Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К КУРСОВОЙ РАБОТЕ

НА ТЕМУ:

Разработка базы данных для хранения и аналитики результатов статистического опроса

Студент _____ _____ Т. А. Казаева
(Группа) (Подпись, дата) (И.О. Фамилия)

Руководитель курсовой работы _____ Ю. В. Строганов
(Подпись, дата) (И.О. Фамилия)

Москва, 2022 г.

Содержание

Страница

ВВЕДЕНИЕ

1. Аналитический раздел

В разделе описан способ выражения эмоций в тексте и подходы к его анализу. Приведены существующие размеченные корпуса для русского языка и описаны шкалы оценки тональности. Описано существующее решение для англоязычной лексики и рассмотрены некоторые модели данных, применяемые в no-SQL системах.

1.1 Подходы к анализу эмоций в тексте

Лексическая тональность выражается в тексте на уровне лексем или коммуникационных фрагментов. Коммуникативные фрагменты – это отрезки речи различной длины, которые хранятся в памяти говорящего в качестве стационарных частиц его языкового опыта и которыми он оперирует при создании и интерпретации высказываний. [1] Тональность текста в целом определяется лексической тональностью единиц, составляющих текст и правилами их сочетания.

Традиционный подход к определению эмоциональной оценки текста чаще всего заключается в использовании одномерного эмотивного пространства: позитив-негатив, то есть хорошо-плохо. Тональность текста определяется тремя факторами: субъектом тональности, тональной оценкой и объектом тональности. [2]. Под субъектом тональности понимается автор высказывания, под объектом тональности – объект, о котором высказывается субъект и под тональной оценкой – эмоциональное отношение автора к такому объекту.

Лексический уровень языковой системы характеризуется разнообразием экспрессивных слов. Однако, их денотация может изменяться в зависимости от культурного фона или географического положения.

1.1.1 Мониторинг индекса настроений

Лексический фон вбирает в себя те ассоциативные сведения, которые накапливаются у носителей языка в процессе применения слова. [3] Соотвест-

венно, аппроксимация оценки экспрессивного слова неразрывно связана с регионом, в котором проживает субъект тональности.

В работе [4] исследовалась субъективная оценка благополучия, основанная на данных ВКонтакте об активности пользователей. Для формирования выборки были предложены критерии, приведенные ниже.

1. Географическая репрезентативность. Регионы представлены равномерно на всей территории РФ — все федеральные округа, все климатические и культуро-исторические зоны.

2. Социально-экономическая репрезентативность. Согласно типологии [5], населенные пункты были поделены по уровню социально-экономического развития на четыре типа, представленных ниже.

    a) Крупные города – Москва и города-миллионники, города с населением свыше 500 тыс. человек.

    b) Средние по размеру индустриальные города с населением от 20 до 25 – 300 тыс. человек.

    c) «Периферия» – деревни, сёла и небольшие города.

    d) Республики Северного Кавказа и Юга Сибири.

Результаты исследования показали существенную разницу между регионами России. Например, индекс социального (не)благополучия в Республике Алтай составил $-0,4632$, а в Алтайском крае $-18,2462$. Такая существенная разница, скорее всего, будет влиять на общий уровень энтузиазма, и, в результате, на измерение оценки тональных прилагательных. При подготовке обучающего набора данных следует уделить особое внимание разметке корпусов в соответствии с личными характеристиками субъекта – пол, возраст, географическое положение и т. п.

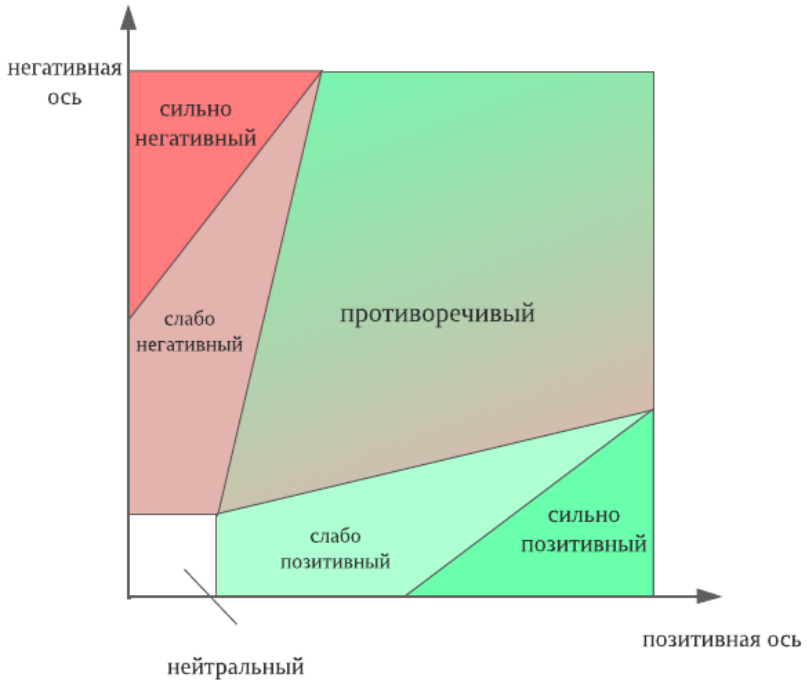
В настоящее время результаты анализа онлайн-текстов не могут рассматриваться как полноценная альтернатива классическим подходам оценки мнений на основе массовых опросов. [6] Для получения более точных результатов автоматического извлечения тональности требуются заранее подготовленные данные – размеченный словарь эмоций или размеченный корпус. Разметка должна предполагать не только обобщение данных до

боле крупных групп населения (например, согласно типологии [5]), но и ассоциировние с социально-демографическими группами, учет пола и возраста.

1.1.2 Шкалы оценки тональных слов

В большинстве случаев для оценки тонального слова используется одномерная шкала (1.1(а)).

(а)



(б)

Рисунок 1.1 – Эмотивное пространство – плоское(а), объемное(б)

На такой шкале расположены шесть классов оценки: «сильно позитивный», «слабо позитивный», «сильно негативный», «слабо негативный», «нейтральный» (отсутствие выраженного эмотивного окраса) и «противоречивый» (в равной мере присутствует негативный и позитивный эмотивный окрас). Однако, в таком случае возникает неопределенность в отношении нулевого, «центрального» значения – ноль на плоской шкале может обозначать как отсутствие эмотивного окраса, так и присутствие позитивного и негативного окраса в равной мере.

Для разрешения этой неопределенности в качестве шаклы оценки используется двумерное пространство, представленное на рис. 1.1(б). Границы между классами изображены приблизительно, симметрия предположительна.

Минимальное количество классов в корпусе – два (позитивный и негативный). [7] Встречаются также случаи пятибальной шкалы оценивания, где «3» – это центральное значение и десятибальной шкалы, где «5» – это центральное значение.

1.1.3 Существующие размеченные эмоциональные корпуса

Корпус ROMIP-2012. [8] Чтобы составить коллекцию, один эксперт разметил отзывы на книги, фильмы и цифорвые камеры. Для проверки ответов систем тестовая коллекция поступила на оценку одному экспертам. В их задачу входило отобрать из всех имеющихся постов такие посты, которые релевантны заданным предметным областям, содержат оценку упоминаемых объектов, а также классифицировать отобранные посты по трем шкалам: двухбалльной (позитивный, негативный), трехбалльной (позитивный, негативный, удовлетворительный), пятибалльной (отлично, хорошо, средне, плохо, ужасно).

Корпус SentiRuEval-2015. [9] В 2015 году было проведено тестирование автоматического анализа тональности русскоязычных текстов. Исследование было разделено на две части. В рамках первой части участникам было предложено найти слова и выражения, обозначающие важные ха-

рактеристики сущности (аспектные термины), и классифицировать их по тональности и обобщенным категориям. Разметка экспертом проводилась по четырехбалльной шкале (позитивный, негативный, противоречивый и нейтральный), далее была проведена проверка. Корпус для первой части состоял из отзывов на автомобили.

В рамках второго задания анализировался корпус публикаций из социальной сети «Твиттер», разметка была проведена тремя экспертами. Для разметки была использована четырехбалльная шкала.

Корпус RuSentiment. Корпус составлен из публикаций социальной сети ВКонтакте. [10] Был размечен тремя экспертами по трехбалльной шкале – позитивный, негативный и нейтральный.

Корпус Kaggle Russian News Dataset. Интернет-ресурс Kaggle [11] содержит корпус, составленный из новостей Казахстана на русском языке. Был размечен по трехбалльной шкале – позитивный, негативный и нейтральный. Метод аннотации и ресурсы неизвестны.

Корпус LinisCrowd. PolSentiLex [12] – тональный словарь, ориентированный на тексты социальных медиа, разработанный в рамках сотрудничества с Лаборатории интернет-исследований (ЛИНИС) НИУ ВШЭ. Для него была сформирована коллекция документов, посвященных социально-политической тематике. В качестве источника данных использовались записи блог-платформы Живой Журнал и социальной сети Фэйсбук. Далее был создан краудфандинговый веб-ресурс [13], позволяющий добровольцам размечать слова и тексты онлайн, а исследователям и практикам – использовать результаты разметки.

Корпус tweets. Tweets [14] – русскоязычный корпус сообщений социальной сети «Twitter». Является одним из немногих на данный момент корпусом текстов на общую тематику. Корпус был разделен на три класса: позитивно окрашенные, негативно окрашенные и нейтральные. Сбор и разметка сообщений производились с помощью специального скрипта и привлечения экспертов. При анализе корпуса была выявленна склонность использовать чаще ту или иную часть речи в зависимости от эмотивной окраски сообщения.

1.2 Формализация задачи

Задача сводится к поиску экспертов. Входными данными в такой задаче является исходная коллекции данных, полученных из массового опроса, а выходными – более крупные группы населения, состоящих из экспертов, обладающих определенной характеристикой. В такой задаче выделяют два основных подхода – первый предполагает выполнение поиска экспертов в две стадии: первичный поиск документов в соответствии с пользовательским запросом и последующий поиск людей в найденных документах. [15] Второй подразумевает построение описания для каждого человека, а поиск людей фактически производится в этих описаниях. [16] Для рассматриваемой задачи требуется использовать второй подход.

Важная особенность модели – это способ моделирования связи между лексикой и человеком. Такая связь выстраивается не только в зависимости от топологических особенностей автора в подсети термина. Здесь подсеть термина – это граф, в котором узлы представляют собой людей, поставивших определенную отметку прилагательному.

1.3 Обзор существующего решения

Британская международная компания «YouGov» провела исследование, в котором выявило, насколько различается численная оценка эмотивных прилагательных в зависимости от географического положения субъекта тональности. [17] В рамках исследования «YouGov» продемонстрировали респондентам список эмотивных прилагательных и предложили оценить каждое из них по шкале 0-10, где 0 – это «сильно негативный», а 10 – «сильно позитивный». Результаты исследования изображены на рис. 1.2.

Исследование было проведено в двух странах - США и Великобритании. Исследование выявило, что респонденты из Великобритании более пессимистичны, чем респонденты из Штатов. В списке было 31 прилагательное, которое в среднем оценили на 8 из 10, но респонденты из Великобритании дали 28 из них более низкий балл. Однако, 9 самых позитивных прилагательных респонденты из Великобритании оценили

более высоко.

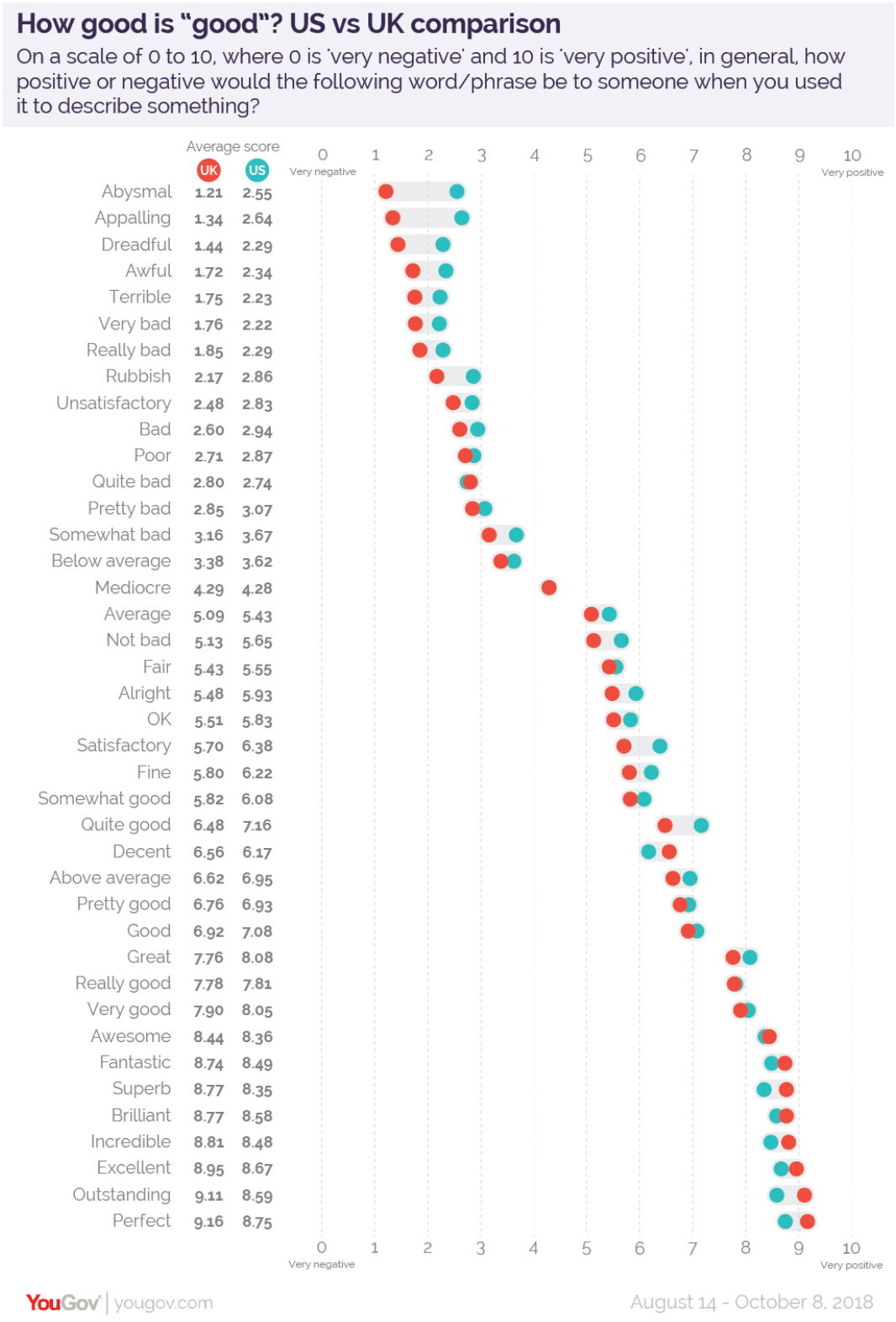


Рисунок 1.2 – Результаты исследования «YouGov»[17]

Самое большое различие наблюдается в более негативных прилагательных – разрыв оценки в слове «abysmal» составляет 1.34 балла – респонденты из Великобритании оценили на 1.21, из Штатов – на 2.55.

Похожие результаты наблюдаются и при оценке слова «appalling» – респонденты из Штатов присвоили слову в среднем более высокую оценку, чем респонденты из Великобритании – различие составляет 1.3 балла.

Самое небольшое различие наблюдается при оценке слова «mediocre» – разница не дотягивает и до половины балла. В Штатах слово оценили на 4.28, а в Великобритании – на 4.29.

1.4 Использование нереляционных баз данных

Нереляционные базы данных могут хранить неструктурированные данные в виде целостной сущности – база данных NoSQL не накладывает ограничений на типы хранимых данных. Потребность в использовании нереляционной базы данных обусловлена тем, что данные, полученные в результате выявления объектов и связей между ними принадлежат классу, который находится «между» Data Mining и Text Mining – Linked Objects mining(далее LOM-mining). [18]

1.4.1 Обзор моделей данных

Существует множество моделей данных в нереляционных базах данных. *Колоночное хранилище.* В колоночных нереляционных базах данных данные хранятся в ячейках, сгруппированных в колонки, а не в строки данных. Колонки логически группируются в колоночные семейства. Колоночные семейства могут состоять из практически неограниченного количества колонок, которые могут создаваться во время работы программы или во время определения схемы. Чтение и запись происходит с использованием колонок, а не строк.

Система ключ-значение. Представляет собой большую хэш-таблицу. Каж-

дое значение сопоставляется с уникальным ключом, и хранилище ключей использует этот ключ для хранения данных, применяя к нему некоторую функцию хэширования. Выбор функции хэширования должен обеспечить равномерное распределение хэшированных ключей по хранилищу данных. [19]

Документо-ориентированные базы данных. Такие базы данных представляют собой усложненную систему «ключ-значение», которая позволяет к каждому ключу привязывать вложенные данные.

Графовая модель Графовые базы данных предназначены для хранения взаимосвязей и навигации в них. В графовых базах данных используются узлы для хранения сущностей данных и ребра для хранения взаимосвязей между сущностями. Для запросов, соответствующих графовой модели, поиск в такой базе может быть эффективнее, чем в реляционной.

Ниже приведена таблица некоторых no-SQL баз данных.(1.1).

Таблица 1.1 – Хранение данных в различных no-SQL базах данных

База данных	*Модель данных*
Cassandra	Семейства столбцов
Redis	Ключ-значение
MongoDB	Документо-ориентированая
Neo4j	Графовая
Scalaris	Ключ-значение

Наиболее подходящими средствами для задачи LOM-mining можно назвать методы из теории графов и теории множеств, [18] поэтому наиболее подходящая модель хранилища данных – графовая.

Вывод

Были проанализированы некоторые размеченные текстовые корпуса для русского языка. Ни один из проанализированных корпусов не учитывал регион проживания субъекта и его личностные характеристики. Чтобы получить более точную тональную оценку, необходимо при разметке учитывать

ассоциирование с социально-демографическими группами, пол и возраст.

Для решения обозначенной проблемы необходимо разработать систему для получения данных, необходимых при построении корпуса, обладающего более тонкой разметкой, учитывающей множество различных характеристик, описанных в разделе. Задача сводится к поиску экспертов при помощи построения описания для каждого человека и поиска людей в рамках этого описания.

Программный продукт, разрабатываемый в данной работе, должен содержать базу данных для хранения результатов статистического опроса о тональности прилагательных.

Тестирование должно представлять собой сопоставление оценочного прилагательного с численной оценкой по плоской шкале от 0 до 10, где 0 – это «сильно негативный», 10 – это «сильно позитивный».

База данных должна хранить информацию о респондентах и результатах тестирования. Аналогично исследованию [17], значимы следующие характеристики респондента: гендерная принадлежность, возраст, населенный пункт и место обучения.

2. Конструкторский раздел

3. Технологический раздел

4. Исследовательский раздел

ЗАКЛЮЧЕНИЕ

Список литературы

1. *Гаспаров Б.* Язык, память, образ. Лингвистика языкового существования. — Новое литературное обозрение, 1996.
2. *Пазельская А. Г.*, *Соловьев А. Н.* Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2011». Сб. научных статей / Вып. 11. — 2011.
3. *Матвеева Т.* Экспрессивность русского слова. — LAP LAMBERT Academic Publishing, 2013.
4. *Щекотин Е.*, *Мягков М.*, *Гойко В.*, *Кашпур В.*, *Г.Ю. К.* Субъективная оценка (не)благополучия населения регионов РФ на основе данных социальных сетей // Мониторингобщественного мнения : Экономические исоциальные перемены. — 2020.
5. *Zubarevich N.* Russia 2025: Scenariosfor the Russian Future. // / под ред. M. Lipman, N. Petrov. — Palgrave Macmillan, London, 2013. — гл. Four Russias: Human Potential and Social Differentiation of Russian Regions and Cities. с. 67—85.
6. *Дудина В. И.*, *Юдина Д. И.* Извлекая мнения из сети Интернет: могут ли методы анализа текстов заменить опросы общественного мнения? // Мониторинг общественного мнения : Экономические и социальные перемены. — 2017.
7. *Котельников Е. В.* Текущее состояние русскоязычных корпусовдля анализа тональности текстов // Computational Linguistics and Intellectual Technologies:Proceedings of the International Conference Dialogue 2021. — 2021.
8. *Четверкин И.*, *Браславский П. И.*, *Лукашевич Н.* Sentiment Analysis Track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog-2012». — 2012.

9. *Loukachevitch N.*, *Blinov P.*, *Kotelnikov E.*, *Rubtsova Y.*, *Ivanov V.*, *Tutubalina E.* SentiRuEval: Testing object-oriented sentiment analysis systems in Russian // Computational Linguistics and IntellectualTechnologies: Proceedings of the International Conference «Dialog-2015». — 2015.

10. *Rogers A.*, *Romanov A.*, *Rumshisky A.*, *Volkova S.*, *Gronas M.*, *Gribov A.* RuSentiment: An Enriched Sentiment Analysis Datasetfor Social Media in Russian // Proceedings of the 27th International Conference onComputational Linguistics. — 2018.

11. Sentiment Analysis in Russian. — дата обновления: 04.08.2022. — URL: `https://www.kaggle.com/c/sentiment-analysis-in-russian`.

12. *Koltsova O. Y.*, *Alexeeva S. V.*, *N. K. S.* An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016". — 2016.

13. Общедоступный тональный словарь PolSentiLex и краудсорсинговая платформа для его создания. — дата обновления: 04.08.2022. — URL: `http://linis-crowd.org/`.

14. *Рубцова Ю.* Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора // инженерия знаний и технологии семантического веба. — 2012.

15. *Petkova D.*, *Croft W. B.* Hierarchical language models for expert finding in enter-prise corpora // Proceedings of the IEEE International Conference on Tools withArtificial Intelligence. — 2006.

16. *Balog K.*, *Azzopardi L.*, *Rijke M. de.* Formal models for expert finding in enterprisecorpora // Proceedings of the Annual International ACM SIGIR Conference on Re-search and Development in Information Retrieval. — 2006.

17. How good is «good»? — дата обновления: 02.01.2022. — URL: `https://yougov.co.uk/topics/lifestyle/articles-reports/2018/10/02/how-good-good`.

18. *Попов И.*, *Фролкина Н.* Анализ связанных объектов и визуализация результатов // Доклады международной конференции Диалог 2004. — 2004.

19. Нереляционные данные и базы данных NoSQL. — дата обновления: 04.09.2022. — URL: `https://docs.microsoft.com/ru-ru/azure/architecture/data-guide/big-data/non-relational-data`.