их контекст. Под контекстным окружением понимается «мешок слов» составляющий фрагмент текста с анализируемым словом. Тонкая оценка близости контекстов даёт оценку семантической (смысловой) близости, а относительно грубая – тематической близости. Таким образом, под тематикой можно понимать относительно огрублённую оценку связанности анализируемых контекстов. Регулируя степень огрубления можно варьировать получаемое количество топиков.

Пусть имеется корпус анализируемых документов. По данному корпусу формируется его целевой словарь V размером n, который будет использоваться для последующего анализа. Если словарь слишком большой, то он может быть сокращён по мере уменьшения значимости входящих в него слов. Целевой словарь может формироваться как из слов, так и из словосочетаний. На основе данного корпуса должна быть получена симметричная квадратная матрица C частот взаимных встречаимостей в документах всех попарных комбинаций слов целевого словаря. Будем считать каждую i-ую вектор-строку матрицы C разряженным векторным представлением i-го слова из целевого словаря V. На основе косинусной меры вычисляется квадратная симметричная матрица D размером n x n. Каждый произвольный элемент матрицы D с координатами (i,j) содержит значение косинусного расстояния между векторными представлениями i-го и j-го слова.

При решении задачи тематического моделирования объектами кластеризации являются слова в составе целевого словаря V. Для пояснения сути работы алгоритма дадим следующее определение: Сгусток (протокластер) - подмножество объектов из V где все попарные расстояния, между которыми не превышают некоторого порога d1. С точки зрения кластеризации сгустки должны включать в себя как можно большее количество объектов с как можно меньшей суммой расстояний между всеми попарными комбинациями включённых в него объектов. При этом один и тот же объект если он удовлетворяет ограничению на d1 может быть включён в состав различных сгустков. Иными словами, процесс формирования одного сгустка никак не влияет на процесс формирования любого другого сгустка. Если объект был включён в один сгусток ничто не мешает нам в последующем включить его и в другой сгусток если соблюдается ограничение на d1.

Так как одни и те же объекты могут быть отнесены в различные сгустки, то это приводит к тому, что различные сгустки могут иметь полную или частичную общность своих составов. Измеряя степень общности составов двух различных сгустков можно делать выводы о степени их связанности. Степень связанности между двумя произвольными сгустками удобно определять по модифицированной мере Жаккара. Введя пороговое значение меры Жаккара (d3) можно принимать решение о связанности или несвязанности двух произвольных сгустков.

Кластер верхнего уровня - это последовательность (граф) связанных между собой сгустков.

Суть предложенного алгоритма заключается в следующем: алгоритм разбивает всё имеющееся множество объектов (слов) на взаимно пересекающиеся сгустки (протокластеры) с ограничением на максимальное попарное расстояние d1. При формировании сгустков окрестность задаётся радиусом d2. Затем полученные сгустки по мере (d3) общности их составов связываются в итоговые кластеры.

Основополагающей является процедура нахождения сгустка в заданной окрестности объекта.

Окрестность объекта задаётся dm - квадратной матрицей попарных расстояний от анализируемого объекта до всех объектов в его окрестности. Максимальное внутрикластерное расстояние задаются параметром d1. В формируемых сгустках расстояние между любыми двумя объектами не превышает d1. Параметр start\_ind - индекс объекта который будет использован как начальная точка роста при формировании сгустка. В качестве результата процедура возвращает индексы всех объектов, включённых в сгусток:

def single\_clot(dm, d1, start\_ind):

n = dm.shape[0]

if start\_ind < 0 or start\_ind > n-1:

raise ValueError('start\_ind is out of bounds')

R = np.array([start\_ind])

C = np.delete(np.arange(n),start\_ind)

while len(C) > 0:

C = C[np.sum(dm[R][:,C] <= d1, axis=0) == len(R)]

if len(C) > 0:

dist\_sum = np.sum(dm[R][:,C], axis=0)

best\_ind = np.argsort(dist\_sum)[0]

R = np.append(R,C[best\_ind])

C = np.delete(C,best\_ind)

return R

Следующая процедура запускает процесс выявления сгустков для каждого из объекта в его окрестности заданной окружностью с центром в рассматриваемом объекте и радиусом d2. Максимально допустимое попарное расстояние внутри сгустка также задаётся параметром d1. Входными данными являются: D - матрица попарных межобъектных расстояний между объектами; use\_medoid - если True, то в качестве точек роста сгустков будут использоваться медоиды. В противном случае - центры окружностей. Каждый сгусток задан множеством индексов, включённых в него объектов. В качестве результата процедура возвращает массив, в котором каждый i-ый элемент является сгустком в окрестности i-го объекта.

def all\_clots(D, d1, d2, use\_medoid = True):

n = D.shape[0]

global\_inds = np.arange(n)

clots = [np.array([0])] \* n

for i in nb.prange(n):

local\_inds = global\_inds[D[i] <= d2]

if len(local\_inds) > 0:

dm = D[local\_inds][:,local\_inds]

if use\_medoid:

start\_ind = np.argmin(np.sum(dm, axis=0))

else:

start\_ind = np.where(local\_inds == i)[0][0]

clot = single\_clot(dm, d1, start\_ind)

if len(clot) > 0:

clots[i] = local\_inds[clot]

else:

clots[i] = np.empty(0,dtype=nb.int64)

else:

clots[i] = np.empty(0,dtype=nb.int64)

return clots

В последствии выполняется процедура поиска уникальных сгустков. В результат включаются только сгустки размер, которых (количество включённых объектов) не меньше min\_size.

Возвращает набор уникальных сгустков представленных в виде list() of set(), где set() - множество индексов составляющих каждый сгусток. Один и тот же объект может быть включён в несколько сгустков:

def unique\_clots(clots, min\_size = 1):

R = set()

for clot in tqdm(clots):

if (len(clot) > 0) and (len(clot) >= min\_size):

R.add(frozenset(clot))

return np.array([set(v) for v in R])

На заключительном этапе выполняется процедура связывания сгустков в единый кластер верхнего уровня по мере их взаимных пересечений. Параметрами данной процедуры являются clots - массив сгустков в котором i-ый элемент является множеством индексов объектов, составляющих данный сгусток; n\_jobs - количество параллельных обработчиков, которые будут использованы для расчёта матрицы расстояний между сгустками. Если n\_jobs = -1, то будет задействовано всё возможное количество обработчиков. Одномоментно в кластер связываются все сгустки если они по модифицированной мере Жаккара имеют долю пересечение не менее 1-d3. Сгустки могут быть связаны как напрямую, так и опосредованно. В случае прямой связи сгустки непосредственно имеют достаточную долю пересечения. Опосредованная связь между двумя сгустками устанавливается если между ними имеется цепочка напрямую связанных сгустков (т.е. они опосредованно связываются через цепочку напрямую связанных сгустков). Таким образом все множества напрямую или опосредованно связанных сгустков образуют единые кластера. Сгустки можно рассматривать как нечёткие кластера первого уровня, которые по мере их пересечения объединяются в более высокоуровневые кластера с относительно более сложной пространственной структурой. Смысл данного метода кластеризации заключается в разбиении всего множества объектов на сгустки - простые взаимно пересекающиеся подмножества объектов с последующим установлением связей между ними, что позволяет аппроксимировать простыми взаимно пересекающимися сгустками кластеры, имеющие сложные пространственные структуры. В результате функция возвращает массив каждый элемент которого является кластером, состоящим из индексов, включённых в него объектов:

def clots\_binding(clots, d3, n\_jobs = -1):

# a, b - индексы сгустков, между которыми вычисляется расстояние по модифицированной мере Жаккара

def jaccard\_dist(a, b):

return 1 - len(clots[np.int(a[0])].intersection(clots[np.int(b[0])])) / min(len(clots[np.int(a[0])]),len(clots[np.int(b[0])]))

m = len(clots)

inds = np.reshape(np.arange(len(clots),dtype = np.int), (-1, 1))

dm = pairwise\_distances(inds, metric=jaccard\_dist, n\_jobs = n\_jobs)

labels = np.arange(m,dtype = np.int)

inds = np.arange(m,dtype = np.int)

for i in tqdm(range(m)):

old\_labels = labels[dm[i] < d3]

for j in old\_labels:

np.place(labels, labels==j, labels[i])

R = np.array([np.array(list(set().union(\*clots[labels == i]))) for i in np.unique(labels)])

return R

Также были реализованы различные методы взвешивания слов в составе топиков, слов вне состава топиков, но по мере их согласованности с топиками, а также и самих топиков:

1. Взвешивание объектов в кластере (метод 1.1): объект получает тем больший вес чем меньше его суммарное расстояние до остальных объектов в кластере.
2. Взвешивание объектов в кластере (метод 2.1): объект получает тем больший вес чем больше его суммарная совместная частота с каждым из объектов в кластере.
3. Взвешивание полного набора объектов по мере их соответствия имеющимся кластерам (метод 1.2). Объект получает тем больший вес чем меньше его суммарное расстояние до объектов сопоставляемого кластера. Возвращает матрицу размером NxM, где N - количество кластеров, M - количество объектов, элементы матрицы - коэффициенты соответствия (веса) между i-ым кластером и j-ым объектом.
4. Взвешивание полного набора объектов по мере их соответствия имеющимся кластерам (метод 2.2). Объект получает тем больший вес чем больше его суммарная совместная частота с каждым из объектов в сопоставляемом кластере. Возвращает матрицу размером NxM, где N - количество кластеров, M - количество объектов, элементы матрицы - коэффициенты соответствия (веса) между i-ым кластером и j-ым объектом.
5. Взвешивание каждого кластера по мере его соответствия полному исходному множеству объектов (метод 1.3). Кластер получает тем больший вес чем меньше среднее расстояние до всех объектов полного исходного множества. Попарные расстояния между объектами представлены матрицей D.
6. Взвешивание каждого из полученных кластеров по мере их соответствия полному исходному множеству объектов (метод 2.3). Кластер получает тем больший вес чем больше среднее частота встречаемости объектов рассматриваемого кластера с со всеми объектами полного исходного множества. Попарные частоты встречаемости объектов представлены в матрице.

Были разработаны две служебные процедуры:

1. Выбирает n первых ключевых слов для каждого из топика-кластера из слов, входящих в кластер по максимуму их веса полученных на основе методов 1.1 и 2.1.
2. Выбирает n первых дополнительных объектов (слов - кандидатов для включения в кластер) для каждого из кластера из числа всех объектов, не входящих в рассматриваемый кластер. Выбор осуществляется по соответствующим весам полученных на основе методов 1.2 и 2.2.

В качестве эксперимента был проведён тематический анализ полного корпуса новостных публикаций по одному из СМИ tengrinews.kz. Анализируемый корпус содержал 199362 публикаций за десятилетний период с 2008 по 2018 годы. Были использованы следующие значения параметров алгоритма: d1 = d2 =0.75, d3 = 0.5, min\_size = 20, use\_medoid = False

В результате проведённого тематического моделирования в анализируемом корпусе было выявлено 204 топика. Как сами топики, так и входящие в них слова были взвешены вышеперечисленными методами. Также для каждого топика автоматически был сформирован заголовок (название) состоящее из 3 наиболее весомых в данном топике слов. Также для каждого топика автоматически были определены по 3 кандидата на включение выбранные среди слов, не вошедших в данный топик (по методам 1.2 и 2.2). Результаты тематического моделирования можно скачать по следующей ссылке [64].

На основе полученных весов тематик (методы 1.3 и 2.3) было проведено ранжирование тематик по убыванию степени их согласованности со всем корпусом анализируемого СМИ. В начале списков стоят тематики являющиеся "центральными" для анализируемого СМИ (в данном случае - tengrinews.kz) и далее по убыванию. В конце находятся наиболее "нехарактерные" для данного СМИ тематики. Под «центральными» понимаются тематики наиболее лучше согласующиеся с полным набором тематик. Т.е. проходят «красной нитью» через остальные тематики.

В приложении два файла [64]. В них одинаковый набор тематик, но их взвешивание и ранжирование проведено двумя различными методами:

1. На основе расстояний между векторными представлениями (метод 1.3)
2. На основе частот взаимных встречаимостей (метод 2.3)

Результаты ранжирования тематик по методу 1.3 размещены в Приложении Е.

Как видно при субъективной визуальной оценке приведённые результаты как самого тематического моделирования, так и ранжирования слов и топиков носят довольно интерпретируемый, связанный и осмысленный характер.

На текущий момент начат комплекс работ по объективной сравнительной оценке качества разработанного алгоритма тематического моделирования по стандартным критериям, наиболее зарекомендовавшим себя в данной области. Предварительные результаты показывают, что разработанный алгоритм имеет значительное преимущество как по временным затратам, так и по качеству полученных результатов. Результаты сравнений будут опубликованы в виде отдельной статьи.

## Автоматическое установление тональности текстов методом их концептуального анализа

Одним из значимых информативных признаков для классификации текстовой информации является её тональность. В рамках данного проекта рассматривались, анализировались и разрабатывались различные подходы для численной оценки тональности анализируемых текстов [65, 66, 67]. В частности, был разработан и апробирован метод автоматического установления тональности текстов методом их концептуального анализа [67].

В статье [67] изложено решение задачи создания лингвистического инструментария и методики автоматического определения тональности новостных сообщений, связанных с качеством жизни рядового гражданина. Был определен подход для решения задачи, разработано программное обеспечение и методы автоматизированного создания словарей объектов и словарей оценочных предикатов, а также словарей модификаторов меры оценки. Проведенный эксперимент подтвердил правильность предложенной методики оценки событий, освещенных в новостных сообщениях и работоспособность программного комплекса.

Эта методика, при соответствующей выборе объектов оценки событий, может быть использована при создании тональных портретов конкретных авторов по совокупности их публикаций, а также тональных портретов различных новостных агрегатов по совокупности освещаемых ими событий в конкретный временной интервал.

Предлагаемый метод, основанный на методах автоматизированного составления словарей по текстам документов, методах семантико-синтаксического анализа и использования онтологических ресурсов, является универсальным поскольку позволяет проводить детальных семантический анализ коллекций документов по различным многомерным критериям. Комплекс тональных словарей большого объема и средства формализации текстов позволят проводить этот анализ с достаточно высокой точностью.

## Генерализация

Одним из важных признаков манипулятивности при оценке текстовых публикаций является степень генерализации (обобщения высказываний - дедукции). В рамках данного проекта был проведён комплекс работ, направленный на разработку методов численной оценки степени генерализации, их алгоритмической реализации и апробации на сформированном корпусе [68,69,70].

В частности, даётся определение понятия генерализации (обобщения), анализируется феномен «обобщения» в текстах СМИ и делается попытка синтеза лингвистической модели для ее распознавания и последующей численной оценки. Существуют исследования, которые рассматривают обобщение как признак предвзятости или как идеологическую стратегию, влияющую на мнение читателей. Однако каких-либо исследований по автоматическому распознаванию обобщения в текстах не выявлено. Предложена схема аннотации, на основе которой определены лингвистические правила для ручного определения обобщения и его последующего автоматического распознавания в текстах. Эксперимент по обнаружению обобщения в текстах официальных СМИ проводился с использованием правил.

## Степень информативности публикаций

Следующим не менее важным признаком является степень информационной насыщенности анализируемых текстовых публикаций. В рамках данного проекта был предложен и разработан метод для автоматической численной оценки степени информативности публикаций [71].

## Классификация публикаций по типам, жанрам и стилям

Публикации в СМИ преследуют разные цели. Цель одних – только информировать читателя о событии: кто сделал, что сделал, где и когда оно произошло. Другие не только информируют о событии, но и добавляют к деталям справочную информацию, обобщают информацию, делая выводы. Такие публикации содержат и факты, и мнение автора. Третьи только отталкиваются от конкретного факта и пытаются трактовать события нужным им образом, возможно, манипулируя мнением читателя.

Предлагается подход для автоматического распознавания таких типов публикаций, основанный на выделении признаков [72, 73]. Исследуем и выделяем ряд признаков, присущих трем типам публикаций. И на небольшом корпусе исследуем их применимость для автоматической классификации текстов по их типам.

Был сформирован небольшой корпус, достаточный для того, чтобы исследовать признаки и провести эксперимент. Источники имеют принципиально разную направленность: спортивный новостной сайт, конспирологический сайт, сайт с научно-популярными статьями, и сайт о бизнесе. Мы отобрали из этих источников статьи, в которых наиболее видна типизация публикаций: информативная публикация, аналитическая или артистическая.

Предлагаемый нами подход можно использовать для разного рода задач. Например, для классификации фейковых новостей или поиска пропаганды, или поиска самых свежих новостей, новостей финансового рынка или другой информации.

В своей работе мы собрали корпус из открытых источников официальных и полуофициальных СМИ и разделили его по типам; исследовали и выделили признаки; исследовали с какой точностью признаки автоматически распознают типы.

Мы полагаем, что конспирологические и аналитические статьи могут отличаться от информативных такими статистическими признаками как среднее количество слов в предложении, количество слов в документе, количество предложений, среднее количество абзацев. Новостные статьи, в свою очередь, более информативные, содержат больше цифр и именованных сущностей. Также мы предполагали, что кавычки, как признак прямой речи, чаще всего встречаются в аналитических и конспирологических текстах. Лексические признаки включают частоту встречаемости глаголов, частоту встречаемости прилагательных и существительных в документе. Мы ожидали, что большое количество глаголов в тексте будет ассоциироваться с более информативными новостными текстами. Ранее в своих работах мы предполагали, что существительные во множественном числе и наличие слов-квантификаторов как *все, каждый, любой, всегда, вечно, никогда, постоянно, никто, ничего* – это признаки генерализованных текстов и было интересно анализировать их распределение в жанрах. Окончательно было отобрано двенадцать стилистических и лексических признаков.

Затем мы экспериментировали с разными вариациями признаков, чтобы посмотреть какая совокупность признаков показывает наиболее высокую точность, мы тренируем модель для всех совокупностей признаков. Совокупность четырех признаков: количество кавычек, цифр и чисел, среднее количество слов в предложении, частота встречаемости глаголов и прилагательных показала самую высокую точность 85%. Для совокупности всех признаков алгоритм показал 67%. Таким образом совокупность четырех признаков в целом дает более высокую точность распознавания, чем один признак или все признаки.

Дополнительные эксперименты по распознаванию источников текстов показали, что алгоритм распознает новостные статьи с точностью 88.5%, конспирологические статьи с результатом 100%, бизнес статьи с результатом 67.7% и научно-популярные с результатом 75%. Более высокие результаты автоматического распознавания новостных и конспирологических статей могут быть результатом того, что имеющиеся в них синтаксические признаки более выражены, на наш взгляд. Тогда как научно-популярные и бизнес статьи различаются больше по лексическому составу. Это требует дополнительного изучения и ввода новых признаков, например, как частота встречаемости уникальных слов.

## Вероятная деструктивность

В рамках данного проекта было проведен анализ методов, обеспечивающих выявление деструктивных сообщений в интернет-среде [74, 75]. Показано, что для их выявления применяются подходы, основанные на анализе синтаксических закономерностей; анализе семантической информации, заключенной в тексте, и соотнесении ее с текстовым корпусом; методах краудсорсинга; выявлении закономерностей поведения пользователей в социальных сетях; рассмотрении дополнительной информации и т.п. Хорошие результаты могут быть достигнуты при некоторых условиях: наличии доступа к трафику социальных сетей и других новостных интернет-ресурсов, возможности организовать или получить результаты краудсорсинга и т.п. При ограничениях на выполнение перечисленных условий выявление деструктивных сообщений можно выполнять, опираясь на косвенные признаки. Приведены результаты выявления деструктивных новостей в корпусе новостных сообщений, размещенных на сайтах Республики Казахстан.

Приведены результаты выявления деструктивных новостей в корпусе новостных сообщений на русском языке, размещенных на сайтах Республики Казахстан с использованием заранее составленных словарей, содержащих слова, указывающие на следующие свойства текста: наличие проверяемых фактов, политизированность, призыв к действию, негативная или позитивная тональность, манипулятивность.

Результаты применения алгоритма показывают, что в некоторых случаях он дает хорошую степень совпадения с экспертными оценками. С другой стороны, он позволяет ранжировать тексты и в ситуации, когда тексты весьма близки по своим характеристикам. Приведенные выше тексты взяты из одного достаточно надежного источника, поэтому достоверность и объективность текстов близки. Вместе с тем данный подход требует весьма тщательного составления словарей и применения дополнительных алгоритмов для определения более широкого круга параметров текстов.

## Оценка выраженности в анализируемом тексте информативных критериев на основе словарного подхода

Одним из методов оценки выраженности в анализируемом тексте информативных критериев (признаков) является словарный подход [50].

Подготовительным этапом, обусловливающим решение задач по автоматизации оценки новостных текстов и распознаванию информативных признаков, является комплекс работ по созданию словарного массива, отбору языковых единиц и способов (правил) их употребления в новостных текстах. В связи с этим целью данной работы является разработка методов формализации правил и словарей для распознавания в корпусе новостных текстов информативных признаков и определения степени их выраженности.

Были определены 12 информативных признаков, для распознавания которых необходимо формирование словарей и формализация лингвистических шаблонов и правил.

Объектом анализа выступили 2 000 текстов из 5 информационных новостных интернет-источников: «KazakhSTAN 2.0», «Central Asia Monitor», «Zakon.kz», «Радио Азаттык», «Tengrinews.kz».

Для достижения цели в текущем году были поставлены и решены следующие задачи:

1. Проведение глубинного анализа сформированного корпуса публикаций СМИ по 9 информативным признакам (3 признака в 2018 г., 9 – в 2019 г.).
2. Осуществление разметки текстового корпуса по обозначенным информативным признакам.
3. Формирование и систематизация словарей по определенным информативным признакам на базе анализируемых текстов.
4. Выявление закономерностей для формализации правил распознавания отдельных информативных признаков.
5. Получение экспертных заключений о реализуемости выявленных закономерностей в рамках проекта.
6. Описание методов формализации правил на основе лингвистического анализа текстов.

Таким образом, на основе сводного анализа и коррекции результатов работы группы лингвистов по исследованию и детальной разметке 2000 медиа-текстов, получены следующие результаты:

1. Сформирован словарь из 8961 языковых единиц (3091 слов и 5870 словосочетаний) по информативному признаку «тональность».
2. Сформирован словарь из 166 языковых единиц (30 слов и 136 словосочетаний) по информативному признаку «политизированность».
3. Сформирован словарь из 1543 языковых единиц (652 слова и 891 словосочетание) по информативному признаку «наличие манипулятивных техник».
4. Сформирован словарь из 2639 языковых единиц (по 6 параметрам) по информативному признаку «социальная значимость».
5. Сформирован словарь из 1398 слов и 5501 словосочетания по информативному признаку «казахстанское содержание».
6. Сформирован словарь компетентных источников (именованные сущности: организации, должности и персоналии), состоящий из 2378 слов и словосочетаний, 88 слов-связок, используемых при ссылке на источник.
7. Сформирован словарь из 1202 слов и 4221 словосочетания в рамках информативного признака «наличие проверяемых фактов в разрезе субъект/предмет – действие – место – время».
8. Сформирован словарь из 4001 слов и 6409 словосочетаний по информативному признаку «ключевые слова» в текстах.
9. Сформирован словарь по информативному признаку «предвзятость» – 32 слова и 423 словосочетания.
10. Сформирован словарь по информативному признаку «призыв к действию» – 89 слов и 1427 словосочетаний.
11. Выявлены закономерности для формализации правил по определению информативных признаков тональности, политизированности, наличия манипулятивных техник, ссылок на компетентные источники, наличия проверяемых фактов в разрезе «субъект/предмет – действие – место – время», соответствия заголовка публикации содержанию, предвзятости и призыва к действию.

Была произведена чистка, систематизация и внесены дополнения в вышеуказанные словари.

Также в ходе анализа корпуса была произведена разметка текстов с выделением текстовых фрагментов, включающих маркеры информативных признаков.

По результатам закономерностей, выявленных в ходе лингвистического анализа корпуса, были получены экспертные заключения об их реализуемости в рамках проекта от российских и казахстанских экспертов.

Также были описаны методы формализации правил на основе закономерностей, выявленных в ходе лингвистического анализа корпуса.

Полученные результаты исследования в целом подтвердили необходимость и возможность выявления заданных информативных признаков на языковом уровне для автоматизированного распознавания искомых параметров текста.

В число однозначно определяемых с помощью лингвистических средств информативных признаков вошли тональность, социальная значимость, казахстанское содержание, ссылки на компетентные источники и призыв к действию.

В то же время, как показали результаты работы данного этапа исследования, часть информативных признаков, отобранная для оценки новостных текстов, должна определяться и распознаваться с применением комбинированных средств – лингвистических способов и машинных алгоритмов. В число таких признаков входит политизированность, наличие манипулятивных техник, ключевые слова, предвзятость. В рамках этой работы необходимо искусственное моделирование конструкций, построенных по лингвистическим правилам и шаблонам на базе моделей и словоупотреблений, определенных в ходе глубинного анализа новостных текстов. Например, анализ слова развитие в сочетании с другими словами, анализ частиц не в сочетании с глаголами с позитивной и негативной тональностью, с прилагательными, местоимениями, существительными, сочетание не с другими частями речи, повтор и т.д.

В целях более эффективной работы по определению наличия казахстанского содержания текста необходимо включение в материал для автоматизированного распознавания, помимо извлеченных из 2000 текстов казахстанских персон, топонимов, предметов обихода и т.д., в силу ограниченности объемом проанализированного материала универсального перечня наименований географических объектов, известных персон, должностных лиц, наименований казахстанских государственных органов, учреждений, организаций и бизнес-структур. Подобные словари в разы повысят «узнаваемость» казахстанского содержания в тексте. Результаты проведенных исследований были опубликованы в [76, 77, 78].

# ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ

## Назначение, цели и задачи разрабатываемой системы

Система оценки влияния открытых текстовых информационных источников на социум или Система анализа информационных трендов (NLPMonitor) предназначена для сбора, классификации, анализа текстовых публикаций казахстанского сегмента электронных средств массовой информации для принятия социально-значимых управленческих решений в стимулировании устойчивого развития личности ведомственными государственными организациями, такими как Министерство образования и науки, Министерство труда и социальной защиты и т.п., и/ или коммерческими организациями.

Цель создания системы: Разработать систему для стимулирования устойчивого развития личности, которая в режиме реального времени показывает пользователю актуальные социально-значимые темы в заданной социальной сфере или области с оценкой влияния открытых текстовых информационных источников на социум с использованием технологий больших данных.

Критерии оценки влияния открытых текстовых информационных источников на социум:

* степень позитивной направленности публикации;
* степень негативной направленности публикации;
* степень генерализации;
* степень политизированности;
* степень субъективности (степень выраженности мнения автора по предмету публикации);
* степень ангажированности (предвзятости, пристрастности);
* степень актуальности (степень актуальности информации на сегодняшний день);
* степень оригинальности (уникальности);
* степень привлекательности публикации (насколько публикация может способна привлечь к себе внимание потенциальных читателей);
* степень казахстанского содержания;
* степень провокационности (манипулятивности);
* степень связанности с другими публикациями по обсуждаемым вопросам;
* степень противоречия другим публикациям.

Задачи системы:

1. Сбор, хранение и отбор уникальных публикаций с интернет пространства в базу данных системы
2. Распределение публикаций по темам: кластеризация, классификация, определение тематических комбинаций, ранжирование и фильтрация (по социальным сферам, областям, отраслям и т.п.)
3. Определение информационных поводов
4. Вычисление степеней информативных признаков публикации, таких как: достоверность, резонансность, тональность, объективность, медиа-вовлеченность
5. Определение информационных трендов
6. Формирование всевозможных аналитических графиков, инфографиков, статистик для дальнейшего принятия управленческих решений.

## Архитектура информационной системы

В рамках поставленных задач была разработана архитектура информационной системы (Рисунок 5) для мониторинга, сбора, хранения, обработки, визуализации и анализа больших объёмов текстовых данных [80], в частности данных из медиа-пространства Казахстана NLPMonitor, а также программная реализация разработанной архитектуры, развёрнутая на имеющемся аппаратном обеспечении и запущенная в эксплуатацию.

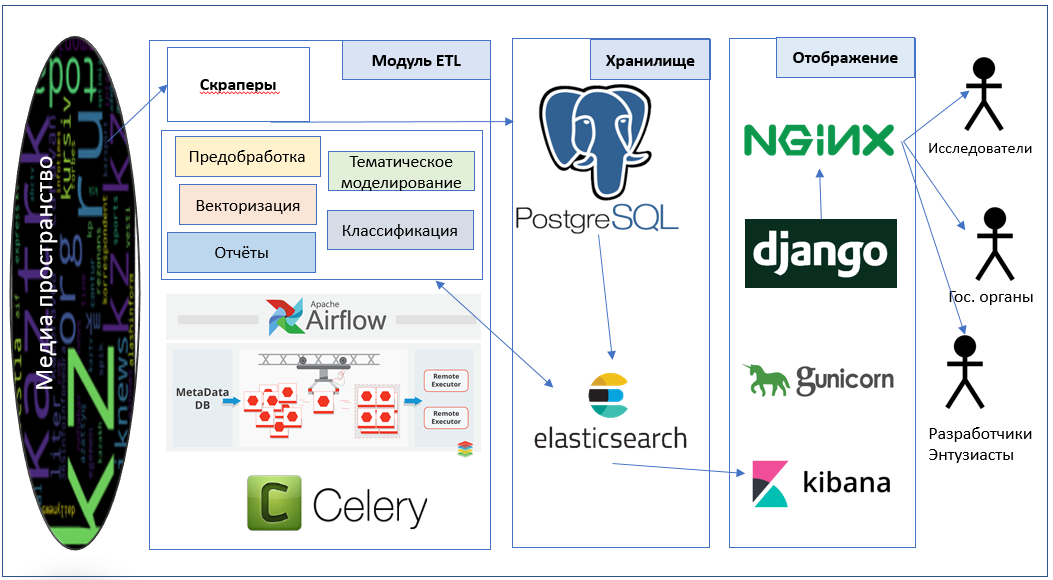


Рисунок 5 – Архитектура информационной системы NLPMonitor

Ввиду необходимости хранения и обработки больших объёмов данных, а также применения ресурсозатратных алгоритмов машииного (в том числе глубокого) обучения, архитектура была разработана с учётом потребности в масштабируемости и распределённости. Рассмотрим основные компоненты данной системы, показанные на рисунке 5:

* *Хранилище*

В системе предусмотрено три вида хранилищ:

1. *PostgreSQL* – выполняет роль персистентного хранилища для структурированных данных. Основные типы данных, хранящиеся в этой базе:
   1. Новости и метаданные;
   2. Обработанные данные на уровне разных базовых единиц анализа (токен/слово/фраза/предложение/текст), в том числе векторизации, результаты лемматизации, очистки и пр.;
   3. Результаты проведения тематического моделирования;
   4. Результаты классификации новостей по различным признакам (тональность, политизированность, социальная значимость и пр.) [49].
2. *ElasticSearch* – in-memory NoSQL хранилище, предназначенное для хранения неструктурированных или слабоструктурированных данных, а также быстрого поиска (в том числе полнотекстового), фильтрации и потокового доступа. ElasticSearch выполняет несколько функций:
   1. Основное хранилище для доступа, поиска и фильтрации данных конечным пользователем;
   2. Основное хранилище для ETL (Extract-Transform-Load) процессов обработки данных – в том числе запись любых промежуточных результатов в свободной форме;
   3. Хранилище для кеширования определённых результатов вычислений, необходимых для построения дашбордов и отчётов в системе.

ElasticSearch дублирует данные, хранящиеся в PostgreSQL как персистентном хранилище, поскольку ElasticSearch является in-memory базой данных без гарантий относительно персистентности и целостности данных.

1. *Redis* – быстрое key-value хранилище, используемое для кеширования отдельных страниц и элементов, а также для кеширования сессий авторизации. В Redis хранятся служебные данные, а также кеш страниц и элементов, к которым происходит частый доступ.

Все три основных хранилища системы могут быть легко масштабированы на несколько отдельных машин, поддерживается как шардирование, так и репликация, при этом ElasticSearch и Redis показывают близкое к линейному увеличение производительности при горизонтальном масштабировании.

* *Обработка данных*

При разработке архитектуры для обработки данных (ETL) были выявлены следующие основные потребности:

1. Возможность распараллеливания вычислений, в том числе на нескольких машинах;
2. Возможность гибкого планирования выполнения различных задач по обработке данных;
3. Возможность мониторинга выполнения задач в реальном времени, в том числе оперативное информирование об исключениях;
4. Гибкость в используемых инструментах и технологиях.

В ходе анализа для удовлетворения всех этих потребностей была выбрана программная платформа с открытым исходным кодом Apache Airflow. Рассмотрим основные компоненты данной платформы:

1. *Airflow-worker* – основной компонент, выполняющий обработку данных. Может быть горизонтально масштабирован, в том числе на отдельные сервера/облачные VM. В текущем варианте архитектуры в образ контейнера Airflow-worker заранее встраиваются необходимые зависимости, однако принципиально процесс инъекции зависимостей может происходить различным образом, в том числе путём динамичного получения Docker контейнеров из публичных либо приватных репозиториев.
2. *Airflow-scheduler* – компонент, отвечающий за назначение задач Airflow-workerам в порядке, определённом Airflow DAGами. Airflow DAG – нецикличный направленный граф, описывающий порядок выполнения определённых задач, а также содержащий информацию о расписании, приоритетах, поведении в случае исключений и пр.
3. *Airflow web server* – веб-интерфейс, позволяющий отслеживать и контролировать ход выполнения задач.

Для хранения служебных данных, таких как состояния выполнения задач, используется отдельный кластер PostgreSQL. Для запуска и отслеживания прогресса задач используется связка Celery+Redis.

Основной компонент для выполнения задач в Airflow – DAG – Directed Acyclic Graph – направленный нецикличный граф, описывающий зависимости, порядок выполнения, расписание и приоритет различных задач. В отличие от аналогичных систем, Airflow DAGи полностью описываются на стандартном языке Python, что позволяет, в том числе, создавать динамические задачи и группы задач, например для распараллеливания вычислений, мета-оптимизации и задач, требующих корректировку структуры выполнения в зависимости от внешних данных (например скрапинг). Для добавления DAGов на боевой сервер, члены исследовательской могут совершить push в отдельны git репозиторий (<https://github.com/KindYAK/NLPMonitor-DAGs>), после чего на основном сервере произойдёт автоматическое получение нового кода, с возможностью ручного, либо автоматического запуска по расписанию.

* *Интерфейс*

Интерфейс системы – стандартный HTML+CSS+JS веб-сайт с доступом по протоколу HTTP. Веб-приложение реализовано на Python фреймворке Django, в качестве веб-сервера выступает Gunicorn, реверс-прокси Nginx. Веб-приложение имеет доступ как к персистентному хранилищу PostgreSQL, так и к ElasticSearch. Веб-приложение реализует ряд страниц для фильтрации, поиска, доступа к различным дашбордам и отчётам.

* *Развёртка (deploy) системы*

NLPMonitor реализован в виде модульной системы, основанной на принципе контейнеризации. Контейнеризация (виртуализация на уровне ОС) в отличии от классической виртуализации (на уровне аппаратного обеспечения) позволяет обеспечить изоляцию и масштабируемость отдельных компонентов системы при минимальных потерях в производительности и возможности полноценного использования всех аппаратных элементов системы.

NLPMonitor разработан на базе Docker CE Linux контейнеров, локальная развёртка для целей разработки происходит с помощью docker-compose, на основном сервере используется Docker stack. В официальном репозитории системы (<https://github.com/KindYAK/NLPMonitor>) в папке configs расположены файлы конфигурации для развёртки с помощью docker-compose/docker stack, а в папках airflow-worker, visartm и web/docker расположены Dockerfile и файлы необходимые для сборки кастомных контейнеров. На данный момент все версии образы собранных кастомных контейнеров хранятся в развёрнутом на отдельном сервере приватном Docker Registry.

Одним из преимуществ контейнеризации является возможность простого масштабирования элементов системы, в том числе шардирования хранилищ. Также планируется применение оркестратора контейнеров Kubernetes для реализации возможности простого масштабирования системы на вычислительный кластер с автоматической балансировкой нагрузки между серверами в кластере.

Проведена работа по формированию базы данных статей СМИ (корпус). Разработаны и реализованы алгоритмы по автоматическому сбору статей с 54 новостных сайтов казахстанского сегмента СМИ. Проведена работа по агрегации и последующей обработке собранных статей. В результате собрано и обработано 2 387 974 уникальных статей. Сформирован новостной корпус казахстанского сегмента СМИ с помощью специально разработанных ПО. Всего было использовано 54 агентства (новостных сайта), из них: новостные агентства – 29, печатные – 15, телевидение - 8 и интернет – 2.

Всего собрано статей: уникальных по сайтам – 2 388 534, уникальных по всему корпусу – 2 387 974. Объём данных составил 9.5 Гбайт.

## Определение необходимой и достаточной функциональности разрабатываемой информационной системы для качественного решения поставленных задач проекта

При проектировании информационной системы была определена необходимая и достаточная функциональность разрабатываемой информационной системы для качественного и достаточного решения поставленных задач проекта. Хотя здесь и определяется необходимая и достаточная функциональность, но в системе всегда будет присутствовать «избыточная» функциональность так как для добавления новой и расширения существующей функциональности необходимо проводить апробацию различных методов/алгоритмов и их сравнение по степени эффективности, а уже затем имплементировать наиболее эффективные из них тем самым сокращая избыточность функциональности системы.

Далее предлагается определённый перечень необходимого и достаточного функционала информационной системы для качественного решения имеющихся задач проекта (в дальнейшем его можно расширять, но важно именно определить функционал, который с наименьшими затратами решает поставленные в проекте задачи):

1. Автоматических сбор (скрапинг) публикаций из открытых текстовых информационных источников в сети Интернет и формирования на их сонове анализируемого корпуса. Скрапинг должен осуществляться по заранее определённому списку источников.
2. Формирование словаря базовых единиц анализа (БЕА), который может включать: отдельные слова, устойчивые словосочетания, именованные сущности и не должен включать малоинформативную лексику (предлоги, вводные слова и их сочетания и др.). С целью оптимизации в словарь БЕА можно фильтровать по лексическим и морфологическим признакам (например, включать в словарь только определённые части речи; добавлять/удалять именованные сущности, вводить ограничения на частотность включаемых элементов и т.д.).
3. Классический Topic Modeling на основе сформированного ранее словаря БЕА.
4. Формирование словаря тематических комбинаций (СТК): Комбинаторный анализ взаимной сочетаемости топиков в публикациях анализируемого корпуса. Так как каждая отдельная публикация в корпусе может содержать несколько топиков, то можно проанализировать какие бывают сочетания топиков. Каждому отдельному топику присваивается уникальный идентификатор. По результатам анализа получаем список возможных комбинаций идентификаторов с указанием количества публикаций соответствующих данному сочетанию топиков.
5. Разбить сформированный СТК по прикладным классам. Например, в качестве прикладных классов можно выбрать следующие:

* социальная значимость;
* стимулирование устойчивого развития личности;
* социальная напряженность;
* политика;
* казахстанское содержание;
* образование;
* развитие, наука, инновации, технологии;
* кризисные явления и происшествия;
* позитив;
* негатив;
* и др.

Разбиение по прикладным классам может осуществляться экспертно, на основе предобученных моделей либо по словарям. Также при разбиении можно указывать/вычислять вес данной тематической комбинации в данном прикладном классе.

1. Анализ имеющегося корпуса в разрезе прикладных классов:

* объём публикаций соответствующих разным прикладным классам и динамика его изменения во времени в разрезе всего корпуса
* объём публикаций соответствующих разным прикладным классам и динамика его изменения во времени в разрезе конкретного СМИ (источника)
* ранжирование СМИ по мере соответствия различным прикладным классам
* ранжирование публикаций по мере соответствия различным прикладным классам
* кластеризация публикаций по мере соответствия [прикладным классам]/[тематическим комбинациям]
* выявление аномалий (нестандартного сочетания тематик/классов или нестандартных сочетаний и динамик анализируемых признаков)
* выявление тематик, их сочетаний и классов находящихся в тренде, набирающих популярность.

# МОНИТОРИНГ И КОРРЕКТИРОВКА ОТКЛОНЕНИЙ В РАБОТЕ РАЗРАБОТАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ

Мониторинг за состоянием вычислительных процессов происходит стандартными методами web-административной панели Airflow, а также с помощью системы Sentry.io, интегрированной во все компоненты системы, включая Airflow и веб-интерфейс системы. Sentry.io позволяет автоматизировать сбор информации об ошибках и исключениях, включая данные логгирования и информативные дампы памяти на момент ошибки. Возможно отслеживание статистики по частоте возникновения тех или иных ошибок, а также гибкий поиск и фильтрация. Также реализованы автоматическая система оповещения об исключительных ситуациях на электронные адреса ответственных лиц.

Проведен мониторинг и анализ разработанной информационной системы, на основе которых рекомендовано рассмотреть возможности написания собственных кодов для использования дэшбордов, пользовательского интерфейса и для наличия возможности конфигурации функционала под специфику проекта; включения разработанных в ходе проекта алгоритмов тематического моделирования LDA, машинного обучения распознавания тональности, определения полярности тональности и степени объективности новостных текстов лексико-словарным подходом и создания собственной библиотеки с размещением сформированных в рамках исследования словарей и экспериментальных корпусов.

Дальнейшие направления развития системы NLPMonitor

Основными направлениями для дальнейшего развития системы являются:

1. Разработка кластера на базе Kubernetes для автоматической балансировки нагрузки между несколькими физическими серверами;
2. Полная интеграция CUDA-совместимых вычислений в большую часть алгоритмов и вычислительных моделей;
3. Разработка продвинутого пользовательского интерфейса для доступа и визуализации информации;
4. Разработка системы распознавания аномалий и новизны (novelty detection) с автоматическим оповещением соответствующих ответственных лиц;
5. Автоматическое формирование аналитических отчётов и рекомендации для поддержки принятия управленческих решений.

# СОЗДАНИЕ НЕОБХОДИМЫХ ТЕХНИЧЕСКИХ И ЭКСПЕРТНО-АНАЛИТИЧЕСКИХ УСЛОВИЙ ДЛЯ РАЗРАБОТКИ ИНФОРМАЦИОННОЙ СИСТЕМЫ ОЦЕНКИ ВЛИЯНИЯ ОТКРЫТЫХ ТЕКСТОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ НА СОЦИУМ

## Методы кластерного анализа и их применение при анализе корпуса публикаций СМИ

Так как в рамках данного проекта решаются задачи, связанные с обработкой нескольких миллионов текстовых документов – публикаций в СМИ, то становится очевидным что наиболее целесообразным подходом является групповой анализ публикаций. Так как анализировать/классифицировать каждую публикацию в отдельности в разрезе множества информативных признаков особенно для экспертов является очень трудозатратной задачей. Напротив, при групповом анализе публикации автоматическими методами, с помощью алгоритмов кластеризации разбиваются на группы/кластеры по мере близости наборов своих информативных признаков. Таких кластеров-групп получается значительно меньше нежели исходных объектов – публикаций. Поэтому их групповой анализ уже можно экспертно проанализировать, проинтерпретировать и визуализировать в информационной системе.

В рамках данного проекта алгоритмы кластеризации применяются для решения следующих задач:

1. Кластеризация публикаций СМИ по инфоповодам, тематикам, темам, источникам, действующим объектам/субъектам и др. [83, 84].
2. Кластеризация базовых единиц анализа (слов, устойчивых словосочетаний, n-грамм и др.) на основе их векторных представлений, полученных методами дистрибутивной семантики [81, 82].
3. Агломеративная кластеризация тематик и их сочетаний с целью построения тематических онтологий.
4. Кластеризация СМИ по их агрегированным векторным представлениям.
5. Кластеризация индивидуумов (потребителей информации СМИ) по анализу их «вектора информационных предпочтений» и «вектора индивидуальных признаков».

В процессе кластерного анализа приходится сталкиваться необходимостью решения следующих основных проблем, диктуемых спецификой предметной области:

* + 1. Сбор кластеризируемых данных;
    2. Предобработка, нормализация и параметризация в виде вектора численных признаков (различные прикладные задачи требуют применения различного состава информативных признаков);
    3. Определение оптимального состава вектора информативных признаков;
    4. Выбор или разработка алгоритма, обладающего требуемыми характеристиками;
    5. Выработка критериев численной оценки качества результатов кластеризации.

Вариабельность вышеперечисленных пунктов напрямую связана со спецификой решаемой прикладной задачи.

В настоящее время в мире существует огромное количество разнообразных новостных медиа ресурсов. Со временем их количество только растет. Распространение новых технологий обеспечивают большую доступность информации, упрощение в ее распространении и в ее производстве. Каждый день медиа ресурсы производят громадный объем новостной информации, касающейся различных происходящих или произошедших в мире событий, высказанных мнений, идей. Они могут относится к различным вещям или материям в зависимости от сферы интереса, специализации или редакторской политики ресурса.

Зачастую пользователю требуется найти определенную информацию касательно интересуемого информационного повода, для этого он обращается к новостным медиа ресурсам, где в новостных публикациях пытается найти необходимое. Ему на основе найденного требуется выработать объективный взгляд, мнение, свою точку зрения по интересующему поводу. С целью этого для полного понимания, получения полной всесторонней картины, пользователю необходимо найти публикации от как можно большего числа медиа ресурсов, в которых могут быть сформированы различные мнения, точки зрения касательно интересующего информационного повода. При получении информации от какого-то одного ресурса или связанной группы ресурсов можно приобрести однобокое смещенное представление. Но и возможно также подвергнуться, при неэтической политике медиа ресурса, манипулированию c его стороны и навязыванию какой-то однобокой выгодной ему точки зрения.

При обращении к поисковому запросу, то даже при успешном его формулировании, поисковая система зачастую выдает большое количество информации, в том числе информацию не относящуюся (несвязанную) либо косвенную к интересуемой информации. Пользователю в этом объеме информации легко затеряться, ему требуется приложить большой объем усилий для переработки результатов выдачи, подобная работа требует большого количества времени и сил. Для более эффективной работы пользователю было бы удобно под рукой иметь интеллектуальную систему, которая автоматически бы по указанному информационному поводу (по выделению определенной публикации с интересуемым информационным поводом) находила бы в пространстве Интернета напечатанные публикации с тем же информационным поводом, произведенных различными медиа ресурсами.

## Плотное векторное представление текстов на основе анализа взаимной встречаемости слов и тематического моделирования в условиях BigData

В рамках данного проекта разработан метод плотной тематической векторизации текстов, содержащихся в большом в текстовом корпусе (BigData). В основе данного подхода лежит использование векторного представления слов на основе расчета матрицы их взаимной встречаемости в текстах с последующим применением к ним алгоритма нечеткой кластеризации для получения групп тематически близких слов. Полученные группы, построенные на основе “больших данных”, далее уже используются для получения плотного векторного представления текстов. Данный подход рассматривался в отношении новостных публикаций СМИ. Также в статье продемонстрировано, что подход практически реализуем в отношении “больших данных”.

У данного подхода одновременно три преимущества, которыми обладают вышеуказанные типы. Первое, он предоставляет текст в виде плотного вектора, что удобно в вычислительных целях. Второе, он векторно может одинаково представлять тематически близкие тексты даже при использовании в них слов различных по словоформе, при условии, что они тематически связанные. Третье, результаты векторного представления являются интерпретируемыми.

При применении алгоритмов кластеризации к разработанному векторному представлению, в случае если векторное представление тестовой информации присваивает схожим по тематике или смыслу текстам близкие (по некоторой метрике) вектора, а отличающимся текстам присваивает различные векторные представления в зависимости от степени различия, получим кластера из текстов, которые будут схожими по тематике и может быть даже по смыслу. Проведённые эксперименты показали, что разработанное векторное представление обладает указанным свойством.

*Первый этап:* составление словаря

Первоначальная задача стоит в определении всех слов, из которых составлены тексты корпуса. Для этого формируется словарь слов-имен существительных (включая именные сущности), используемых в рассматриваемых текстах. В словарь включаются все слова в текстах за исключением высокочастотных слов, служащим только функциональным целям, так называемых stopwords. Также исключаются из рассмотрения очень редкие слова. В итоге полученное множество слов будет представлять из себя сформированный словарь слов (целевых единиц анализа).

*Второй этап:* группировка слов словаря согласно их тематической связанности

На данном этапе требуется сгруппировать все слова словаря согласно их тематической связанности. Тема в словаре Ожегова определяется как предмет, основное содержание рассуждения, изложения, творчества и т.п. Здесь же под тематикой понимается совокупность сочетаний слов, которые либо используются часто в одном и том же контексте, либо используются постоянно в одних и тех же текстах с повторяющейся частотой совместного употребления. Группировка слов по тематикам согласно приведенному определению, позволит, как будет показано дальше, “сжать” векторное представление и также представить тематически связанные тексты схожим образом. Поэтому идея состоит в том, чтобы первоначально определить группы слов, которые тематически или возможно даже семантически близки.

Для группировки слов по тематике первоначально формируется матрица взаимной встречаемости слов в текстах корпуса, размерностью равной числу текстов на длину словаря. Элементы данной матрицы равны либо частоте встречаемости соответствующих слов в текстах, либо их значениям по статистической мере *tf-idf,* при этом считается, что два слова встречаются, если они использованы в одном и том же тексте в независимости от места расположения в нем по отношению друг к другу, то есть текст рассматривается как мешок слов. Для получения векторного представления слов словаря транспонированную матрицуумножим на нее же саму.

(3)

В результате получим квадратную симметричную матрицус размерностью равной длине словаря на длину словаря, где каждая строка или столбец есть векторное представление слова. Данное векторное представление слова характеризует контекстную связанность данного слова с другими словами словаря в отношении встречаемости в текстах корпуса. Векторные представление слов, имеющих одинаковый уровень связанности с другими словами и тем самым, с большей вероятностью относящихся к одной тематике, будут схожими и значение косинусной близости у них будет выше по сравнению с другими. Таким образом, рассчитав матрицу косинусных расстояний векторов слов (строки матрицы ) словаря по формуле,

(4)

можем применить к ней алгоритм последующей кластеризации. Так как одно слово потенциально может относиться к нескольким различным тематикам, то необходимо применять алгоритм “нечеткой” кластеризации, где в результате каждый элемент может относиться не только к одному кластеру, как в случае применения алгоритма “четкой” кластеризации. В результате применения алгоритма “нечеткой” кластеризации получим кластера слов, которые будут связаны по тематике и при этом слова в зависимости от значения, в котором употребляются, попадут в соответствующую группу, а не исключительно только в одну.

*Алгоритм “нечеткой” кластеризации*

В качестве алгоритма “нечеткой” кластеризации выбран алгоритм, у которого минимальное число параметров, который строит кластера, основываясь на расстояниях между элементами, который не требует задания итогового числа кластеров в виде параметра, так как заранее их число неизвестно, который также предполагает возможность реализации процесса работы алгоритма на основе параллельных вычислений и обеспечивает относительно быструю работу алгоритма по отношению при работе с большими данными. Использован алгоритм кластеризации, описанный в разделе 2.2.3 «Алгоритм тематического моделирования на основе кластерного подхода». Нечёткость алгоритма заключается в том, что он может отнести один объект в несколько разных кластеров.

*Третий этап:* агрегирование полученных кластеров для получения тематических категорий

Так как число полученных на втором этапе кластеров в может быть очень высоким (может превышать размер словаря), то для того, чтобы обеспечить в дальнейшем плотное векторное представления требуется произвести объединение тематически похожих кластеров слов для получения векторов с размерностью не превосходящую выбранного порога. В дальнейшем полученные кластера слов будем называть тематическими категориями и их число будет определять размерность векторного представления (оно будет в точности ему равняться).

*Четвертый этап:* формирование векторного представления текстовой информации

Формироваться векторное представление текстовой информации будет через определение степени взаимосвязи текста с каждой выявленной на третьем этапе тематической категорией. Те тексты, которые являются схожими по тематике будут обладать схожим уровнем взаимосвязанности с каждой отдельной категорией. Степень взаимосвязи текста с каждой тематической категорией можно рассчитать через модифицированную меру Жаккара.

(5)

Другой вариант модифицированной меры Жаккара, которая учитывает значения *tf-idf* слов публикации, представлен формулой:

(6)

Формула (6) в случае, если у текста и категории общее слово, добавляет в числителе не просто единицу, а умножает ее на значение *tf-idf* слова, тем самым в случаях, когда общее слово (для категории и текста) часто встречается во всем количестве текстов, то оно вносит меньший вклад по сравнению с общими словами, которые встречаются реже во всем количестве текстов, но относительно часто в рассматриваемом тексте.

Высокое значение меры Жаккара между категорией и текстом будет говорить о высоком уровне связанности тематики, определяемой категорией, и текстом. В результате получится число различных тематических взаимосвязей текста с категориями равным числу категорий и все полученные значения можно рассматривать как один вектор, представляющий данный текст. Таким образом, размерность полученного вектора будет равна числу тематических категорий, определенных на третьем этапе.

Была разработана методика оценки качества кластеризации новостных публикаций. Данная методика основана на сравнении результатов кластеризации с результатами тематической разметки корпуса выполненной экспертами.

Качество результатов кластеризации при использовании разработанного векторного представления намного превосходит результаты кластеризации, получаемые при случайном построении кластеров. Результаты кластеризации в сравнении с результатами, получаемыми при использовании векторов, основанных на моделях дистрибутивной семантики, и векторов с размерностью равной размеру словаря оказываются между ними, где лучший результат показывает последний вариант представления. Также замечено, что чем больше тематических категорий используются (тем самым размерность вектора выше) тем результаты в случае кластеризации при разработанном представлении становятся лучше.

В дальнейшем можно исследовать результаты использования разработанного векторного представления в других конкретных задачах области обработки текстов естественных языков и проанализировать получаемые результаты с теми, которые показывают на данный момент самые лучшие передовые подходы.

В настоящее время готовится к публикации статья, в которой будут подробно изложены описания разработанных алгоритмов, а также проведённые эксперименты, по сравнительной оценке, качества.

## Кластеризация новостных публикаций по информационным поводам

В рамках данного проекта разработан алгоритм группировки новостных публикаций в соответствии с информационными поводами в них содержащимися с помощью методов кластеризации в развёрнутом виде опубликованный в [82]. Применение данного подхода может позволить автоматически предоставлять по интересующему информационному поводу целую группу публикаций, относящиеся к различным новостным медиа ресурсам и к данному информационному поводу. В зависимости от характера информационного повода предоставленная группа публикаций может содержать публикации, где встречаются различные точки зрения в отношении данного информационного повода, в том числе отличную точку зрения или обладать более, или менее развернутой информацией или с дополняющей, или корректирующей информацией.

Разработка данного подхода включала в себя три этапа: первый этап, разработка/выбор подходящего для целей задачи представления новостных публикаций, то есть определение подходящих признаков, второй этап - разработка меры близости (или функции дистанции), позволяющей на основании представления количественно определять смысловое (согласно информационному поводу) сходство представлений новостных публикаций. Третий этап – выбор/разработка алгоритма кластеризации для группировки публикаций в кластеры. В данном подходе ключевое место занимает разработанная функция дистанции, которая при определенном сочетаний представлений текста и заголовков публикации, обеспечивает группировку публикаций.

Рассматриваемые/разработанные варианты подходов аппробировались на подготовленном экспертами корпусе, состоящем из 822 новостей на предмет изучения их работоспособности, эффективности и выявления оптимальных. Была получена оценка времени работы подходов, проведен анализ полученных результатов и сравнение подходов.

По результатам аппробации различных подходов к размеченному корпусу был выбран подход, показавший наилучшие результаты. Ключевое новшество данного подхода - разработанная функция расстояния по определению семантической близости текстовой информации. Функция расстояния состоит их двух компонент: первая - мера близости Жаккара в отношении заголовков, которые представляются как множество слов, вторая – определяет близость текстов. Вторая компонента может быть в двух вариантах, оба показали близкие результаты. Первый и второй варианты используют сжатые вектора (основанные на моделях дистрибутивной семантики) всех слов, при этом в первом варианте используется функция расстояния WMD (Word Mover’s Distance), а в втором варианте сжатые вектора складываются для получения средневзвешенных векторов и между ними рассчитывается евклидово расстояние.

Как было представлено в опубликованной статье [83] разработанный подход по группировке новостных публикаций по информационным поводам успешно применен к размеченному экспертами корпусу новостных публикаций. В дальнейшем предполагается исследовать вопрос о применимости предложенного подхода к корпусу “больших данных’ как в отношении общей работоспособности подхода, так и в пригодности в вычислительном плане. Дополнительно также можно исследовать разработанный подход в плане большей оптимальности рассмотрев более совершенные или модифицированные алгоритмы кластеризации, другие меры близости, разработать другие функции расстояния для поиска оптимальной комбинации с целью улучшения результатов по выбранным критериям, рассмотреть алгоритмы по определению уникальных публикаций.

## Метод кластеризации новостных сообщений СМИ на основе их концептуального анализа

В рамках данного проекта разработан метод кластеризации новостных сообщений СМИ на основе их концептуального анализа опубликованный в [83].

В статье изложено решение задачи кластеризации сообщений СМИ на основе разработанной авторами методики автоматического вычисления меры смысловой значимости наименований понятий документов, использующей их статистические, синтаксические и семантические признаки, и технологий автоматического составления декларативных средств для кластеризации документов, базирующихся на методах их семантико-синтаксического и концептуального анализа.

При решении задачи кластеризации сообщений СМИ показана принципиальная возможность ее реализации на основе разработанной авторами методики автоматического вычисления меры смысловой значимости наименований понятий документов и технологий автоматического составления декларативных средств для кластеризации документов, базирующихся на методах их семантико-синтаксического и концептуального анализа.

На основе предложенной методики вычисления меры смысловой значимости наименований понятий и созданных в процессе проведения настоящего исследования программных и декларативных средств был проведен эксперимент по обработке представительного массива сообщений СМИ. Анализ полученных результатов показал, что при автоматическом установлении смысловой значимости текстовых наименований понятий использование семантических коррелирующих коэффициентов понятий повышает точность установления смысловой схожести между документами.

При этом на количественные характеристики результатов обработки текстов принятые ресурсные ограничения влияют незначительно.

## Алгоритм кластеризации больших данных высокой размерности на основе метода декомпозиции

В связи с тем, что в рамках реализации задач данного проекта мы столкнулись с необходимостью кластеризации «больших данных» высокой размерности нами был разработан и реализован соответствующий алгоритм, опубликованный в [81] и в [84].

Кластеризация крупномасштабных наборов данных (BigData) в настоящее время является актуальным предметом и обсуждается в литературе как c прикладной точки зрения, так и в теоретических аспектах. Одним из широко распространенных способов оценки качества кластеризации является критерий суммы квадратов расстояний (Sum of Squire Distance - SSD). В этом исследовании мы показываем, что можно достичь высококачественных результатов с точки зрения критерия SSD, применяя итеративную процедуру кластеризации к подмножествам входного набора данных. Также демонстрируем, что время кластеризации можно использовать как дополнительный параметр, который позволяет повысить качество кластеризации.

Алгоритмы и методы кластеризации в последнее время привлекают большое внимание как эффективные инструменты в теоретических и прикладных задачах машинного обучения, и которые позволяют обнаруживать закономерности в необработанных/ слабоструктурированных данных. Другая мотивация связана с необходимостью обработки больших наборов данных для получения естественной группировки данных. Следовательно, одним из основных аспектов для методов кластеризации является их масштабируемость.

*Существующие методы для решения задачи кластеризации больших данных*

Имеется ряд исследований, направленных на улучшение качества за счет стоимости и сложности времени [85, 86]. Этот тип методов обычно имеет существенный недостаток: практически невозможно кластеризовать средние и большие наборы данных (приблизительно 105 ~ 107 элементов и более). Эти методы не могут работать с огромными базами данных, потому что вычислительная сложность по времени / пространству(памяти) растет (полиноминально) очень быстро. Следовательно, имеет смысл искать алгоритмы, с хорошим согласованием между эффективной масштабируемостью и качеством кластеризации [87-90]. Одним из известных методов кластеризации данных является алгоритм *k*-средних, который широко используются благодаря своей простоте и хорошим характеристикам [91]. Ряд алгоритмов и технологий усовершенствовали этот метод при помощи кластеризации по частям [92]. Существуют алгоритмы, которые используют подход декомпозиции данных, например, mini batch *k*-means [93]. Взвешенная версия алгоритма *k*-средних дана с приложениями в [89]. Мета-эвристики могут существенно помочь тогда, когда точное решение сложно или дорого с точки зрения используемого машинного пространства / процессорного времени. Разработаны и реализованы некоторые эвристики для *k*-средних, которые ускоряют вычисления: Часть данных эвристик занимается ускорением сходимости метода, другая отбрасывает избыточные либо малозначимые промежуточные вычисления: Следующие примеры мета-эвристик показали свою эффективность при кластеризации больших данных:

* удаление на каждой итерации паттернов данных, которые маловероятно изменят свою принадлежность к тому или иному кластеру, как в [90];
* используя неравенство треугольника в [94];
* комбинации различных техник [95, 96, 97].

*Неформальная постановка задачи*

В данной работе мы используем модификацию *k*-means++ для построения алгоритмической мета-эвристики, которая на каждом шаге использует некоторые подмножествах из всего набора данных. Цель данной работы являлось создание метода декомпозиции для алгоритма *k*-средних на крупномасштабных наборах данных для инициализации центроидов, чтобы получить качественные результаты в отношении критериев MSSD (минимальная сумма квадратов расстояний). Другими словами, мы используем метод нахождения инициализации *k*-средних таким образом, чтобы он был близок к оптимальному при этом обладая высокой скоростью расчета. Используется мета-эвристика в задаче кластеризации *k*-средних путем обработки полученных данных во вторичной (высокоуровневой) процедуре кластеризации. Другой целью данной работы является исследования эффективности полученной эвристики и изучении поведения полученного алгоритма при изменениях мета-параметров алгоритма.

Идея нашего подхода в том, что сначала мы разделяем весь набор данных, случайным образом на подмножествах фиксированного размера, так называемых *пакетах* данных. Мы задействуем либо все объекты набора данных, либо некоторую их часть таким образом, чтобы используемые подмножества оставались представительными. Следующим шагом является кластеризация методом *k*-средних данных этих пакетов. Оценивая *SSD* по полученным кластерам, мы делаем эвристику для “хорошей” инициализации алгоритма для всего набора данных.

*Параметры модели:*

* *k* – заданное количество кластеров;
* *N* – количество объектов во всем наборе данных;
* *d* – размер одного пакета данных (количество объектов в одном пакете). В нашей работе размеры пакетов выбирались пропорционально всему набору данных. Например, разбиение на 5 пакетов означает, что используем пакеты длинной [N / 5] объектов.
* *n* – количество пакетов, используемых для независимой инициализации *k*-средних, фаза 1;
* *m* (≥*n*) - число пакетов, используемых для кластеризации в фазе 2. При этом объединение объектов всех *m* пакетов не обязано быть равным всему набору данных.

Мы рассмотрели следующие два способа для генерации пакетов:

1. Сегментация всего набора данных на пакеты: создается случайная перестановка объектов в используемом наборе данных. Набор данных сегментируется в последовательные пакеты размера *d*. Мы называем этот способ декомпозицией на унифицированные пакеты.
2. Для каждого пакета выбирается *d* случайных объектов из всего набора. Повторяя это, генерируется необходимое количество пакетов (в разных пакетах объекты могут повторяться, однако в каждом из пакетов все объекты уникальны). Мы называем это способом генерации случайных пакетов.

*Алгоритм кластеризации больших данных*

Алгоритм состоит из двух частей. В первой части (фаза 1) применяется алгоритм *k*-средних++ на каждом из *n* пакетов независимо. Схема алгоритма показана на рисунке 6. Цель данной фазы - найти усредненный набор центроидов, соответствующий пакетам 1, …, *n*.

Центроиды всех пакетов (*n* штук) объединяются в один новый набор данных. Каждому элементу из получившегося набора данных присваивается вес, соответствующий нормализованному значению *SSD* для тех пакетов, в котором он рассматривается как центроид. Вес *i*-го объекта рассчитывается следующим образом:

(7)

где *SSDi* - это значение *SSD* для пакета, из которого *i*-й центроид взят в качестве объекта. Затем с помощью *k*-средних новый набор данных делится на *k* кластеров с учетом весов объектов *wi*. В случае вырождения *k*- средних, т.е. когда алгоритм сходиться к разбиению, в котором присутствуют кластеры размера ≤ 1; в таком случае алгоритм инициализируется повторно. Полученные центроиды могут использоваться для:

1. Инициализации *k*-средних на всем наборе данных (качественное приближение).
2. Оценки *SSD* по всему набору данных.
3. Инициализации алгоритма, описанного в следующей части (фаза 2).

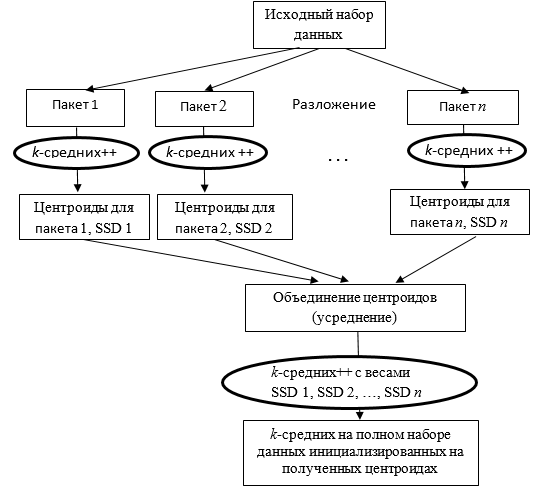


Рисунок 6 – Схема параллельной кластерной декомпозиции на пакетах .

Центроиды вычисленные на предыдущем этапе являются усреднением центроидов для случайно выбранных подмножеств исходного набора данных. При условии, что исходные пакеты были достаточно представительных размеров, данный набор центроидов может служить приближением к оптимальному значению кластерного разбиения и могут быть использованы, как указано в пунктах 1 и 2, для всего набора данных. Однако, если размеры пакетов, либо структура данных таковы, что получившиеся значение глобального *SSD* не дает явно выраженного минимума, то для этого случая предлагаем следующее улучшение алгоритма. Будем использовать полученные центроиды в качестве начальных приближений для *k*-средних на дополнительно сгенерированных пакетах. При обработке последующих пакетов *n* + 1, *n* + 2, ..., *m* (фаза 2) мы рассмотрели два варианта: параллельный (2.1) и последовательный (2.2).

*(2.1) Параллельный вариант*

Центроиды, полученные в предыдущей фазе 1, используются для инициализации *k*-средних в каждом последующем пакете *n* + 1, *n* + 2, ..., *m*. Получающиеся центроиды и SSD сохраняются, если нет дегенерации центроидов. Условием остановки является указанное ограничение либо по времени вычисления, либо по количеству обрабатываемых пакетов. Подобно фазе 1, мы делаем кластеризацию на объединенном наборе центроидов. Последующее использование его результатов аналогично пунктам 1 и 2 фазы 1.

*(2.2) Последовательный вариант*

Ниже приведена последовательная версия алгоритма. Данная версия схематически представлена на рисунке 7.

Центроиды из фазы 1 используются для кластеризации *k*-средних для пакетов *n* + l, …, *n* + *l*,… Критерием остановки для данного варианта выступает некоторый фиксированный временной интервал (как правило сравнимый с интервалом вычисления k-средних++ на всем наборе данных). В данном варианте результаты кластеризации на новых пакетах аддитивно добавляются к уже полученным. Если во время кластеризации не происходит вырождения, то используются получившиеся центроиды и соответствующие им значения *SSD*, в противном случае возвращаемся к фазе 1. Чтобы получить новые центроиды, была проведена кластеризация с весами на объединенном наборе центроидов (по аналогии с фазой 1). До тех пор, пока не вышло заданное время, процедура повторяется на новых сгенерированных пакетах данных как описано в фазе 1. Наконец, когда срабатывает ограничение по времени, использование полученных результатов, а именно, последних вычисленных центроидов, аналогично пунктам 1 и 2 фазы 1.

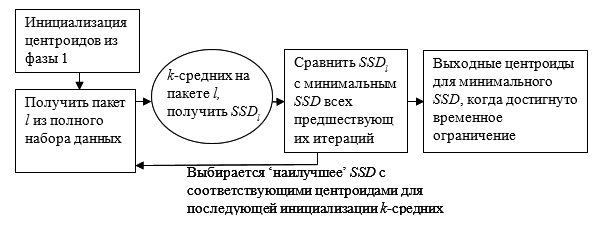


Рисунок 7 – Схема последовательной кластерной декомпозиции на пакетах

*n*+1, *n*+2, ..., *n+l*, *..*.

*Используемый инструментарий и материальная база*

Реализация данного алгоритм выполнена на программном языке Python, используя библиотеки, находящиеся в открытом доступе: numpy, scikit-learn, math, multiprocessing. Материальную базу составляют десктоп компьютеры Intel core i7-4790K 4.0 GHz с 6-ю ядрами, ноутбук Intel core i5 7200U 2.5GHz, а также удаленный высокоскоростной сервер, 40 ядер, 600Гбайт ОЗУ.

*Результаты*

В Таблице 1 представлены результаты экспериментов нашего алгоритма на наборах данных различных размеров. Для каждого из наборов данных генерируются пакеты, размеры которых ранжированы следующим образом: 1/5 × *N*, 1/10 × *N*, ..., 1/100 × *N*, с шагом деления 1/(5*j*) × *N, j =* 1, …, 20.

Таблица 1 – Результаты сравнения улучшенным алгоритма 2 на разных наборах данных и алгоритма *k*-средних. Показаны соответствующие значения *SSD* критерия и время вычисления (схождения алгоритма).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Размер набора данных | Число экспериментов | Улучшение, в сравнении с *k*-средним |  | Число случаев превышения времени |  |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 1590 | 33 (2.1%) | 0.964 | 0 | 0.330 |
|  | 1186 | 35 (2.9%) | 0.989 | 17 | 0.433 |
|  | 312 | 14 (4.5%) | 0.992 | 15 | 0.051 |
|  | 1140 | 345 (30.3%) | 1.000 | 10 | 0.386 |

Продолжение Таблицы 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 |
|  | 300 | 160 (53.3%) | 1.000 | 12 | 0.394 |
|  | 82 | 40 (48.8%) | 0.996 | 41 | 0.971 |
| \* | 30 | 26 (86.7%) | 1.000 | 30 | 1.650 |
| \*\* | 260 | 165 (63.5%) | 1.000 | 0 | 0.160 |

*Описание наборов данных:*

* Первые четыре строки в таблице экспериментов описывают кластеризацию на синтаксических наборах данных с размерами от 1 × 103 элементов до 2 × 106 и с количеством атрибутов от 20 до 100. Рассматривались пакеты разных размеров. Структура синтетических данных – это фиксированное количество синтетических сгустков (blobs) гауссовского распределения.
* Для набора данных размера 50 × 103 число кластеров рассматривалось в диапазоне от 5 до 70 с шагом 2. Синтетические данные представляли собой также случайно сгенерированные плотностные сгустки. Их число перебиралось в диапазоне от 30 до 140. Количество пакетов варьировалось в диапазоне от 5 до 100 с шагом 5.
* Набор данных размера 5 × 106 SUSY получен из открытой базы данных UCI [98].
* \* Набор данных предварительно нормализован.
* \*\* В данном варианте кластеризации искали только два класса; декомпозиция набора данных проводилась на пакеты 1/5 × *N*, 1/10 × *N*, ..., 1/150 × *N*.

*Направление дальнейшей работы*

Данный подход кластеризации предполагает внедрение методологии поиска с чередующимися окрестностями для решения вычислительных задач на больших данных высокой размерности и ее дальнейшая апробация на векторизованном текстовом корпусе. Также планируется интегрирование данной методики в комплекс с альтернативными кластерными методами извлечения информации из текстов.

## Методы, алгоритмы и инструменты предобработки текстовых публикаций

В решения большинства из задач данного проекта необходимо осуществлять обработку и последующий анализ текстовых данных из открытых информационных источников. Предобработка текстовых данных является базовым этапом. От качества предобработки во многом зависит кочечное качество получаемого результата и эффективность разрабатываемой информационной системы. В рамках данного проекта решались следующие задачи предобработки текстовых данных:

* Лемматизация [99];
* Исследовались вопросы применимости имеющихся облачных сервисов для обработки текстов на естественном языке [100];
* Решалась задача автоматического выделения словосочетаний из текстовых корпусов [101];
* Автоматизация процесса извлечения ключевых слов из анализируемых публикаций [102, 103];
* Разрабатывались алгоритмы автоматической суммаризации (реферирования) текстовых документов [102, 103];
* Снятие морфологической неоднозначности [104];
* Исследовались вопросы применимости существующих методов и алгоритмов параметризации и векторного представления текстовых данных в частности с использованием методов дистрибутивной семантики [105].
* Автоматический парсинг зависимостей на синтаксическом уровне (статья в процессе опубликования);
* Распознавание именованных сущностей (статья в процессе опубликования);
* Автоматическая разметка частей речи в тексте (статья в процессе опубликования).

## Автоматическое реферирование текстовых документов с помощью word mover’s distance и извлеченных ключевых слов документа

В рамках данного проекта был разработан метод автоматической суммаризации тестовых документов, опубликованный в [102, 103]. Основная идея метода заключается в использовании ключевых слов, определенных на основе плотного векторного представления слов (word embeddings) и нахождения центроида документа, и ранжирования предложений с помощью Word Mover's Distance (WMD). Результаты ROUGE на наборе данных DUC 2002 показали, что качество результата суммаризаций разработанного подхода находится на одном уровне с другими более сложными современными системами автоматического реферирования, а в некоторых случаях превосходит их. В этой работе также представлен анализ качества оценок суммаризации рассматриваемого набора данных, составленные экспертами.

В данной работе представляется система экстрактивной суммаризации документов, которая использует преимущества векторных представлений на основе нейронных сетей, так и метрики WMD. На первом этапе используются плотные векторные представления слов для извлечения определения центроида документа и ключевых слов документов. Далее WMD применяется для расчета сходства между предложениями и ключевыми словами документов. Для оценки качества работы разработанного метода приведены сравнительные результаты его работы с другими современными методами, основываясь на результатах ROUGE на наборе данных DUC-02.

Рассмотрим разработанный метод по ранжированию предложений для суммаризации документов. Основная идея состоит в выделении тех предложений, которые являются близкими к ключевым словам документов. Ключевые слова определяются как слова, которые наиболее близкие к центру масс представления документа (центроиду документа). В качестве метрики сходства используется Word Mover's Distance (WMD). WMD является одной из лучших метрик для определения смыслового сходства текстов, но процесс вычисление ее является времяемким. Основной причиной этого является то, что WMD основана на решении транспортной задачи. Далее будет представлена система автоматической суммаризации, которая использует преимущества метрики WMD, при этом избегая попарных вычислений, применяет ее к измерению расстояния между предложениями и ключевыми словами документа.

Для апробации данного метода, для представления слов в векторном пространстве использовалась предобученная модель word2vec на GoogleNews2. Предварительная обработка документов произведена была таким образом, чтобы максимизировать количество слов, которые могут быть найдены в словаре данной модели. Первым шагом было проведено разбиение документов на предложения, затем проведена токенизация и удаление stopwords. Понижение регистра всех слов осуществлялось с учетом того, что есть слова, которые хранятся в модели только в верхнем регистре.

Метод экстрактивной суммаризации на основе центроидов был предложен в [106]. Центроид документа является своего рода абстрактным документом, который представляет единственным вектором наиболее значимую информацию исходного документа. Рассмотрим документ с словарем *V* всех слов размером *N*, построим матрицу *E ∈* , где i-й строка матрицы векторное представление i-го слова из V. Центроидом документа будет являться сумма векторных представлений всех уникальных слов в документе.

Ключевые слова – это набор слов ограниченного количества, которые описывают документ. Предполагается, что рассчитанный центроид документа (на предыдущем этапе) представляет центральную тему документа, а слова (то есть их векторное представление) расположенных близко к данному центроиду могут образовывать набор ключевых слов документа. С целью их определения отранжируем все уникальные слова в документе, на основе косинусного сходство между каждым векторным представлением слова и центроидом документа. Выберем 25% наиболее близких к центроиду документа уникальных его слов; они образуют ключевые слова документа.

Для определения предложений документа, которые и будут являться суммаризацией документа, используем метрику сходства WMD, описанную в разделе 2.2:

(8)

где – bag of words i-го предложения, K - набор ключевых слов.

Выбрав ближайшие предложения так, чтобы они по количеству были равны 100 слов (если предложения ровно не укладываются в сто слов, то последнее предложение обрезается), получим резюме документа.

Основываясь на результатах проведённых сравнительных экспериментов [102] разработанный метод генерирует достаточно успешно краткое изложение текста с приемлемой оценкой ROUGE. Более того, он может успешно конкурировать с уже существующими системами.

## Разработка методов оптимизации

Большинство из разрабатываемых в рамках данного проекта алгоритмов имеют наборы настраиваемых входных параметров. Кроме этого при разработке каждого алгоритма необходимо определить критерий для численной оценки качества его работы. Так как у большинства из алгоритмов имеется по несколько настраиваемых параметров, то возникает широкий спектр задач, связанных с их оптимизацией на основе алгоритмов параметрической оптимизации.

В рамках данного проекта разрабатывались и применялись различные оптимизационные методы [107-113]. Широко использовалась оптимизация в решении задач кластерного анализа [107, 109]. Оптимизация рассматривалась для класса задач Minimum Sum-of-Squares Clustering Problem, к числу которых относятся k-means, k-means++, j-means. В большинстве случаев алгоритмы класса k-means останавливаются в области локального минимума по критерию Minimum Sum-of-Squared Distances. Однако применение техник Variable Neighborhood Search позволяет приблизится к глобальному минимуму, не останавливаясь на локальных [112, 113].

Так как в решении различных задач проекта приходится обрабатывать большие объёмы информации в том числе и в процессе параметрической идентификации, часто приходится применять методы распараллеливания алгоритмов обработки и анализа данных. При распараллеливании встают задачи оптимального планирования мультипроцессорных вычислений [109].

В процессе кластеризации как текстовых документов, так и различных других лексических единиц анализа можно строить различные графовые структуры в том числе и онтологии. Например, на основе методов дистрибутивной семантики можно строить различные графовые представления словарей где связи между словами устанавливаются на основе меры их контекстной и семантической близости. После построения подобных лексических графовых структур возникает необходимость решения Hub Location Problem [110].

Для решения задач апроксимации экспериментальных данных применялся метод поиска с переменными окрестностями (VNS - variable neighbourhood search) [112, 113].

## Создания необходимых технических условий

Для успешной и бесперебойной работы разработанной информационной системы, сбора, анализа текстовых данных и тестирования алгоритмов группой инженерно-технических работников улучшены технические условия [114]: установлены Система хранения данных (СХД) и программное обеспечение для виртуализации серверов VMware (гипервизор ESXi и vCenter), проведена модернизация существующего сервера: установлены и настроены HBA-карты для подключения сервера к системе хранения данных. В СХД создано два дисковых массива RAID10, 4.3Тб и 2.2Тб, для каждого сервера и в каждом массиве выделен LUN, а также настроено зонирование потоков данных для обеспечения отказоустойчивости (Рисунок 8). Все сервера и оборудования СХД перенесены в ЦОД на CO-LOCATION и настроен удаленный доступ к каждой единице оборудования из локальной сети ИИВТ.

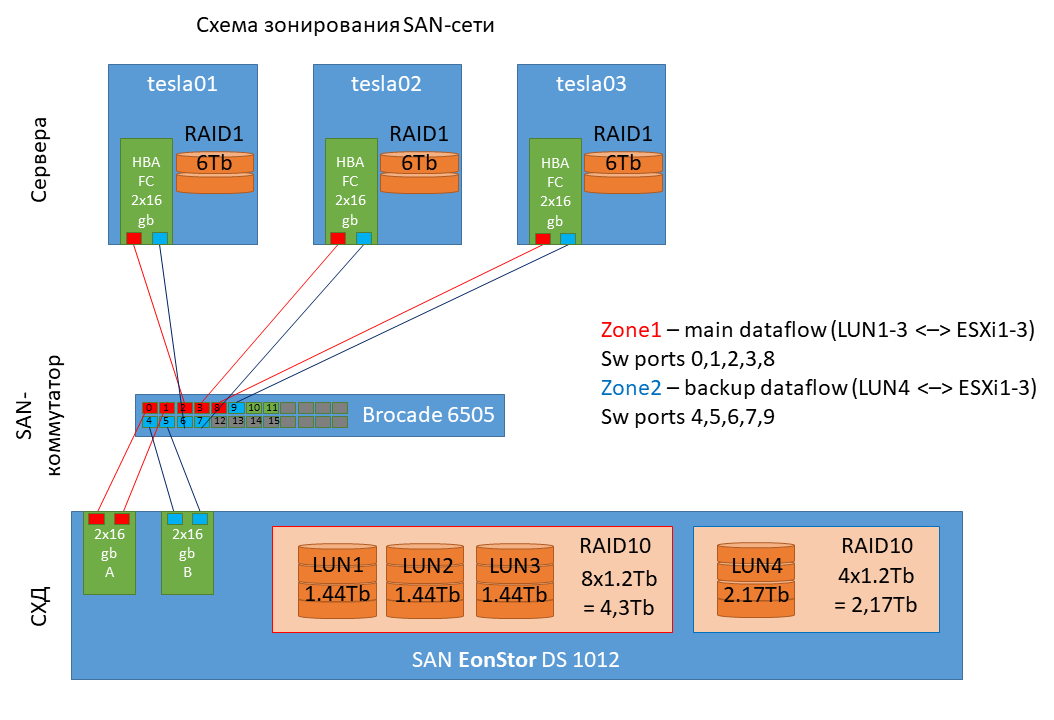


Рисунок 8 – Схема зонирования SAN-сети

Созданы виртуальные машины: рабочие виртуальные сервера для проведения расчетов на CPU и GPU (настроен проброс видеокарт в режиме Passthru); для хранения и обмена данными; для управления виртуализацией; для удаленного доступа к серверам; виртуальный маршрутизатор для обеспечения интернет-траффика. На рабочих серверах с видеокартами установлено и настроено: ОС Ubuntu 16.04, cuda 10.0, cudnn 7.3, Anaconda 3, библиотеки для машинного обучения, JupyterHub, nbextensions. На рабочих серверах с видеокартами настроен доступ из сети интернет по протоколам HTTPS, заведены доменные имена для каждого рабочего сервера и персональные учетные записи для всех разработчиков проекта, настроена политика безопасности доступа к серверу и данным. Для хранения и обмена данными все домашние директории пользователей перенесены на сервер для хранения и обмена данными. Настроен обмен данными между серверами по протоколу NFS.

Для разработки и внедрения информационной системы, согласно проекту, были выделены ресурсы и созданы виртуальные сервера:

1. *Vm-app* – основной сервер приложений. Система для мониторинга, хранения и обработки данных. Содержит базы данных (Postgresql), noSQL хранилище (ElasticSearch+Kibana), веб приложение на Django, стек Airflow+Celery и сопутствующие служебные компоненты (nginx, redis). Всё развернуто в docker контейнерах.
2. *Vm-registry* - приватный Docker registry (реестр) для хранения готовых собранных докер образов.

Была произведена реконфигурация ресурсов имеющихся виртуальных серверов, в связи изменившимися потребностями в вычислительных мощностях. Большая часть виртуальных CPU и RAM была изъята из одного из серверов для математического моделирования и передана в пользу сервера приложений.

## Создание необходимых экспертно-аналитических условий

Целью работы является формирование сбалансированного, представительного, размеченного и проклассифицированного корпуса для автоматизированной оценки влияния отрытых текстовых информационных источников на социум на основе публикуемой информации [50].

В рамках комплекса работ по формированию корпуса публицистических текстов: текстов официальных СМИ, журнальных публикаций, новостных порталов и др., были определены следующие задачи:

* проведение первичного отбора источников и сбор данных;
* определение требований к корпусу: сбалансированность, представительность, основные категории мета-разметки;
* формирование основного, малого и тестового корпусов;
* чистка корпусов и систематизация данных, разметка тестового корпуса.

В ходе проведенного комплекса работ по формированию дополнительного сбалансированного, представительного, размеченного и проклассифицированного корпуса публицистических текстов информационных новостных порталов было сформировано 3 корпуса:

* 1. основной корпус, включающий 1 982 433 статьи из 22 источников с мета-разметкой по 14 категориям;
  2. малый корпус текстов, включающий 245 352 статьи из 5 источников с мета-разметкой по 7 категориям;
  3. тестовый корпус, включающий 35 643 статьи из 5 источников, с разметкой по 39 категориям и классификацией по части из признаков.

*Методология формирования основного корпуса.* Из общего списка доменов, состоящего из 2613 СМИ с указанием домена, типа СМИ (печатные, телевидение, радио, новостные агентства, интернет-СМИ), и масштаба (региональные, центральные), на основе специфики методов машинного обучения при анализе текстовой информации [84], был произведен отбор источников. При анализе и отборе источников учитывались требования по:

* исключению региональных СМИ;
* исключению узкоспециализированных источников;
* обеспечению включения в состав корпуса публикаций на русском языке;
* включению в корпус информационных порталов, специализирующихся на предоставлении ежедневных новостей о событиях в стране;
* включению в корпус публикаций информационных порталов СМИ, различных по уровню популярности среди населения.

Исходя из этих требований, из выборки были исключены региональные СМИ, узкоспециализированные СМИ, СМИ публикующие статьи на других языках, СМИ активность которых приостановлена.

В результате был сформирован основной корпус из 1 982 433 статей, основные характеристики которого представлены в таблице Ж1 приложения Ж.

По результатам работ над формированием дополнительного большого корпуса были выделены следующие категории мета-разметки:

1. Автор публикации – имя автора;
2. Категория/ тема публикации – тематический раздел сайта, к которому относится публикация;
3. Дата выхода публикации;
4. Дата скачивания публикации;
5. Идентификационный номер публикации - автоматически присваиваемый статье идентификационный номер для ее поиска и исключения двойного учета или «неучтения» статьи в процессе анализа;
6. Ссылки - наличие гиперссылок на др. статьи или странице в тексте самой публикации;
7. Количество комментариев;
8. Количество просмотров (на дату скачивания);
9. Заголовок;
10. Подзаголовок;
11. Теги (теги и хеш-теги присвоенные публикации);
12. Текст публикации;
13. Текст публикации в формате html;
14. URL адрес публикации.

В нашем случае репрезентативность в отношении исследуемой области большого корпуса не подлежит сомнению, так как выборка дополнительного корпуса включила исключительно из текстов, представляющих исследуемую область (всю популяцию текстов, открытых информационных, источников, оказывающих влияние на социум Казахстана). Поскольку основной целью проекта является исследование влияния открытых информационных текстовых источников на социум, данная задача была решена полностью: дополнительный корпус представляет собой собрание текстов исключительно из источников данного типа, а также охватывает практически весь период существования таких источников, и практически весь спектр источников, оказывающих влияние на социум.

В рамках исследования, исходя из того, что собранный корпус отражает всю совокупность казахстанских открытых текстовых источников СМИ с информацией новостного характера, оказывающих влияние на социум, требование к сбалансированности всего корпуса необходимо не воспринимать как критически важное, поскольку самое главное условие - отражение объективной реальности, полностью соблюдено (Рисунок Ж1 Приложения Ж).

Необходимо также уточнить, что для целей машинного обучения и тестирования алгоритмов проектируемой в рамках проекта информационной системы, из общего корпуса осуществлена выборка, где принцип сбалансированности в отношении исследуемых признаков соблюден надлежащим образом.

Необходимо отметить, что во многих случаях данные по вышеуказанным категориям мета-разметки отсутствуют. Это обусловлено, прежде всего, существующей архитектурой данных. К примеру, лишь небольшая часть сайтов спроектирована таким образом, что можно собрать данные по количеству просмотров и количеству комментариев к публикациям, размещенным на них. На многих сайтах возможность комментирования публикаций отсутствует. Это значительно усложняет анализ данных, поскольку с точки зрения оценки влияния на социум, количество комментариев и просмотров публикаций является, почти единственным прямым индикатором заинтересованности населения.

Подготовительным этапом, обусловливающим решение задач по автоматизации оценки новостных текстов и распознаванию информативных признаков, является комплекс работ по созданию словарного массива, отбору языковых единиц и способов (правил) их употребления в новостных текстах. В связи с этим целью данной работы является разработка методов формализации правил и словарей для распознавания в корпусе новостных текстов информативных признаков и определения степени их выраженности.

Были определены 12 информативных признаков, для распознавания которых необходимо формирование словарей и формализация лингвистических шаблонов и правил.

Объектом анализа выступили 2 000 текстов из 5 информационных новостных интернет-источников: «KazakhSTAN 2.0», «Central Asia Monitor», «Zakon.kz», «Радио Азаттык», «Tengrinews.kz».

Для достижения цели в текущем году были поставлены и решены следующие задачи:

* 1. Проведение глубинного анализа сформированного корпуса публикаций СМИ по 9 информативным признакам (3 признака в 2018 г., 9 – в 2019 г.).
  2. Осуществление разметки текстового корпуса по обозначенным информативным признакам.
  3. Формирование и систематизация словарей по определенным информативным признакам на базе анализируемых текстов.
  4. Выявление закономерностей для формализации правил распознавания отдельных информативных признаков.
  5. Получение экспертных заключений о реализуемости выявленных закономерностей в рамках проекта.
  6. Описание методов формализации правил на основе лингвистического анализа текстов.

По завершении работы группы лингвистов, выявленные закономерности по 8 признакам из 12 были вынесены на обсуждение. В ходе экспертного обсуждения по выявленным закономерностям были даны комментарии и заключения относительно применимости и реализуемости результатов работы лингвистов, где заключение «реализуемо» означает наличие программных средств, позволяющих реализацию выявленной закономерности в рамках разрабатываемого продукта; заключение «сложно реализуемо» означает, что реализация выявленной закономерности требует построения сложных многоэтапных алгоритмов и временных затрат и, возможно, нецелесообразно для реализации с точки зрения целей и бюджета проекта; заключение «не реализуемо» означает отсутствие программных средств, позволяющих реализацию выявленной закономерности в рамках разрабатываемого продукта; заключение «требует обсуждений» означает необходимость получения дополнительных консультаций с экспертами для вынесения окончательного заключения.

# ЗАКЛЮЧЕНИЕ

В рамках проекта программно-целевого финансирования за 2019 год в соответствие с календарным планом получены следующие основные результаты:

1. Получены методики для оценки влияния открытых информационных источников на социум на основе анализа публикуемой текстовой информации;
2. Начата разработка методики расчета отдельных критериев для оценки влияния открытых текстовых информационных источников на социум;
3. Начато проектирование и разработка информационной системы оценки влияния открытых текстовых информационных источников на социум;
4. Начаты мониторинг и корректировка отклонений в работе разработанной информационной системы оценки влияния открытых текстовых информационных источников на социум
5. Для текущего этапа созданы необходимые технические и экспертно-аналитические условия для разработки информационной системы оценки влияния открытых текстовых информационных источников на социум.

Все поставленные задачи выполнены в полном объеме и в соответствие с календарным планом (Приложение Б).

Результаты НИР будут использованы при разработке информационно-аналитической системы оценки влияния открытых текстовых информационных источников на социум.

Полученные результаты соответствуют текущему уровню научно-технического развития и основаны на последних достижениях в области обработки данных на естественном языке, машинного и глубинного обучения, распознавания образов и др.

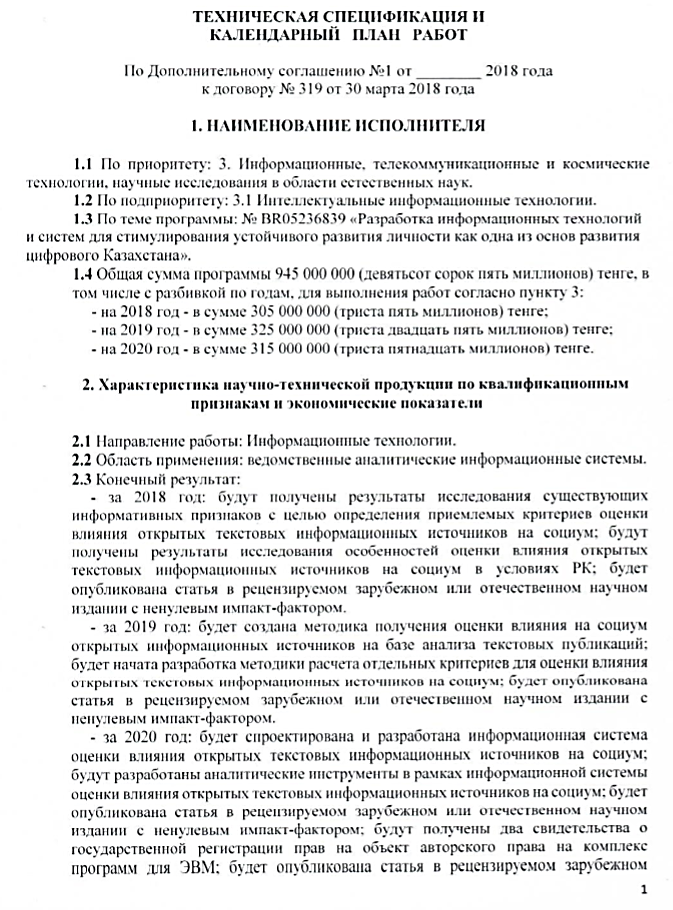
Согласно п. 9 календарного плана проекта, в 2019 году опубликовано 15 статей в зарубежных рецензируемых изданиях с ненулевым импакт-фактором и получено 3 авторских свидетельства на программу для ЭВМ, а так же 5 статей в отечественных научных изданиях, рекомендованных ККСОН и 20 статей в трудах международных конференций.

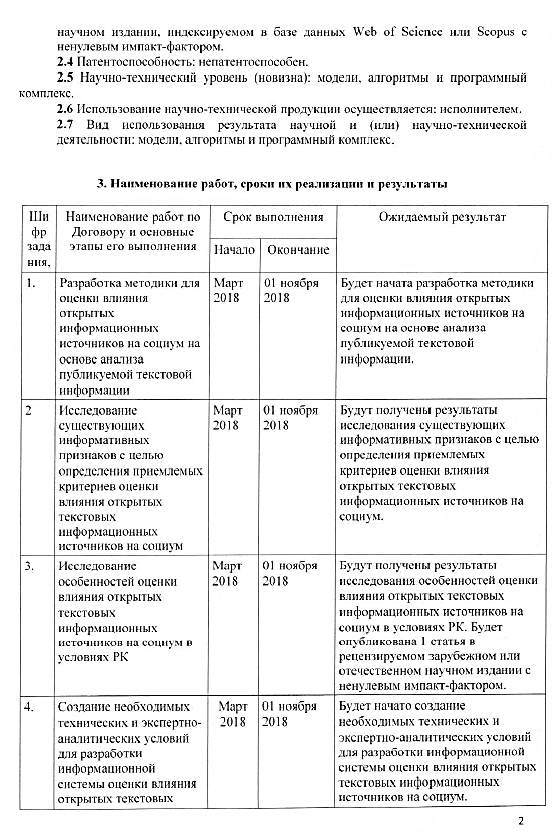
# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

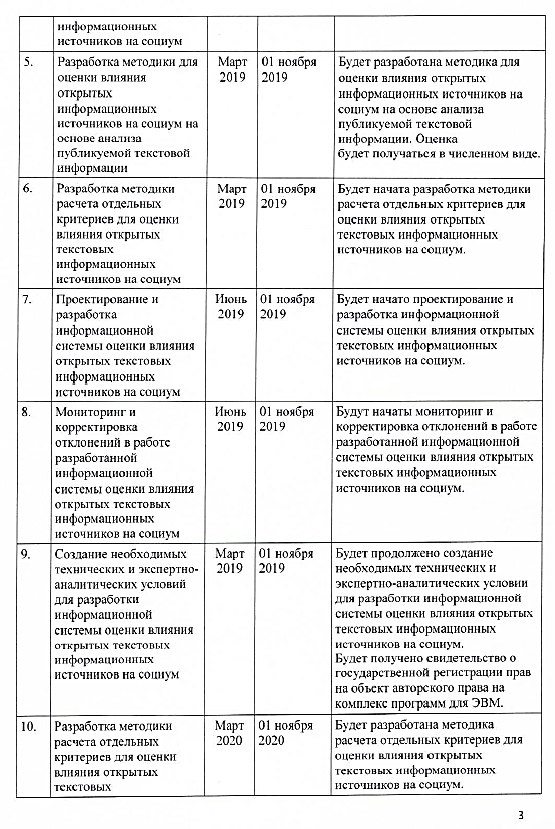
1. Barakhnin V. B., Duisenbayeva A. N., Kozhemyakina O.Yu., Yergaliyev Y. N. and Muhamedyev R. I. The automatic processing of the texts in natural language. Some bibliometric indicators of the current state of this research area// Journal of Physics: Conference Series. -2018 // <https://doi.org/10.1088/1742-6596/1117/1/012001>: 21.10.2019
2. Барахнин В.Б., Дуйсенбаева А., Кожемякина О.Ю., Кучин Я.И., Якунин К.О., Мухамедиев Р.И. The studies of dinamics’ changes in publication activity in the field of natural language processing// Матер. 17-ой Междунар. науч. конф. "Информационные Технологии и Менеджмент 2019". - Рига, Латвия, 2019. -С. 71-74
3. Барахнин В., Дуйсенбаева А., Кожемякина О., Кучин Я., Якунин К., Мухамедиев Р. Изменение публикационной активности в области обработки естественного языка // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 130-135
4. А.с. № 4505. Программа расчета динамических библиометрических показателей D1 и D2 (Calculation of dynamic scientometric indicators D1, D2 – CalcDStMi)/ Якунин К.О., Мухамедиев Р.И., Кучин Я.И.; опубл. 12.07.2019. – 1с.
5. Global research "2019 Edelman Trust Barometer"// <https://www.edelman.com/trust-barometer> : 18.07.2019
6. Miller D. Promotional strategies and media power // Briggs, A., Cobley, P. (eds.), THE MEDIA: AN INTRODUCTION. - London, United Kingdom: Longman, 1998. -С. 65-80
7. Bushman B., Whitaker J. Media Influence on Behavior // Encyclopedia of Human Behavior, -2012. -P. 571-575
8. Don W., Zongchao C., Cylor, S. Media Effects. In: James, D. INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES, Second Edition, - Elsevier, 2015. - С. 29-34
9. Hoon K., Jong Y., Sangheon K., Libor M., In S.: Human-machine interaction: A case study on fake news detection using a backtracking based on a cognitive system // Cognitive Systems Research, -2019. -Т. 55. –С. 77-81
10. Bushman B., Whitaker, J. Media Influence on Behavior // Reference Module in Neuroscience and Biobehavioral Psychology, -2017.
11. Машечкин И.В., Петровский М.И., Царёв Д.В. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования // Вычислительные методы и программирование, -2013. - Т. 14(1). –С.91-102
12. Пархоменко П., Григорьев А.А., Астраханцев А.А. Обзор и экспериментальное сравнение методов кластеризации текстов // Труды институт системного программирования РАН, -2017. –Т. 29(2). –С. 161-200
13. Vorontsov K., Anna, P.: Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование, - 2012. -Т. 4(4). –С. 693-706
14. Vorontsov, K., Frei, O., Apishev, M., Romov, P., & Dudarenko, M. (2015). BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. Analysis of Images, Social Networks and Texts, -2015. –С. 370–381 // doi:10.1007/978-3-319-26123-2\_36  : 21.10.2019
15. Blei D., Ng A., Jordan M. Latent Dirichlet allocation// Journal of Machine Learning research. -2003. –Т.3. –С. 993-1022
16. Scott J. A decision support system for supplier selection and order allocation in stochastic, multi-stakeholder and multi-criteria environments// International Journal of Production Economics, №166, -2015. –С. 226-237
17. Mardani A., Jusoh A., Zavadskas E., Cavallaro F., KhalifahZ. Sustainable and renewable energy: an overview of the application of multiple criteria decision-making techniques and approaches //Sustainability. -2015. -Т. 7(10). –С. 13947-13984
18. Wanderer T., Stefan H. Creating a spatial multi-criteria decision support system for energy related integrated environmental impact assessment// Environmental impact assessment review. -2015. -№52. -С. 2-8
19. Hoceini Y., Mohamed C., Moncef A. Towards a new approach for disambiguation in NLP by multiple criterian decision-aid// The Prague Bulletin of Mathematical Linguistics. -2011. -№ 95. –С. 19-32
20. Kumar A. A review of multi criteria decision making (MCDM) towards sustainable renewable energy development// Renewable and Sustainable Energy Reviews. -2017. -№69. -С. 596-609
21. Yager R. On ordered weighted averaging aggregation operators in multicriteria decision making// IEEE Transactions on systems, Man, and Cybernetics. -1988. -№18 (1). –С. 183-190
22. Hansen, P., Franz O. A new method for scoring additive multi‐attribute value models using pairwise rankings of alternatives// Journal of Multi‐Criteria Decision Analysis. -2008. -№15(3‐4). –С. 87-107
23. Figueira J., Vincent M., Bernard R. Electre methods // Multiple Criteria Decision Analysis: State of the Art Surveys. Part of the International Series in Operations Research & Management Science book series, -2005. –Т.78. –С. 133-153
24. Lai Y., Ting-Yun L., Ching-Lai H. Topsis for MODM // European journal of operational research. -1994. -№76(3). –С. 486-500
25. Brans J., Ph V. A Preference Ranking Organization Method (The PROMETHEE Method for Multiple Criteria Decision-Making)// Management science. -1985. -№ 31 (6). –С. 647-656
26. Saaty T. Group decision making and the AHP// The Analytic Hierarchy Process. Springer, Verlag. -1989. –С. 59-67// <https://doi.org/10.1007/978-3-642-50244-6_4>: 21.10.2019
27. Abaei M. Developing a novel risk-based methodology for multi-criteria decision making in marine renewable energy applications// Renewable energy. -2017. №102. –С. 341-348
28. Mukhamediyev R. Tools and Methods of Information Support for the Development of Renewable Energy// Renewable Energy and Power Supply Challenges for Rural Regions, IGI Global, -2019. –С. 39-65
29. Charabi Y., Adel G. PV site suitability analysis using GIS-based spatial fuzzy multi-criteria evaluation// Renewable Energy. -2011. -№36 (9). -С. 2554-2561
30. Jelassi T., Vladimir O. A framework for building an expert system for MCDM models selection// Improving Decision Making in Organizations, -1989. -С. 553-562
31. Dalkey N., Olaf, H. An experimental application of the Delphi method to the use of experts// Management science, -1963. -№ 9(3). -С. 458-467
32. Okoli C., Suzanne P. The Delphi method as a research tool: an example, design considerations and applications// Information & management, -2004. -№42 (1). -С. 15-29
33. Allen E., Christopher S. Likert scales and data analyses// Quality progress, -2007. -№40 (7). -С. 64-65
34. Mokken, R. A theory and procedure of scale analysis with applications in political research// De gruyter mouton, Series:Methods and Models in the Social Sciences. -2011
35. Orme B. Maxdiff analysis: Simple counting, individual-level logit, and HB // Sawtooth Software, Inc. -2009
36. Bradley R., Milton T. Rank analysis of incomplete block designs: I. The method of paired comparisons// Biometrika. -1952. -№ 39(3/4). -С. 324-345
37. Saati T., Andreychikova O. Об изменении неосязаемого. Подход к относительным изменениям на основе главного собственного вектора матрицы парных сравнений// Cloud of science. -2015. -№2(1). –С. 5-39
38. Яндекс назвал самые популярные запросы казахстанцев // <https://kapital.kz/lifestyle/74330/yandeks-nazval-samye-populyarnye-zaprosy-kazahstancev.html>: 20.09.2019
39. Байманов Д. Волнующие казахстанцев вопросы назвал Президент РК. // <https://www.inform.kz/ru/volnuyuschie-kazahstancev-voprosy-nazval-prezident-rk_a3449476> : 20.09.2019
40. Жулмухаметова Ж. Тройку самых острых вопросов, волнующих казахстанцев, назвали социологи. // <https://informburo.kz/novosti/troyku-samyh-ostryh-voprosov-volnuyushchih-kazahstancev-nazvali-sociologi.html> : 20.09.2019
41. Барахнин В.Б., Кучин Я.И., Мухамедиев Р.И. К вопросу о постановке задачи выявления фейковых новостей и алгоритмах их мониторинга // Матер. III Междунар. науч. конф. «Информатика и прикладная математика», посв. 80-летнему юбилею проф. Бияшева Р.Г. и 70-летию проф. Айдарханова М.Б. - Алматы, 2018. -С. 113-118
42. Jaccard P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines// Bull Soc Vaudoise Sci Nat. -1901. -№37 (140), -С. 241-272
43. Kravtsov G. Обучаемая модель вычислений на классификациях// Yelektronne mo-delyuvannya. -2018. -№40(3). -С. 63-76
44. Articles and term dictionaries in Russian language// <https://drive.google.com/open?id=1CDEs6ooXtCq6v_jg-xSUVuPS1AE1ZSOn> : 18.07.2019.
45. R. I. Mukhamediev, R. Mustakayev, K. Yakunin, S. Kiseleva and V. Gopejenko. Multi-Criteria Spatial Decision Making Support Ssystem for Renewable Energy Development in Kazakhstan // IEEE Access. -2019. - T. 7. - С. 122275-122288
46. Мусабаев Р., Мухамедиев Р., Кучин Я., Сымагулов А., Якунин К., Мурзахметов С. Метод мультимодальной оценки средств массовой информации на основе тематической модели корпуса текстов // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий». – Алматы. - 2019. -C. 239-247
47. Мусабаев Р.Р., Мухамедиев Р.И., Кучин Я.И., Сымагулов А., Якунин К.О. On a method of multimedial media ranking using corpus based topic modeling// Матер. 17-ой Междунар. науч. конф. "Информационные Технологии и Менеджмент 2019". - Рига, Латвия, 2019. -С. 64-67
48. Мухамедиев Р.И., Мусабаев Р.Р., Булдыбаев Т., Кучин Я., Сымагулов А., Оспанова У., Якунин К., Мурзахметов С., Сагындык Б. Эксперименты по оценке средств массовой информации на основе тематической модели корпуса текстов // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 419-431
49. А.с. №4523. Программа мультимодальной оценки СМИ на базе тематической модели корпуса текстов (Multi Modal Mass Media Assesment – M4A)/ Якунин К.О., Мухамедиев Р.И., Кучин Я.И., Мурзахметов С.Б., Сымагулов А., Мусабаев Р.Р.; опубл. 15.07.2019. -1 с.
50. Атанаева М., Булдыбаев Т. и др. Отчет о НИР АО «ИАЦ» за 2019 год по научно-технической программе «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана». Нур-Султан. – 2019. // <http://iict.kz/wp-content/uploads/2019/10/iac-report-2019_11.10_final.docx> : 23.10.2019
51. Мейрамбеккызы Ж., Аманбай А., Мусабаев Р. Автоматическое формирование тематических словарей социально-значимых понятий // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 233-239
52. Meirambekkyzy Zh., Mussabaev R.R. Automatic development of topic dictionaries of socially significant concepts // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 65-70
53. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. - 2003. -Vol. 3. -С. 993–1022
54. К. В. Воронцов. Вероятностное тематическое моделирование // Электронный ресурс <http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf>: 14.08.2019
55. Korenčić D., Ristov S., Najder J.E. Document-based topic coherence measures for news media text // Expert Systems with Applications. – 2018. – Vol, 114. – P. 357-373.
56. Атанаева М.К., Булдыбаев Т.К., Оспанова У.А., Акоева И.Г., Нурумов К.С., Баймаханбетов М.А. Методика для определения информативных признаков новостных текстов и проверка их значимости // Научный аспект, – 2019. - Т.3(3). –С. 277-295
57. Buldybayev T., Baimakhanbetov M., Nurumov K., Akoyeva I., Ospanova U., Atanayeva M. Analysis of digital media: a comprehensive assessment of informative features significance // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 31-44
58. Атанаева М.К., Булдыбаев Т.К., Оспанова У.А., Акоева И.Г., Нурумов К.С., Баймаханбетов М.А. Определение тональности и объективности новостных текстов словарным подходом // Научный аспект. -2019. -Т.3(3). -С.296-308
59. Нурумов К.С., Атанаева М.К., Булдыбаев Т.К., Оспанова У.А., Акоева И.Г. Комплексная оценка значимости информативных признаков для оценки влияния на социум на основе анализа публикуемой текстовой информации // Вестник ЕНУ им. Л.Н. Гумилева, Серия Журналистика. -2018. -№3 (124). -С. 44-52
60. Атанаева М.К., Оспанова У.А., Булдыбаев Т.К., Акоева И.Г., Нурумов К.С. Использование лингвистических подходов в определении классов и информативных признаков в текстах // Вестник КазНПУ, Серия филологическая. -2018. -№2 (64). -С. 166-171
61. Toleu A., Tolegen G., Mussabayev R. Initial exploration for Topic modeling with word sense // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 31-44
62. Vorontsov K., Potapenko A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // Ignatov D., Khachay M., Panchenko A., Konstantinova N., Yavorsky R. (eds) Analysis of Images, Social Networks and Texts. AIST 2014. Communications in Computer and Information Science. Springer, Cham. -2014. –Vol. 436. –C. 29-46
63. R. Rehurek, P.Sojka. Software framework for topic modelling with large corpora // The LREC 2010 workshop on new challenges for NLP frameworks, University of Malta, -2010. -C. 45-50
64. Мусабаев Р. Результаты тематического моделирования по корпусу tengrinews. Алматы. -2019. // <http://iict.kz/wp-content/uploads/2019/10/results-for-tengrinews.zip> : 23.10.2019
65. Shalkarbayuli A., Kairbekov A., Amangeldi Y. Comparison of traditional machine learning methods and Google services in identifying tonality on Russian texts// Journal of Physics: Conference Series. -2018. –Vol.1117. -C.1-9
66. Nurumov K., Baimakhanbetov M., Buldybayev T., Akoyeva I., Ospanova U. An approach for detecting sentiments in mass media publications using content specific Informative features // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 21-31
67. Хорошилов А.А., Мусабаев Р.Р., Красовицкий А.М., Уалиева И.М., Мейрамбеккызы Ж. Автоматическое установление тональности текстов методом их концептуального анализа // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 365-375
68. A.M. Krassovitsky, I.M. Ualiyeva, Zh. Meirambekkyzy, R.R. Mussabayev. In the search of a linguistic model for automatic generalization recognition in media texts // International Journal of Advances in Electronics and Computer Science. -2019. -Vol. 6(3). -C.38-43
69. Красовицкий А.М., Уалиева И.М., Мейрамбеккызы Ж., Мусабаев Р.Р. Основанный на лексиконе подход в оценке обобщений в русскоязычных СМИ // Междунар. науч. журнал «Современные информационные технологии и ИТ образование». -2018. -Т.14(3). -C. 563-567
70. Krassovitsky A., Ualiyeva I., Meirambekkyzy Zh., Mussabayev R. Automatic generalization recognition in media texts /// Proceedings 20th International Conference on Computational Linguistics and Intelligent Text Processing. -La Rochelle, France, 2019.
71. Ualiyeva I.M., Mussabayev R.R. Features analysis for detecting information-dense texts in online news media // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 74-79
72. Барахнин В.Б., Кожемякина О.Ю., Рычкова Е.В., Пастушков И.С., Борзилова Ю.С. Извлечение лексических и метроритмических признаков, характерных для жанра и стиля и их комбинаций в процессе автоматизированной обработки текстов на русском языке // Междунар. науч. журнал «Современные информационные технологии и ИТ образование» – 2018. – Т. 14(4). – С.876-883. – <http://dx.doi.org/10.25559/SITITO.14.201803.876-883>: 21.10.2019
73. Уалиева И.М., Мусабаев Р.Р. Исследование признаков для автоматической идентификации жанров публикаций в СМИ // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 331-336
74. Барахнин В.Б., Мухамедиев Р.И., Мусабаев Р.Р., Кожемякина О.Ю., Исаева А., Кучин Я.И., Мурзахметов С.Б., Якунин К.О. Методы выявления деструктивных новостей в медиапространстве // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 205-219
75. Барахнин В.Б., Якунин К. Methods to identify the destructive information // Journal of Physics: Conference Series. - 2019.
76. Атанаева М.К., Оспанова У.А., Булдыбаев Т.К., Акоева И.Г., Нурумов К.С. Определение индикаторов оценки достоверности публикаций в СМИ для разработки методики определения степени вероятной достоверности отдельной публикации // Вестник КазНПУ, Серия филологическая. -2018. -№2 (64). -С. 158-165
77. Оспанова У.А., Атанаева М.К., Акоева И.Г., Булдыбаев Т.К. Identification of features of probable reliability of news // Современное педагогическое образование. -Москва, 2019. -№6. -С. 44-49
78. Атанаева М.К., Булдыбаев Т.К., Оспанова У.О., Акоева И.Г., Нурумов К.С., Баймаханбетов М.А. Определенеие закономерностей подозрительности в новостных текстах // Вестник ЕНУ им. Л.Н. Гумилева, Серия журналистика. - Нур-Султан, 2019. -№2 (127). -С 8-19 // DOI: <https://doi.org/10.32523/2616-7174-2019-127-2-8-20> : 21.10.2019
79. Оспанова У.А., Атанаева М.К., Акоева И.Г., Булдыбаев Т.К. Informative features of bias and reliability of electronic Mass Media // Социология. - г. Москва, 2019. -№2. –С. 259-269
80. А.с. № 2894. NLP-Preprocessor / Якунин К., Мухамедиев Р., Кучин Я., Мурзахметов С., Сымагулов А., Мустакаев Р., опубл. 18.04.2019. -1 с.
81. Mladenovic N., Krassovitskiy A., Mussabayev R. Decomposition approach for clustering problem // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 15-21
82. Козбагаров О.Б., Мусабаев Р.Р. Кластеризация новостных публикаций по информационным поводам // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 340-350
83. Захаров В.Н., Мусабаев Р.Р., Красовицкий А.М., Козловская Я.Д., Хорошилов А-р А., Хорошилов А-й А. Метод кластеризации новостных сообщений средств массовой информации на основе их концептуального анализа // Системы и средства информатики, Т.29 № 3, 2019. -С. 52-66 // DOI: 10.14357/08696527190305: 21.10.2019
84. Младенович Н., Красовицкий А., Мусабаев Р. Метод декомпозиции в задаче кластеризации больших данных // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 525-533
85. Karlsson, C. (ed.): Handbook of Research on Cluster Theory // Edward Elgar Publishing, - 2010.
86. Krassovitskiy A., Mussabayev R. Energy-Based Centroid Identification and Cluster Propagation with Noise Detection.// Nguyen N., Pimenidis E., Khan Z., Trawiński B. (eds) Computational Collective Intelligence. ICCCI 2018. Lecture Notes in Computer Science. -Springer, Cham. -2018. -Т. 11055. –С. 523-533
87. Bagirov, A.M.: Modified global k-means algorithm for minimum sum-of-squares clustering problems. Pattern Recognition. -2008. –Т. 41 (10). -С. 3192—3199 // http://www.sciencedirect.com/science/article/pii /S0031320308001362 : 09.06.2019
88. Bahmani B., Moseley B., Vattani A., Kumar R., Vassilvitskii S., Scalable k-means++ // Proc. VLDB Endow. -2012. -№5(7). -С. 622–633
89. Capó M., Pérez A., Lozano J.A., An efficient approximation to the k-means clustering for massive data. Knowl.-Based Syst. -2017. –Т. 117. –С. 56–69
90. Chiang M.C., Tsai C.W., Yang C.S., A time-efficient pattern reduction algorithm for k-means clustering // Inform. Sci. -2011. -№181(4). -С. 716–731
91. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J. and Steinberg, D. Top 10 Algorithms in Data Mining // Knowl. Inf. Syst. -2007. -№14(37). // DOI 10.1007/s10115-007-0114-2, 2008: 21.10.2019
92. HajKacem M.A.B., N’Cir CE.B., Essoussi N. Overview of Scalable Partitional Methods for Big Data Clustering // Nasraoui O., Ben N'Cir CE. (eds) Clustering Methods for Big Data Analytics. Unsupervised and Semi-Supervised Learning. -Springer, Cham, 2019. –С. 1-23
93. Comparison of the K-Means and MiniBatchKMeans clustering algorithms// <https://scikit-learn.org/stable/auto_examples/cluster/plot_mini_batch_kmeans.html> : 09.06.2019
94. S.J. Redmond, C. Heneghan, A method for initialising the K-means clustering algorithm using kd-trees// Pattern Recogn. Lett. -2007. -№28(8). -С. 965–973
95. Comparing different clustering algorithms on toy datasets// URL: http:// scikit-learn.org/stable/auto\_examples/cluster/plot\_cluster\_comparison.html : 09.06.2019
96. X. Cui, P. Zhu, X. Yang, K. Li, C. Ji, Optimized big data K-means clustering using MapReduce. J.// Supercomput. -2014. -№70(3). –С. 1249–1259
97. T. Washio, K. Nakanishi, H. Motoda. Deriving Class Association Rules Based on Levelwise Subspace Clustering. // In: Jorge A.M., Torgo L., Brazdil P., Camacho R., Gama J. (eds) Knowledge Discovery in Databases.PKDD 2005. Lecture Notes in Computer Science. - Springer, Berlin, Heidelberg. 2005. -Т. 3721. –С. 692-700
98. Online clustering data sets UCI// URL https://archive.ics.uci.edu/ml/datasets.html. : 09.06.2019
99. Akhmetov I.R., Krassovitsky A., Ualiyeva I., Myssabayev R.R. Lemmatization of Russian Language by Tree Regression Models // Proceedings 20th International Conference on Computational Linguistics and Intelligent Text Processing, -La Rochelle, France, 2019.
100. Мухамедиев Р.И., Сымагулов А., Кучин Я., Абдуллаева С., Абдолдина Ф.Н. Облачные сервисы для обработки текстов на естественном языке // Междунар. науч. журнал «Современные информационные технологии и ИТ образование». – 2018. – Т.14(4). –С.859-869
101. Барахнин В.Б., Кожемякина О.Ю., Рычкова Е.В., Борзилова Ю.С. Автоматическое выделение словосочетаний из текстов славянского происхождения: сравнение подходов // Cloud of Science,– 2018. –Т.5(4). –С.713-728
102. Seitkali D., Mussabayev R. Using Centroid Keywords and Word Mover’s Distance for Single Document Extractive Summarization // ACM International Conference Proceeding Series, 3rd International Conference on Natural Language Processing and Information Retrieval, NLPIR 2019, -Tokushima, Japan, 2019. –С. 149-152. // DOI:10.1145/3342827.3342852: 21.10.2019
103. Мусабаев Р.Р., Сейткали Д. Реферирование текстового документа с помощью word mover’s distance и извлеченных ключевых слов документа // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 247-253
104. Tolegen G., Toleu A., Mussabayev R. Мoted perceptron approach for morphological disambiguation // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч.2. -С. 68-74
105. Мусабаев Р.Р., Козбагаров О.Б., Касымжанов Б.К., Мукашев А.Ш., Ибраева В.М., Меркебаев А. Обучения дистрибутивных моделей для предобработки текстов в WORD2VEC и GLOVE // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч.1. -С. 405-414
106. Dragomir R. Radev, Hongyan Jing, Malgorzata Stys, and Daniel Tam. Centroid-based summarization of multiple documents // Inf. Process. Manage. - 2004. – С. 919-938
107. Pereira T., Aloise D., Brimberg J., Mladenović N. Review of Basic Local Searches for Solving the Minimum Sum-of-Squares Clustering Problem // In: Pardalos P., Migdalas A. (eds) Open Problems in Optimization and Data Analysis. Springer Optimization and Its Applications, Springer, Cham. -2018. –Vol. 141. –С. 249-270
108. Mladenovic N., Alkandari A., Pei J., Todosijevic R., Pardalos P.M. Less is more approach: basic variable neighborhood search for the obnoxious p-median problem // International transactions in operational research, Intl. Trans. in Op. Res. -2019. -№00. –С.1–14
109. Cadi AA. El., Ratli M., Mladenović N. New MIP model for multiprocessor scheduling problem with communication delays // In: Pardalos P., Migdalas A. (eds) Open Problems in Optimization and Data Analysis. -Springer Optimization and Its Applications, -Springer, Cham. – 2019. -Т. 141. –С. 129-149
110. Brimberg J., Mladenovic N., Todosijevic R., Uroševic D. A non-triangular hub location problem // Optimization Letters, Springer Verlag, - 2019. –C. 1-20
111. Pei J., Drazic Z., Drazic M., Mladenovic N., Pardalos P.M. Continuous Variable Beighborhood Search (C-VNS) for Solving Systems of Nonlinear Equations // Informs journal on computing, 2019. -Т.31(2). -С. 193-410
112. Ivanov S. V., Kibzun A. I., Mladenović, N. Variable Neighborhood Search for a Two-Stage Stochastic Programming Problem with a Quantile Criterion. Automation and Remote Control // Automation and Remote Control, 2019. -№ 80(1), -С. 43–52
113. Ivanov S. V., Kibzun A. I., Mladenović, N., Urosevic D. Variable neighborhood search for stochastic linear programming problem with quantile criterion // Journal of global optimization, --2019. Т. 74(3). -С. 549-564
114. Mussabayev R., Kassymzhanov B., Mukashev A., Ibrayeva V., Merkebayev A. Creation of necessary technical and expert-analytical conditions for development of the information system of evaluating open text information sources' influence on society // Proceedings 15th International Asian School-Seminar "Optimization Problems of complex systems", - Novosibirsk, Russia, 2019.

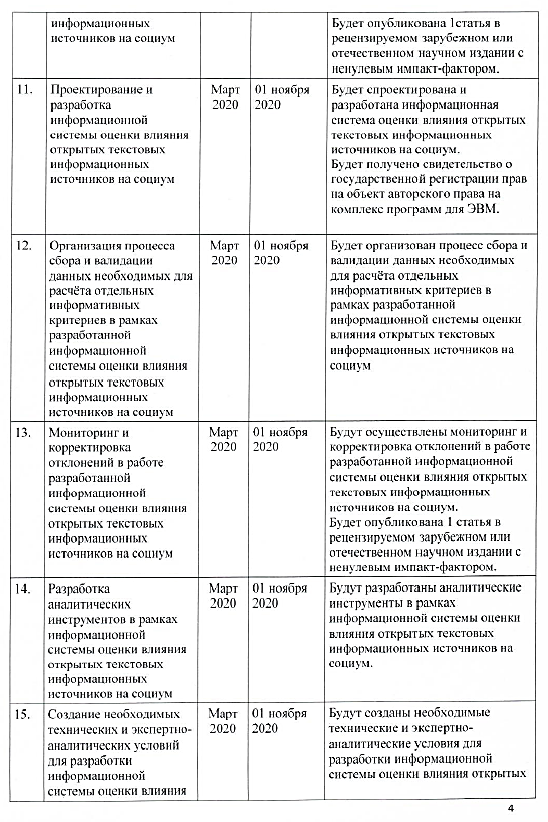
# ПРИЛОЖЕНИРЕ А

# Календарный план работы











# ПРИЛОЖЕНИЕ Б

# Список публикаций

1. Toleu A., Tolegen G., Mussabayev R.: KeyVector Unsupervised Keyphrase Extraction Using Weighted Topic via Semantic Relatedness // Computación y Sistemas, 2019. -Vol. 23(3). -P. 861–869 // doi: 10.13053/CyS-23-3-3264 : 23.10.2019 (Impact Factor – 0.53)
2. Shalkarbayuli A., Kairbekov A., Amangeldi Y. Comparison of traditional machine learning methods and Google services in identifying tonality on Russian texts//Journal of Physics: Conference Series. -2018. -Vol. 1117(1), -P.1-9 (Impact Factor = 0.48)
3. Barakhnin V. B., Duisenbayeva A. N., Kozhemyakina O.Yu., Yergaliyev Y. N. and Muhamedyev R. I. The automatic processing of the texts in natural language. Some bibliometric indicators of the current state of this research area//Journal of Physics: Conference Series. -2018. // <https://doi.org/10.1088/1742-6596/1117/1/012001> : 21.10.2019 (Impact factor = 0.48)
4. Pereira T., Aloise D., Brimberg J., Mladenović N. Review of Basic Local Searches for Solving the Minimum Sum-of-Squares Clustering Problem // In: Pardalos P., Migdalas A. (eds) Open Problems in Optimization and Data Analysis. Springer Optimization and Its Applications, Springer, Cham, 2018. -Vol 141. –P. 249-270 // <https://doi.org/10.1007/978-3-319-99142-9_13>: 21.10.2019 (Impact Factor – 0,31)
5. Cadi AA. El., Ratli M., Mladenović N. New MIP model for multiprocessor scheduling problem with communication delays // In: Pardalos P., Migdalas A. (eds) Open Problems in Optimization and Data Analysis. Springer Optimization and Its Applications. Springer, Cham, -2018. –Vol. 141. –P. 129-149 <https://doi.org/10.1007/978-3-319-99142-9_8> (Impact Factor – 0,31)
6. Brimberg J., Mladenovic N., Todosijevic R., Uroševic D. A non-triangular hub location problem // Optimization Letters, Springer Verlag, 2019. –P. 1-20 // <https://doi.org/10.1007/s11590-019-01392-2> : 21.10.2019 (Impact Factor – 1,399)
7. Pei J., Drazic Z., Drazic M., Mladenovic N., Pardalos P.M. Continuous Variable Beighborhood Search (C-VNS) for Solving Systems of Nonlinear Equations // Informs journal on computing. -2018. -Vol.31(2). -P. 193-210. // <https://doi.org/10.1287/ijoc.2018.0876> : 21.10.2019 (Impact Factor – 2.39)
8. Mladenovic N., Alkandari A., Pei J., Todosijevic R., Pardalos P.M. Less is more approach: basic variable neighborhood search for the obnoxious p-median problem // International transactions in operational research, Intl. Trans. in Op. Res. -2019. –Vol.00. –P. 1–14 // <https://doi.org/10.1111/itor.12646> :21.10.2019 (Impact Factor – 2.51)
9. Ivanov S. V., Kibzun A. I., Mladenović, N. Variable Neighborhood Search for a Two-Stage Stochastic Programming Problem with a Quantile Criterion. Automation and Remote Control // Automation and Remote Control. -2019. -Vol. 80(1), -P. 43–52. // <https://doi:10.1134/s0005117919010041> : 21.10.2019 (Impact Factor – 0.86)
10. Ivanov S. V., Kibzun A. I., Mladenović, N., Urosevic D. Variable neighborhood search for stochastic linear programming problem with quantile criterion // Journal of global optimization, 2019. -Vol. 74(3). –P.549-564 // <https://doi.org/10.1007/s10898-019-00773-2> : 21.10.2019 (Impact Factor – 1.631)
11. A.M. Krassovitsky, I.M. Ualiyeva, Zh. Meirambekkyzy, R.R. Mussabayev. In the search of a linguistic model for automatic generalization recognition in media texts // International Journal of Advances in Electronics and Computer Science, 2019. Vol. 6(3). PP.38-43, (JIFACTOR – 2.68)
12. Красовицкий А.М., Уалиева И.М., Мейрамбеккызы Ж., Мусабаев Р.Р. Lexicon-based approach in generalization evaluation in russian language media // Международный научный журнал «Современные информационные технологии и ИТ образование». -2018. -Т. 14(3). -С. 563-567 (Импакт-фактор РИНЦ - 0,229)
13. Мухамедиев Р.И., Сымагулов А., Кучин Я., Абдуллаева С., Абдолдина Ф.Н. Облачные сервисы для обработки текстов на естественном языке // Международный научный журнал «Современные информационные технологии и ИТ образование». – 2018. – Т.14(4). – С.859-869 (Импакт-фактор РИНЦ - 0,229)
14. Барахнин В.Б., Кожемякина О.Ю., Рычкова Е.В., Пастушков И.С., Борзилова Ю.С. Извлечение лексических и метроритмических признаков, характерных для жанра и стиля и их комбинаций в процессе автоматизированной обработки текстов на русском языке // Международный научный журнал «Современные информационные технологии и ИТ образование» – 2018. – Т.14(4). – С.876-883. // <http://dx.doi.org/10.25559/SITITO.14.201803.876-883> : 21.10.2019 (Импакт-фактор РИНЦ 2017- 0,229)
15. Барахнин В.Б., Кожемякина О.Ю., Рычкова Е.В., Борзилова Ю.С. Автоматическое выделение словосочетаний из текстов славянского происхождения: сравнение подходов // Cloud of Science. – 2018. – Т.5(4). –  С.713-728 (Импакт-фактор РИНЦ 0.533)
16. Нурумов К.С., Атанаева М.К., Булдыбаев Т.К., Оспанова У.А., Акоева И.Г. Комплексная оценка значимости информативных признаков для оценки влияния на социум на основе анализа публикуемой текстовой информации // Вестник ЕНУ им. Л.Н. Гумилева, Серия Журналистика, -2018. -№3 (124). -С. 44-52
17. Атанаева М.К., Оспанова У.А., Булдыбаев Т.К., Акоева И.Г., Нурумов К.С. Определение индикаторов оценки достоверности публикаций в СМИ для разработки методики определения степени вероятной достоверности отдельной публикации // Вестник КазНПУ, Серия филологическая, -2018. -№2 (64). -С. 158-165
18. Атанаева М.К., Оспанова У.А., Булдыбаев Т.К., Акоева И.Г., Нурумов К.С. Использование лингвистических подходов в определении классов и информативных признаков в текстах // Вестник КазНПУ, Серия филологическая, 2018. -№2 (64). -С. 166-171
19. Оспанова У.А., Атанаева М.К., Акоева И.Г., Булдыбаев Т.К. Identification of features of probable reliability of news // Современное педагогическое образование. – Москва, 2019. -№6. -С. 44-49
20. Атанаева М.К., Булдыбаев Т.К., Оспанова У.О., Акоева И.Г., Нурумов К.С., Баймаханбетов М.А. Определение закономерностей подозрительности в новостных текстах // Вестник ЕНУ им. Л.Н. Гумилева, Серия журналистика. -Нур-Султан, 2019. -№2 (127). -С 8-19
21. Оспанова У.А., Атанаева М.К., Акоева И.Г., Булдыбаев Т.К. Informative features of bias and reliability of electronic Mass Media // Социология. -Москва, 2019. -№2. -C. 259-269
22. Атанаева М.К., Булдыбаев Т.К., Оспанова У.А., Акоева И.Г., Нурумов К.С., Баймаханбетов М.А. Методика для определения информативных признаков новостных текстов и проверка их значимости // Научный аспект. -2019. -Т.3(3). -С. 277-296 (ИФ РИНЦ 0,03)
23. Захаров В.Н., Мусабаев Р.Р., Красовицкий А.М., Козловская Я.Д., Хорошилов А-р А., Хорошилов А-й А. Метод кластеризации новостных сообщений средств массовой информации на основе их концептуального анализа // Системы и средства информатики, 2019. -Т.29 (3). -С. 52-66
24. Seitkali D., Mussabayev R. Using Centroid Keywords and Word Mover’s Distance for Single Document Extractive Summarization // ACM Intern. Conf. Proc. Series, 3rd Inter. Conf. on NLP and Information Retrieval, NLPIR 2019. -Tokushima, Japan, 2019. –P. 149-152 (Scopus)
25. Барахнин В.Б., Дуйсенбаева А., Кожемякина О.Ю., Кучин Я.И., Якунин К.О., Мухамедиев Р.И. The studies of dinamics’ changes in publication activity in the field of natural language processing// Матер. 17-ой Междунар. науч. конф. "Информационные Технологии и Менеджмент 2019". - Рига, Латвия, 2019. -С. 71-74
26. Мусабаев Р.Р., Мухамедиев Р.И., Кучин Я.И., Сымагулов А., Якунин К.О. On a method of multimedial media ranking using corpus based topic modeling// Матер. 17-ой Междунар. науч. конф. "Информационные Технологии и Менеджмент 2019". - Рига, Латвия, 2019. -С. 64-67
27. Mladenovic N., Krassovitskiy A., Mussabayev R. Decomposition approach for clustering problem // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 15-21
28. Nurumov K., Baimakhanbetov M., Buldybayev T., Akoyeva I., Ospanova U. An approach for detecting sentiments in mass media publications using content specific Informative features // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 21-31
29. Toleu A., Tolegen G., Mussabayev R. Initial exploration for Topic modeling with word sense // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 31-44
30. Барахнин В., Дуйсенбаева А., Кожемякина О., Кучин Я., Якунин К., Мухамедиев Р. Изменение публикационной активности в области обработки естественного языка // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 130-135
31. Мейрамбеккызы Ж., Аманбай А., Мусабаев Р. Автоматическое формирование тематических словарей социально-значимых понятий // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 233-239
32. Мусабаев Р., Мухамедиев Р., Кучин Я., Сымагулов А., Якунин К., Мурзахметов С. Метод мультимодальной оценки средств массовой информации на основе тематической модели корпуса текстов // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 239-247
33. Мусабаев Р.Р., Сейткали Д. Реферирование текстового документа с помощью word mover’s distance и извлеченных ключевых слов документа // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 247-253
34. Уалиева И.М., Мусабаев Р.Р. Исследование признаков для автоматической идентификации жанров публикаций в СМИ // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 331-336
35. Хорошилов А.А., Мусабаев Р.Р., Красовицкий А.М., Уалиева И.М., Мейрамбеккызы Ж. Автоматическое установление тональности текстов методом их концептуального анализа // Матер. науч. конф. ИИВТ МОН РК «Современные проблемы информатики и вычислительных технологий», - Алматы, - 2019. -С. 365-375
36. Meirambekkyzy Zh., Mussabaev R.R. Automatic development of topic dictionaries of socially significant concepts // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 1. -С. 65-70
37. Buldybayev T., Baimakhanbetov M., Nurumov K., Akoyeva I., Ospanova U., Atanayeva M. Analysis of digital media: a comprehensive assessment of informative features significance // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 31-44
38. Tolegen G., Toleu A., Mussabayev R. Мoted perceptron approach for morphological disambiguation // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 68-74
39. Ualiyeva I.M., Mussabayev R.R. Features analysis for detecting information-dense texts in online news media // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 74-79
40. Барахнин В.Б., Мухамедиев Р.И., Мусабаев Р.Р., Кожемякина О.Ю., Исаева А., Кучин Я.И., Мурзахметов С.Б., Якунин К.О. Методы выявления деструктивных новостей в медиапространстве // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 205-219
41. Козбагаров О.Б., Мусабаев Р.Р. Кластеризация новостных публикаций по информационным поводам // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 340-350
42. Мусабаев Р.Р., Козбагаров О.Б., Касымжанов Б.К., Мукашев А.Ш., Ибраева В.М., Меркебаев А. Обучения дистрибутивных моделей для предобработки текстов в WORD2VEC и GLOVE // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 405-414
43. Мухамедиев Р.И., Мусабаев Р.Р., Булдыбаев Т., Кучин Я., Сымагулов А., Оспанова У., Якунин К., Мурзахметов С., Сагындык Б. Эксперименты по оценке средств массовой информации на основе тематической модели корпуса текстов // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 2. -С. 419-431
44. Младенович Н., Красовицкий А., Мусабаев Р. Метод декомпозиции в задаче кластеризации больших данных // Матер. IV Междунар. науч.-практич. конф. «Информатика и прикладная математика", посвященной 70-летнему юбилею проф. Биярова Т.Н., В. Вуйцика и 60-летию проф. Амиргалиева Е.Н. - Алматы, Казахстан, 2019. -Ч. 1. -С. 525-533
45. Баймаханбетов М. А., Нурумов К. С., Оспанова У. А., Булдыбаев Т. К., Акоева И. Г. Влияние метода ADASYN на широко распространенные метрики эффективности машинного обучения // Современные информационные технологии и ИТ-образование, -2019. -Т.15(2). –С. 290-297
46. Атанаева М.К., Булдыбаев Т.К., Оспанова У.А., Акоева И.Г., Нурумов К.С., Баймаханбетов М.А. Определение тональности и объективности новостных текстов словарным подходом // Научный аспект, -2019. -Т.3(3). -С.296-308 (ИФ РИНЦ 0,03)
47. А.с. № 2894. NLP-Preprocessor / Якунин К., Мухамедиев Р., Кучин Я., Мурзахметов С., Сымагулов А., Мустакаев Р., опубл. 18.04.2019. -1 с.
48. А.с. №4523. Программа мультимодальной оценки СМИ на базе тематической модели корпуса текстов (Multi Modal Mass Media Assesment – M4A)/ Якунин К.О., Мухамедиев Р.И., Кучин Я.И., Мурзахметов С.Б., Сымагулов А., Мусабаев Р.Р.; опубл. 15.07.2019. -1 с.
49. А.с. № 4505. Программа расчета динамических библиометрических показателей D1 и D2 (Calculation of dynamic scientometric indicators D1, D2 – CalcDStMi)/ Якунин К.О., Мухамедиев Р.И., Кучин Я.И.; опубл. 12.07.2019. – 1с.

# ПРИЛОЖЕНИЕ В

# Список использованных зарубежных информационных ресурсов

* 1. Springer Link // <https://link.springer.com/>
  2. Scopus // <https://www.scopus/>
  3. Web of Science // <https://apps.webofknowledge.com/>
  4. Cornell University Library // <https://arxiv.org/>
  5. РМЭБ // <http://rmebrk.kz/>
  6. Mendeley // <https://www.mendeley.com/>
  7. Cyberleninka // <https://cyberleninka.ru/>
  8. Google Scholar // <https://scholar.google.com/>
  9. ResearchGate // <https://www.researchgate.net/>

## ПРИЛОЖЕНИЕ Г

## Основные параметры текстов СМИ.

Таблица Г1 – Список параметров для классификации резонансных\не резонансных публикаций.

|  |  |
| --- | --- |
| Параметр | Вес (CW) |
| Медиа-охват (количество просмотров) за определённый период | 0.99 |
| Количество репостов информации за определённый период | 0.99 |
| Количество шейров информации за определённый период | 0.577505268 |
| Количество комментариев публикации за определённый период | 0.806507958 |
| Охват тем, вызывающих резонанс | 0.956947071 |
| Тональность комментариев (Негативный/нейтральный/позитивный | 0.653676928 |

Таблица Г2 – Список параметров для классификации социально значимых публикаций.

|  |  |
| --- | --- |
| Параметр | Вес (CW) |
| Наличие/отсутствие ссылки в публикации на адекватный источник | 0.061593752 |
| Ссылка на первоисточник | 0.16683118 |
| Освещение одного и того же события различными изданиями: кросс-чекинг, наличие дискурса с другими статьями | 0.21234652 |
| Аутентичность фото, которое подтверждает событие | 0.142772968 |
| Аутентичность видео, которое подтверждает событие | 0.14244636 |
| Наличие факта последующей правки в уже опубликованной статье | 0.045305297 |
| Соответствие заголовка публикации содержанию медиа текста | 0.108959251 |
| Наличие/отсутствие автора (его имя при наличии) | 0.086821073 |
| Наличие отсутствие проверяемых фактов в статье | 0.202623491 |
| Репутация издания (сайты новостных и информационных агентств, электронных СМИ, иных аналогичных сайтов), на которых опубликована информация | 0.154870571 |
| Медиа-охват (количество просмотров) за определённый период | 0.211239893 |
| Количество репостов информации за определённый период | 0.193185632 |
| Количество шейров информации за определённый период | 0.174316737 |
| Количество комментариев публикации за определённый период | 0.130902894 |
| Тональность публикации (Негативный/нейтральный/позитивный) | 0.069418546 |
| Ненавистнические комментарии (hate detection) | 0.083327509 |
| Политизированность медиа-текста | 0.103780306 |
| Степень генерализации: сильно выраженная, слабо выраженная, нет | 0.063146285 |
| Наличие призыва к действию | 0.140608592 |
| Тональность комментариев (Негативный/нейтральный/позитивный) | 0.087363666 |
| Значимость для социально активных групп | 0.137737477 |

Таблица Г3 – Список параметров для классификации на достоверный\не достоверный.

|  |  |
| --- | --- |
| Параметр | Вес (CW) |
| Наличие/отсутствие ссылки в публикации на адекватный источник | 0.363788658 |
| Ссылка на первоисточник | 0.359066664 |
| Освещение одного и того же события различными изданиями: кросс-чекинг, наличие дискурса с другими статьями | 0.535835773 |
| Аутентичность фото, которое подтверждает событие. | 0.205189785 |
| Аутентичность видео, которое подтверждает событие. | 0.196834318 |
| Наличие факта последующей правки в уже опубликованной статье. | 0.1662592 |
| Соответствие заголовка публикации содержанию медиа текста | 0.178352836 |
| Наличие/отсутствие автора (его имя при наличии) | 0.1154985 |
| Наличие/отсутствие проверяемых фактов в статье | 0.340389563 |
| Репутация издания *(сайты новостных и информационных агентств, электронных СМИ, иных аналогичных сайтов)*, на которых опубликована информация | 0.463591703 |

Таблица Г4 – Список параметров для классификации на объективный\не объективный.

|  |  |
| --- | --- |
| Параметр | Вес (CW) |
| Наличие в медиа-тексте выраженного личного мнения автора публикации | 0.72462313 |
| Наличие манипуляторных техник в медиа-тексте | 0.99 |
| Степень генерализации: слабо выраженная, сильно выраженная, нет | 0.478821031 |
| Политизированность | 0.99 |
| Тональность публикации. (Негативный/нейтральный/позитивный). | 0.867781263 |
| Тональность комментариев (негативный, нейтральный, позитивный) | 0.628555507 |

# ПРИЛОЖЕНИЕ Д

# Эксперименты по применимости алгоритма мультимодального оценивания СМИ

Эксперимент №1

С целью оценки применимости описанного алгоритма был сформирован очень небольшой корпус документов, условно распределенный между двумя СМИ. Для описанного корпуса были рассчитаны матрицы p1, …, p6 c применением меры Жаккара, довольно часто применяемой как мера близости документов в алгоритмах кластеризации. Кратко исходные данные для работы алгоритма следующие:

Корпус, состоящий из 5 статей (m=5), следующей направленности:

1. Экономика
2. Спорт(бокс)
3. Политика
4. Шоу бизнес
5. Образование и наука

Указанный корпус текстов разбит вручную на 4 тематических кластера (k=4)

1. Политика
2. Спорт
3. Шоу бизнес
4. Экономика и финансы

Рассматриваются два класса статей (c=2):

1. Социально значимый
2. Объективный

и два СМИ.

Определены 4 признака (q=4), для каждого из которых определен свой словарь терминов [40].

1. Манипулятивность
2. Политизированность
3. Негативная тональность
4. Позитивная тональность

Рассчитана матрица условных распределений вероятностей тематик по статьям в виде матрицы p2[1..k][1..m], где k- количество тематик (построчно), m- статей (по столбцам)

p2 = [[0.03, 0.0, 0.03, 0.004, 0.005],

[0.0, 0.01, 0.0, 0.008, 0.005],

[0.0, 0.007, 0.004, 0.03, 0.005],

[0.04, 0.0, 0.007, 0.0, 0.005]]

С использованием аналитического иерархического процесса (AHP) получены веса каждого признака для классов в виде матрицы p3[1..с][1..q], где с – классы, q- признаки

p3 = [[0.55, 0.27, 0.18, 0.18],

[0.23, 0.43, 0.34, 0.34]]

Получены следующие результаты расчета матриц p4[1 … k][1 … c] (условные вероятности распределения тематик по классам), p5 [1 ... m][1 ... c] (условные вероятности распределения документов по классам), p6 [1..m][1..q] (условные вероятности распределения документов по признакам):

p4= [[0.1967 0.2239]

[0.0371 0.0531]

[0.0316 0.0508]

[0.0427 0.0443]]

p5= [[0.007609 0.008489 ]

[0.0005922 0.0008866]

[0.0063263 0.0072303]

[0.0020316 0.0028444]

[0.0015405 0.0018605]]

p6= [[0.0058 0.0149 0.0014 0.0008 ]

[0.00027 0.00017 0.00055 0.00166]

[0.00452 0.01329 0.00094 0.00046]

[0.00102 0.0021 0.00174 0.00328]

[0.00105 0.0025 0.00055 0.00105]]

На основании приведенных значений с применением алгоритма мультимодальной оценки СМИ рассчитаны распределения СМИ по тематикам, классам и признакам в предположении, что имеющиеся статьи предоставлены двумя СМИ, первое из которых опубликовало статьи номер 1 и 3 (экономической и политической направленности), второе является источником статей 2,4,5 (“спорт(бокс)”, “шоу бизнес”, “образование и наука”).

Таблица Д1 – Распределение СМИ по тематикам

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СМИ\Тематики: | Политика | Спорт | Шоу бизнес | Экономика и финансы | Статьи |
| СМИ0 | 0.971327 | 0.057346 | 0.431078 | 0.92300 | [экономическая, политическая] |
| СМИ1 | 0.458623 | 0.927009 | 0.969495 | 0.06101 | [спорт (бокс), шоу бизнес, образование и наука] |

|  |  |  |  |
| --- | --- | --- | --- |
| СМИ\Классы: | Социально значимый | Объективный | Статьи |
| СМИ0 | **0.597028** | **0.701486** | [экономическая, политическая] |
| СМИ1 | **0.460911** | **0.769545** | [спорт (бокс), шоу бизнес, образование и наука] |

Таблица Д2 – Распределение СМИ по классам

Таблица Д3 – Распределение СМИ по признакам

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СМИ/  Признаки: | Манипулятив-ность | Политизирован-ность | Негативная  тональность | Позитивная  тональность | Статьи |
| СМИ0 | **0.675538** | **0.917128** | **0.317990** | **0.165743** | [экономическая, политическая] |
| СМИ1 | **0.273881** | **0.496645** | **0.392071** | **0.863059** | [спорт (бокс), шоу бизнес, образование и наука] |

Эксперимент на «игрушечном» примере показал логичные результаты. То есть статьи соответствующей тематической направленности определили распределение вероятностей гипотез по тематикам, классам и параметрам для двух условных СМИ. Видно, что СМИ0 и СМИ1 отличаются по объективности, манипулятивности и тональности. Например, СМИ0 «пишет» в области экономики и политики, значительно менее позитивно, более политизировано и применяет манипулятивные техники больше, чем СМИ1.

Эксперимент №2

Для второго эксперимента был использован корпус статей новостного агентства Tengrinews (https://tengrinews.kz/), состоящий из 20000 статей за период с 2014 по 2019 годы. Цель эксперимента заключалась в сравнительном анализе публикаций во времени. Статьи были отсортированы по времени и сформированы 5 массивов публикаций: 2014 и 2015 годы, 2015, 2017, 2018, 2019. Набор классов, параметров и словарей остался таким же, как в эксперименте 1. Результаты эксперимента показаны в таблицах 4 и 5

Таблица Д4 – Распределение публикаций по классам

|  |  |  |
| --- | --- | --- |
| СМИ: | Социально значимый | Объективный |
| 2014, 2015 | **0.001396** | **0.987428** |
| 2016 | **0.000669** | **0.993975** |
| 2017 | **0.000542** | **0.995114** |
| 2018 | **0.000719** | **0.993520** |
| 2019 | **0.006352** | **0.942828** |

Таблица Д5 – Распределение СМИ по признакам

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| СМИ: | Манипулятивность | Политизированность | Негативная  тональность | Позитивная  тональность |
| 2014, 2015 | 0.00040 | 0.31335 | 0.99633 | 0.61259 |
| 2016 | 0.00019 | 0.30883 | 0.99826 | 0.63196 |
| 2017 | 0.00015 | 0.31287 | 0.99856 | 0.66602 |
| 2018 | 0.00021 | 0.31293 | 0.99810 | 0.69037 |
| 2019 | 0.00193 | 0.29577 | 0.98261 | 0.70545 |

Несмотря на весьма небольшие словари и ограниченный набор параметров эксперимент выявил постепенное увеличение позитивной тональности публикаций новостного портала во времени и некоторое возрастание социально-значимого контента за последние годы.

Эксперимент 3

Для третьего эксперимента использован набор статей новостного агентства Tengri News (<https://tengrinews.kz/>) состоящий из 1800 статей, размеченных экспертами, что позволило сравнить результаты работы алгоритма с оценками экспертов.

Использованы следующие словари признаков: Факты (657 лексических единиц), Негативная тональность (2344), Позитивная тональность (1207), Политизированность (35), Призыв к действию (145), Манипулятивность (1072). С помощью экспертных оценок сформирована матрица p3 (важность параметров для классов):

[[0.23, 0.18, 0.18, 0.2 , 0.21, 0. ],

[0. , 0.22, 0.22, 0.26, 0. , 0.3 ],

[1. , 0. , 0. , 0. , 0. , 0. ]],

где колонки (слева на право) соответствуют перечисленным выше признакам, а строки (сверху вниз) классам: "Социально значимый", "Объективный", "Достоверный". Отметим, что данная матрица не полна. Не все признаки для соответствующих классов определены. Ноль означает, что данный признак не используется для определения вероятности p5. Например, видно, что социальная значимость текста определяется по пяти признакам, в то время как достоверность с использованием только одного (факты).

На рисунке 3 приведены таблицы, где отображаются 5 новостей с наибольшей и 5 с наименьшей вероятностью соответствия параметру или классу (aggregator score). Оценки экспертов приведены в колонке Expert Score.

|  |
| --- |
|  |
|  |
|  |

Рисунок Д1 – Оценки политизированности, объективности и достоверности текстов СМИ

Результаты применения алгоритма показывают, что в некоторых случаях он дает хорошую степень совпадения с экспертными оценками. С другой стороны, он позволяет ранжировать тексты и в ситуации, когда тексты весьма близки по своим характеристикам. Приведенные выше тексты взяты из одного достаточно надежного источника, поэтому достоверность и объективность текстов близки. Вместе с тем данный подход требует весьма тщательного составления словарей и применения дополнительных алгоритмов для определения более широкого круга параметров текстов.

# ПРИЛОЖЕНИЕ Е

# Результаты ранжирования тематик по методу 1.3

(указаны номера-идентификаторы топиков которые соответствуют указанным в файле Кластер №NNN):

167, 135, 195, 68, 188, 107, 178, 164, 144, 166, 86, 203, 192, 134, 11, 95, 184, 29, 190, 75, 176, 139, 111, 67, 45, 13, 146, 53, 66, 185, 43, 14, 78, 89, 34, 197, 104, 152, 101, 7, 202, 85, 159, 149, 96, 172, 74, 28, 57, 133, 171, 102, 162, 1, 55, 120, 64, 12, 147, 150, 4, 76, 3, 105, 108, 191, 50, 154, 187, 2, 126, 193, 168, 124, 40, 87, 163, 8, 119, 186, 56, 6, 194, 23, 17, 71, 79, 99, 42, 180, 25, 37, 93, 30, 112, 138, 125, 80, 109, 59, 92, 0, 110, 122, 35, 173, 77, 136, 116, 129, 198, 151, 19, 63, 114, 100, 118, 15, 182, 62, 137, 88, 24, 196, 44, 5, 73, 91, 200, 16, 27, 84, 174, 60, 36, 21, 98, 201, 128, 183, 148, 123, 20, 38, 127, 65, 83, 177, 81, 61, 103, 170, 157, 70, 94, 181, 48, 54, 51, 117, 132, 106, 49, 46, 39, 121, 9, 47, 31, 199, 160, 142, 97, 141, 161, 69, 189, 22, 26, 145, 130, 90, 10, 155, 18, 179, 33, 72, 32, 113, 143, 169, 115, 58, 131, 140, 175, 156, 158, 52, 41, 153, 165, 82

Результаты ранжирования тематик по методу 2.3:

202, 164, 188, 203, 134, 192, 107, 184, 10, 153, 100, 198, 29, 187, 59, 171, 143, 64, 3, 199, 176, 157, 155, 138, 50, 92, 123, 167, 108, 96, 136, 173, 170, 126, 194, 69, 97, 135, 49, 65, 190, 162, 93, 104, 141, 196, 174, 148, 23, 83, 68, 12, 201, 44, 45, 189, 163, 13, 183, 105, 41, 177, 168, 0, 161, 160, 9, 118, 142, 175, 150, 11, 25, 73, 28, 2, 182, 129, 144, 180, 89, 109, 8, 26, 106, 66, 127, 120, 165, 18, 131, 85, 1, 116, 27, 16, 80, 57, 88, 140, 152, 63, 75, 169, 159, 14, 110, 34, 102, 156, 151, 17, 115, 53, 181, 42, 191, 40, 81, 99, 172, 137, 39, 166, 101, 112, 128, 35, 111, 52, 48, 186, 32, 72, 122, 76, 178, 125, 15, 19, 149, 95, 90, 91, 193, 185, 36, 70, 87, 103, 117, 98, 94, 7, 5, 51, 31, 33, 119, 21, 200, 79, 158, 67, 82, 60, 121, 147, 56, 43, 77, 62, 47, 30, 145, 130, 24, 71, 58, 139, 132, 74, 55, 78, 86, 133, 37, 46, 179, 61, 113, 38, 114, 154, 146, 84, 124, 6, 4, 195, 22, 197, 20, 54

Ранжирование осуществлялось по весу тематик, которые указаны в файле сразу после номера кластера (тематики), например, Кластер №28 0.44

В данном случае 28 - это идентификатор кластера, а 0.44 его вес (степень согласованности со всем корпусом).

Как видно по первому методу наибольший вес имеет тематика №167:

Кластер №167 1.0

<< ДРУЖБА 0.28, БЛАГОПОЛУЧИЕ 0.25, ПОЖЕЛАНИЕ 0.24 >>

начинание 0.11, импульс 0.1, терпение 0.1, сострадание 0.05, созидание 0.08, гостеприимство 0.08, радость 0.2, становление 0.17, единение 0.04, построение 0.07, процветание 0.23, единство 0.19, елбасы 0.03, проявление 0.09, забота 0.15, устремление 0.05, мудрость 0.18, великое 0.12, нуждающийся 0.0, обретение 0.11, добрососедство 0.1, углубление 0.13, патриотизм 0.16, дружба 0.28, богатство 0.15, спокойствие 0.13, признательность 0.14, значимость 0.14, упрочение 0.13, союзничество 0.07, пожелание 0.24, воля 0.18, этнос 0.07, свершение 0.13, поздравление 0.19, взаимоотношение 0.17, телеграмма 0.09, плечо 0.12, служение 0.05, стремление 0.23, юбилей 0.15, контекст 0.11, благополучие 0.25, абиш 0.15, аспект 0.15, достаток 0.09, интеллигенция 0.06, гордость 0.19, видение 0.13

Следующая по значимости №135:

Кластер №135 0.91

<< ПРОГУЛКА 0.27, УГОЛОК 0.27, ДОСТОПРИМЕЧАТЕЛЬНОСТЬ 0.27 >>

пляж 0.21, островок 0.15, долина 0.14, рай 0.13, узор 0.09, созерцание 0.04, серфинг 0.12, пальма 0.13, нурсаин 0.17, душ 0.04, столетие 0.17, шопинга 0.04, шармэльшейх 0.02, бассейн 0.13, пхукет 0.06, хайнань 0.12, ревматизм 0.0, доминикана 0.04, врата 0.11, купол 0.04, ландшафт 0.21, наследие 0.1, белые 0.16, темнота 0.13, времяпровождение 0.09, пологий 0.09, редкость 0.15, перевал 0.07, поход 0.16, китаец 0.13, подборка 0.17, оазис 0.12, руина 0.13, холм 0.1, фотограф 0.12, улочка 0.13, равнина 0.09, простор 0.18, галька 0.02, экскурсия 0.22, шезлонг 0.06, массаж 0.06, каникулы 0.11, водопад 0.2, троп 0.08, ум 0.15, бунгало 0.07, обилие 0.19, дайвинг 0.17, башня 0.15, континент 0.12, бали 0.1, достопримечательность 0.27, расположение 0.14, эпоха 0.16, окрестность 0.13, досуг 0.11, гид 0.13, каньон 0.11, легенда 0.2, чудо 0.16, магазинчик 0.08, рельеф 0.11, материк 0.09, подножие 0.22, рассказ 0.15, колорит 0.14, лежак 0.12, закат 0.17, городок 0.12, аквапарк 0.04, уникальность 0.15, хребет 0.14, растительность 0.09, развалина 0.08, древние 0.12, развлечение 0.23, романтика 0.09, скала 0.16, туристка 0.04, караван 0.08, ширина 0.14, сувенир 0.12, знакомство 0.14, бухта 0.08, великое 0.15, богатство 0.17, ощущение 0.18, сочетание 0.14, уголок 0.27, деревушка 0.11, вьетнам 0.08, суета 0.13, арка 0.15, интерьер 0.13, хургад 0.06, воображение 0.1, зонтик 0.12, трава 0.12, тайланд 0.04, прелесть 0.11, прогулка 0.27, джунгли 0.17, степь 0.13, лестница 0.14, хижина 0.05, зрелище 0.13, удовольствие 0.18, разнообразие 0.2, лето 0.18, лавка 0.08, уединение 0.06, песок 0.19, пустынь 0.14, пустыня 0.14, лагуна 0.14, когдато 0.19, путешественник 0.2, пейзаж 0.25, верблюд 0.12, ривьера 0.05

А наименее значимой является тематика №82:

Кластер №82 0.15

<< ТУРАГЕЛЬДИЙ 0.43, САМОВЫДВИЖЕНЕЦ 0.42, МЕЛЬДЕШОВ 0.37 >>

гульдана 0.02, мельдешов 0.37, самовыдвиженец 0.42, мусагали 0.26, самовыдвижение 0.21, жаксыбай 0.08, койшиева 0.29, фоос 0.26, госязык 0.16, отен 0.2, лиман 0.18, центризбирком 0.33, кайсаров 0.14, токбаев 0.29, жолдасбек 0.22, дуамбек 0.25, турагельдий 0.43, мукаша 0.07, цик 0.13, масино 0.19, есжан 0.0, амантайкаж 0.34, амантаякаж 0.21, маслихат 0.2, базильбай 0.23, сулеймена 0.09, салим 0.04, турганкул 0.21, турганкулов 0.2, мырзатай 0.24, тургун 0.19, ксдп 0.08, выдвиженец 0.14, выдвижение 0.24

По второму же методу взвешивания наиболее весомой является 202 тематика:

Кластер №202 1.0

<< БЛАГОУСТРОЙСТВО 1.0, БАЙБЕК 0.84, ОСВЕЩЕНИЕ 0.75 >>

освещение 0.75, облик 0.45, обустройство 0.24, передвижение 0.17, есим 0.29, объезд 0.4, застройка 0.35, малый 0.16, удобство 0.23, байбек 0.84, документация 0.29, неудобство 0.18, проектирование 0.24, тротуар 0.64, исекешева 0.46, благоустройство 1.0, расположение 0.16, строитель 0.31, архитектор 0.72, фасад 0.69

Следующая по значимости 164:

<< СКИДКА 1.0, НАБОР 0.95, ВЕЩЬ 0.91 >>

коробка 0.67, ассортимент 0.44, нюанс 0.28, акцент 0.39, карман 0.76, уголок 0.33, обувь 0.89, прилавок 0.37, вещь 0.91, скидка 1.0, вовторое 0.57, аналог 0.53, понятие 0.56, прогулка 0.51, четверть 0.37, удобство 0.5, низка 0.42, разнообразие 0.31, лето 0.6, мебель 0.74, набор 0.95, расположение 0.47, предпочтение 0.59, досуг 0.3, копейка 0.23

А наименее значимой является тематика №54:

Кластер №54 0.08

<< РЫБА 1.0, ИКРА 0.49, РЫБАК 0.24 >>

вереница 0.0, кальмар 0.03, гост 0.02, траулер 0.02, пыльца 0.0, брюхо 0.02, промысл 0.02, сазан 0.04, годность 0.03, фекалия 0.03, белуга 0.05, дилдахмет 0.01, спиннинг 0.01, насос 0.03, биоресурсы 0.01, придаток 0.02, амазонка 0.01, ферма 0.18, сом 0.02, полтонны 0.01, консервант 0.06, улов 0.07, колесико 0.01, пиранья 0.05, сарыс 0.02, мальков 0.05, тасмания 0.0, отросток 0.01, водоросль 0.02, обитатель 0.06, эпидемиология 0.0, личинка 0.04, сытник 0.01, живность 0.01, нерест 0.03, аммиак 0.03, молодь 0.03, выдра 0.01, боцман 0.01, берингов 0.01, зайсан 0.02, лов 0.08, рыбак 0.24, рыбалка 0.12, рыба 1.0, ловля 0.06, окунь 0.03, рыбоводство 0.01, рыболов 0.02, рыбка 0.05, вылов 0.04, рыболовство 0.05, неблагополучие 0.01, балык 0.02, плавник 0.03, браконьер 0.08, уба 0.02, трюм 0.01, браконьерство 0.04, ихтиолог 0.01, гильман 0.01, кигач 0.0, снасть 0.01, икра 0.49, технолог 0.01, баночка 0.06, гипофиз 0.06, судак 0.03, коралл 0.02, угорь 0.02, замор 0.01, щука 0.04, жанбай 0.01, урал 0.06, атолл 0.01, особь 0.04, карась 0.04, удочка 0.02, лоток 0.02, трал 0.01, сульфат 0.01, лещ 0.03, разрезание 0.02, акула 0.15, воспроизводство 0.03

В начале каждой тематики в треугольных скобках <<...>> заглавными буквами указан её заголовок, сформированный из первых трёх наиболее весомых для данной тематики слов.

# ПРИЛОЖЕНИЕ Ж

# 

Таблица Ж1 – Основные характеристики дополнительного большого корпуса

|  |  |  |  |
| --- | --- | --- | --- |
| **№ п/п** | **Источник** | **Период** | **Количество публикаций** |
| 1 | [https://365info.kz](https://365info.kz/) | 05.2014-02.2019 | 57 326 |
| 2 | [https://rus.azattyk.org](https://rus.azattyk.org/) | 08.2003-02.2019 | 52 583 |
| 3 | [https://bnews.kz](https://bnews.kz/) | 04.2012-02.2019 | 7 631 |
| 4 | [https://www.caravan.kz](https://www.caravan.kz/) | 07.2001-02.2019 | 246 125 |
| 5 | [http://dailynews.kz](http://dailynews.kz/) | 02.2012-02.2019 | 25 409 |
| 6 | [https://esquire.kz](https://esquire.kz/) | 05.2013-02.2019 | 1 510 |
| 7 | [http://www.exclusive.kz](http://www.exclusive.kz/) | 10.2006-02.2019 | 7 184 |
| 8 | [https://kz.expert](https://kz.expert/) | 01.2017-12.2018 | 671 |
| 9 | [https://forbes.kz](https://forbes.kz/) | 12.2011-02.2019 | 159 755 |
| 10 | [https://www.inform.kz](https://www.inform.kz/) | 01.2010-02.2019 | 341 232 |
| 11 | [https://kapital.kz](https://kapital.kz/) | 09.2007-02.2019 | 73 468 |
| 12 | [https://www.kazpravda.kz](https://www.kazpravda.kz/) | 07.2013-02.2019 | 432 |
| 13 | [http://www.kp.kz](http://www.kp.kz/) | 12.2011-02.2019 | 12 422 |
| 14 | [https://kursiv.kz](https://kursiv.kz/) | 10.2015-02.2019 | 4 253 |
| 15 | [https://www.nur.kz](https://www.nur.kz/) | 08.2018-02.2019 | 18 003 |
| 16 | [https://ru.sputniknews.kz](https://ru.sputniknews.kz/) | 05.2016-01.2019 | 23 406 |
| 17 | [https://tengrinews.kz](https://tengrinews.kz/) | 01.2009-02.2019 | 198 797 |
| 18 | [https://time.kz](https://time.kz/) | 05.2016-02.2019 | 46 557 |
| 19 | [http://today.kz](http://today.kz/) | 12.2011-05.2017 | 102 456 |
| 20 | [http://vesti.kz](http://vesti.kz/) | 10.2008-08.2018 | 27 740 |
| 21 | [https://vlast.kz](https://vlast.kz/) | 04.2012-02.2019 | 21 051 |
| 22 | [https://www.zakon.kz](https://www.zakon.kz/) | Информация отсутствует | 554 422 |
|  | **Всего** |  | **1 982 433** |

Рисунок Ж1 – Сбалансированность состава основного дополнительного корпуса по источникам, %

1. Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida, Ranking news articles based on popularity prediction [↑](#footnote-ref-1)
2. Roja Bandari, Sitaram Asur, Bernardo A., HubermanThe Pulse of News in Social Media: Forecasting Popularity [↑](#footnote-ref-2)