Документы личного происхождения или в современной терминологии эго-документы давно вошли в перечень исторических источников. Мемуары, дневники, письма, автобиографии, жалобы и челобитные, т.е. все те документы, в которых максимально выражена самопрезентация субъекта всегда пользовались вниманием исследователей. В связи с антропологическим поворотом в изучении истории интерес к ним значительно усилился. Самым ярким примером такого интереса является интенсивное развитие проекта «Прожито» [1], публикующего машиночитаемые тексты дневников, а также и новые издания и исследования, например, серии «Переписка» и «Россия в мемуарах» издательства «Новое литературное обозрение». Отметим также публикации дневников тех, кого иногда называют «маленький человек» [2, 3].

В статье мы предполагаем с использованием современных методов компьютерного анализа текстов проанализировать дневник Владимира Ивановича Чемезова гимназиста, а позднее студента Медико-хирургической академии. Более конкретно речь идет о построении тематической модели дневника для раскрытия в тексте памятника скрытых тем, а также наглядной визуализации полученных результатов.

**Дневник В.И. Чемезова**

Оригинал дневника хранится в Государственной публичной научно-технической библиотеке СО РАН в составе коллекции М.Н. Тихомирова, под наименованием «Дневник студента Медико-хирургической академии П. Орешина. Рукопись 2-й пол. XIX в.» (ГПНТБ СО РАН, собр. М.Н. Тихомирова, № 53). Коллекция М.Н. Тихомирова попала в библиотеку согласно его завещанию. Дневник полностью отсканирован и доступен на сайте библиотеки (http://www.spsl.nsc.ru/rbook/Собрание%20Тихомирова/Tix-053/index.html), также в машиночитаемом виде он представлен на сайте проекта «Прожито» [4]. Автором дневника в библиографическом описании ошибочно указан П. Орешин, поскольку в него вложены два пригласительных билета: одно на имя студента П. Орешина, приглашаемого на курсовое собеседование, второе на имя студента И. Орешина, приглашаемого для участия в выборах старост. Предполагаем, что имеется в виду один и тот же человек, чье имя в одном случае написано неверно. Однако из надписи на корешке тетради «1860 года марта 6. Воскресенье. Записки Чемезова» и из текста дневника становится ясно, что ведет его Володя Чемезов, ученик 3-й Санкт-Петербургской гимназии. Первым, кто детально изучил дневник и идентифицировал автора был Н.Я. Эйдельман, посвятивший ему статью в сборнике «Археография и источниковедение Сибири» [5], затем в популярном виде рассказал о дневнике в статье в журнале «Наука и жизнь» [6], позже включенной в виде главы в книгу «Твой девятнадцатый век» [7]. Он же отметил уникальность данного дневника, содержащего данные по повседневной жизни гимназиста середины XIX века, информацию по истории школы и педагогики, описания отмены крепостного права, студенческих волнений, пересказ свидетельств о временах А.А. Аракчеева.

Дневник представляет собой объемную тетрадь, листы в которой сшиты с нарушением хронологии. Сначала идут записи за первую половину 1862 года (страницы 159-178). Затем собственно начало дневника – записи осени 1860 – лета 1861 гг. (страницы 1-106). Далее следуют таблицы с полученными В.И. Чемезовым итоговыми оценками за семь лет в гимназии, выпускные экзамены и записка с отметками на вступительных экзаменах в Императорскую Медико-хирургическую академию. Новый раздел продолжает записи с лета до конца 1861 года (страницы 107-158). Далее идет раздел «Карты и планы» с единственной нарисованной от руки картой, озаглавленной «Новая граница». Следующими идут записи за вторую половину 1862 года, 1863 и 1864 годы (страница 179-196). Наконец на одиннадцати вложенных листах отдельные записи за 1867, 1870 и 1878 годы. Помимо дневника в рукопись входят письмо 1856 года о студенческих волнениях в Киеве, указанные выше приглашения студенту Орешину, напечатанная на машинке листовка Организационного комитета студентов Петербургского университета от 25 февраля1902 года и черновик студенческого воззвания, отпечатанный на машинке.

Дневник начинается с 21 сентября 1960 (все даты приводятся по старому стилю), записи в него вносятся крайне неравномерно. Автор то надолго забрасывает дневник, то вновь к нему возвращается, стремясь записать то, что его более всего волнует и память о чем он хочет сохранить. К 1860 году относится 28 записей, к 1861 – 50, к 1862 – 28, к 1863 – 8, к 1864 – 3, к 1867 – 4, к 1870 – 1. Последняя запись сделана 4 сентября 1878 года.

«Долго я не писал в этой тетради. Наконец и соскучился. Отчего же и не написать чего-нибудь?», – такими словами начинается текст дневника [4, запись от 3 октября (21 сентября) 1867]. Первые записи дневника пишет пятнадцатилетний подросток, а последнюю уже взрослый мужчина в возрасте 33 лет, ставший врачом, защитивший диссертацию, прошедший войну. Перечень сюжетов весьма обширный, в нем ежедневная рутина, учеба в гимназии, подготовка домашних заданий, походы в театры и на выставки, летний отдых на даче, посещение гостей, отзывы о прочитанных книгах, выписанные в дневник стихотворения, любовные переживания, рассуждения о происходящих в стране событиях и многое другое. Дневник для автора не просто памятные записки, но и своеобразная психотерапия, материал для самоанализа. Весьма примечательной является запись за 11 января 1867 года: «Я думаю снова начать записывать, чтобы легче впоследствии обсудить, что я за человек, куда гожусь и как себя держать в обществе, чтобы быть хоть сколько нибудь для него нужным. Под словом общество я разумею простое светское общество. Я, вообще, хочу записывать свое умственное и если можно выразиться так, чувственное развитие. Это разумеется не [нрзб]можно сделать, как строго записывать своим мысли и чувства, и чрез несколько времени прочитать, чтобы увидеть, что я пережил» [4, запись от 23 января (11 января) 1867]. (В цитатах сохраняется орфография автора.) Учитывая вышеизложенное, нам кажется интересным посмотреть на развитие во времени тем и сюжетов, которые более всего интересовали автора, так что он оставлял о них записи в дневнике.

**Биография В.И. Чемезова**

Рассматривать дневник невозможно, не дав хотя бы краткой биографической справки об авторе и его семье – основных действующих лицах дневника. Подробной биографии В.И. Чемезова не существует. Он не упоминается в дореволюционных биографических словарях. Его дневник опубликован лишь недавно и в широкий научный оборот не вошел. Помимо работ Н.Я. Эйдельмана, нам удалось обнаружить ссылку на анализируемый нами дневник только в статье Р. Васенны [8]. Но имеющийся материал позволяют восстановить, хоть и фрагментарно биографию нашего героя, и получить представления о его семье для лучшего понимания текста дневника. В фундаментальном справочнике «История дореволюционной России в дневниках и воспоминаниях» [9, с. 153] приведено указание на автобиографию В.И. Чемезова в юбилейном сборнике «Двадцатипятилетие деятельности врачей, окончивших курс в Императорской Медико-хирургической академии в 1868 году», одним из составителей которого он являлся [10, с. 120—126]. В автобиографии содержится информация преимущественно с окончания обучения в Академии вплоть до начала 1890-х годов и сфокусированная на его служебной деятельности, стороной обходящая личную жизнь автора. В.И. Чемезов упоминается в справочнике «Русские врачи-писатели» [11, с. 190], имеется и краткий некролог, напечатанный в «Историческом вестнике» [12] Дополняют мозаику разнообразные справочники: родословные книги, адресные книги, списки чинов.

Владимир Иванович Чемезов родился 5 апреля 1845 г. Окончил 3-ю гимназию, где учился как приходящий своекоштный воспитанник. В 1862 году поступил своекоштным студентом в Медико-хирургическую академию. С четвертого курса зачислен стипендиатом военного ведомства. В декабре 1867 года окончил курс с серебряной медалью. Подал докладную записку и был оставлен на три года для дальнейшего обучения. В 1868 году В.И. Чемезов впервые упоминается как лекарь без указания чина, служащий по Военному Министерству, в регулярно издаваемом «Российском медицинском списке, изданный Медицинским департаментом Министерства внутренних дел на 1869 год» в числе медиков «имеющим полное право на производство в России врачебной практики» [13, с. 257]. Год работал в клинике В.Е. Экка. Под его влиянием и влиянием Э.Э. Эйхвальда начал изучать внутренние болезни. Писал диссертацию и держал экзамены на доктора. Экзамены по его отзыву весьма скучные и требующие длительной подготовки. С сожалением указывает, что на подготовку к экзаменам был потрачен целый год. Два оставшихся года посвятил научным изысканиям по следующим темам: «1) ход температуры при легочной чахотке, 2) некоторые физические свойства белков и 3) действие озона на животных» [10, с. 122]. Последняя тема стала основой диссертации, защищенной в мае 1876 году [14].

В 1874 году по предложению Э.Э. Эйхвальда стал его ассистентом на женских медицинских курсах. Руководил практическими занятиями третьего и четвертого курсов женщин-врачей. Это был первый выпуск женщин-врачей в истории России.

В том же 1874 году получил место в лейб-гвардии казачьем Его Величества полку в качестве врача для командировок. Одновременно исполнял должность младшего врача при лазарете лейб-гвардии Московского полка. Это были три очень напряженных года. Как он сам пишет в автобиографии: «При дежурствах через 3-4 дня, редко через неделю, разводах, парадах прошлого царствования, ежедневном осмотре 40-80 больных внутренними болезнями в лазарете, стоило немалого труда вести ежедневные же вечерние занятия с III курсом слушательниц. Да и летом не было свободы, одно лето я провел в петергофском госпитале, заменяя больного товарища (Евдокимова), два других пришлось провести с казаками в лагере в Красном селе» [10, с. 122-123].

С началом русско-турецкой войны с мая до 5 ноября 1877 года был в действующих войсках на балканском фронте. Нес все тяготы войны наравне с офицерами, оказывая помощь раненым на боле боя. За участие в военных действиях он был награжден орденом Владимира 4-й степени с мечами, светло-бронзовой медалью за турецкую войну 1877-78 годов и румынским крестом за переход через Дунай.

10 декабря 1877 года уехал в Санкт-Петербург, поскольку по представлению Э.Э. Эйхвальда был назначен ассистентом I терапевтического отделения Клинического военного госпиталя при Медико-хирургической академии.

На новой должности В.И. Чемезов был сильно загружен работой. «Вообще положение ассистента большой клиники далеко не легкое: обойти больных, приготовить выбранного на лекцию, ответить на вопросы ординаторов, в особенности неопытных, проверить рецепты, заглянуть на амбулаторный прием – требует немало времени. … В конце концов бывали дни, когда приходилось проводить в клинике по 8 часов, неся массу черной работы» [10, с. 124]. И это помимо занятий со студентами 2-4 раза в неделю.

В 1879 году в связи с командировкой Э.Э. Эйхвальда на полтора месяца заменил его в руководстве клиники. Этот опыт очень запомнился. Он описал его следующим образом: «Мне приходилось утром читать лекции (2 в неделю) и делать обходы клиники (тоже 2) в роли профессора, а вечером заниматься с студентами в качестве ассистента (два раза в неделю). … Эти недолгие полтора месяца стоили мне не мало труда и здоровья [10, с. 124-125].

В январе 1890 года получил место врача Военно-медицинской академии. Это стало очередным поворотным пунктом в его карьере и жизни. «…Двадцать два года я не знал, что такое праздники; но теперь наконец имею дни, когда не должен идти на службу и могу более располагать своим временем чем прежде», - писал он о новом назначении [10, с. 125]. В конце 1891 года был избран в консультанты Максимилиановской лечебницы советом консультантов [10, с. 125 -126].

За пределами автобиографии обстоятельства жизни В.И. Чемезова весьма фрагментарны.

С 3 февраля 1902 по 1910 гг. он занимал должность главного врача Клинического военного госпиталя. С 6 декабря 1902 г. – действительный статский советник. Известно, что в это время он женат и имеет двух дочерей [15, с. 35]. Согласно справочнику «Весь Петербург» за 1909 год, проживает он по адресу Боткинская ул., д. 1. У него имеется домашний телефон с номером 7020 [16, с. 856]. Из этого же справочника мы узнаем, что его жену зовут Эмилия Федоровна. Пока нам не удалось выяснить когда он заключил брак, но очевидно, что брак этот был поздний. Примечательный случай позволил Н.Я. Эйдельману получить некоторую информацию о семье В.И. Чемезова. После того как статья о дневнике гимназиста была опубликована в журнале «Наука и жизнь», в редакцию пришло письмо от пожилой ленинградки А. К. Ионовой, в котором сообщалось, что она была знакома с семьей доктора В.И. Чемезова: «Были две дочери, славные девушки Ольга и Вера. Они вместе со мною учились, но в разных классах; я гимназию закончила в 1912 году. У них была красавица мать, большая рукодельница. По окончании гимназии Оля в качестве корреспондентки поехала в Англию, но через пару лет оттуда сообщили о ее смерти... Вера тоже умерла совсем молодой во время эпидемии испанки» [7, с. 743].

В последний год жизни В.И. Чемезов числится в справочнике «Весь Петербург» как тайный советник, проживает по адресу Саперный пер., д. 13, но сохраняет прежний номер телефона [17, с 967]. Уйдя в 1910 году с должности главного врача Клинического военного госпиталя он до конца жизни остается практикующим врачом, в качестве почетного консультанта и члена попечительского совета участвует в работе Максимилиановской лечебницы для приходящих больных, а также является товарищем председателя совет Общества попечения о бедных больных [12].

Умер 7 мая 1911 г. Похоронен на Новодевичьем кладбище в Санкт-Петербурге [18, с. 499].

На страницах дневника упомянуты многочисленные родственники, друзья и знакомые семьи Чемезовых. В большинстве случаев нет возможности установить их личность. Володя Чемезов является младшим и, вероятно, любимым ребенком в большой семье. Родители именуются им в дневнике не иначе как папашенька и мамашенька. Отец – отставной коллежский советник Иван Леонтьевич Чемезов (ок. 1797 – 1884) – представитель выслужного дворянства [19, с.14]. На момент начала ведения дневника уже весьма пожилой мужчина, примерно 60 лет. В дневнике нет никаких упоминаний, что он ходит на службу из чего учитывая возраст мы делаем вывод о том, что он уже ушел в отставку. В «Адрес-календаре» за 1844 год указано, что И.Л. Чемезов в чине надворного советника числится старшим контролером в 3-м отделении Хозяйственного управления Святейшего Синода [20, с. 221]. В 1840 году он, дослужившись до чина коллежского асессора, получает потомственное дворянство [19, с.14]. Мать – Чемезова Вера Николаевна (ок. 1812 – 1887) [19, с.14] примерно 48 лет.

В семье пять детей. Старшая Екатерина (1833-1903), именуемая в дневнике чаще как Каташенька, вышла замуж за Ивана Галактионовича Георгиевского, штаб-лекаря в Царской Славянке и проживает в Павловске [21, с. 75]. Имеет детей и часто навещает родителей. Младшие дочери Мария (1839 –после1910) и София (1840 – 1878) [19, с.14] живут в семье. София работает гувернанткой. Старший сын Константин (1842 – 1908), в дневнике чаще именуемый Котя, в 1860 году закончил туже, что и автор дневника, 3-ю Санкт-петербургскую гимназию, где учился за казенный счет [22, с. 71, 23, с. 38]. В тот же год он поступил в Петербургский университет. По окончании университет он два года будет учится на педагогических курсах при Главном управлении военно-учебных заведений и в 1868 году станет преподавателем русского языка и географии в Петровской Полтавской военной гимназии (позже Петровский Полтавский кадетский корпус), где и проработает 40 лет до своей кончины [24, с.147-149].

В семье работает целый штат прислуги. В квартире Царском Селе – слуга Кирилла. В московской квартире –  Катерина, Николашка, Лерка и Алешка. При этом доходы семьи небольшие. Володя Чемезов постоянно пишет про нехватку денег. Дополнительные средства дает то, что семья сдает в аренду квартиры в Санкт-Петербурге и Царском Селе.

**Метод тематического моделирования**

Вероятностное тематическое моделирование (англ. probabilistic topic modeling) – это метод неконтролируемой классификации коллекции текстовых документов, который «выявляет тематику коллекции документов, представляя каждую тему дискретным распределением вероятностей терминов, а каждый документ – дискретным распределением вероятностей тем» [25, с. 195]. В основе тематического моделирования лежит предположение, что каждый документ из текстового корпуса состоит из одной или нескольких тем в разной пропорции, а каждая тема представляет собой определенный набор слов. Тематическая модель текстового корпуса как раз и выявляет, к каким темам относится каждый документ и какие термины (слова, n-граммы) образуют каждую тему.

Применительно к тематическому моделированию существует много определений «темы», как «скрытых паттернов», «скрытых текстуальных слоев, которые могут присутствовать имплицитно, но не вербализуются эксплицитно», «компактных описаний смысла документов», «вероятностных (нечётких) кластеров семантически связанных терминов» [26, p. 282], но во всех случаях тема – это совокупность слов, которые статистически часто совместно употребляются в документах коллекции, результат одновременной кластеризации и слов, и документов по их семантической близости.

Впервые тематическое моделирование было предложено в 1998 году в работе К. Пападимитриу, П. Рагавана, Х. Томаки и С. Вемполы [27]. В настоящее время тематическое моделирование широко применяется для решения разнообразных прикладных задач, таких как разведочный информационный поиск, классификация и категоризация документов, автоматическая аннотации текстовых корпусов, выявление трендов в научных публикациях и новостных потоках.

Вероятнее всего первым историческим исследованием, использующим метод тематического моделирования был проект по анализу тематики газеты Pennsylvania Gazette за весь период её существования с 1728 по 1800 год, результаты которого были опубликованы в 2006 году [28]. С 2010 года наблюдается рост интереса к тематическому моделированию со стороны гуманитариев, в том числе историков [29, p. 1]. Чаще всего метод тематического моделирования применяется для анализа дневников [30], писем [31, 32], периодических изданий [33, 34]. В настоящее время разработано множество конкретных вариантов построения тематических моделей [35, с. 63], но наиболее популярным алгоритмом остается скрытое размещение Дирихле (Latent Dirichlet Allocation или LDA), предложенный Д. Блейем, Э. Ыном и М. Джорданом [36].

С практической точки зрения тематическое моделирование текстовой коллекции предполагает реализацию нескольких конкретных шагов: 1) предварительную обработку текстов, 2) подбор оптимальных параметров моделирования, 3) визуализацию полученных результатов.

**Предварительная обработка текста.**

Предварительная обработка текста является одним из важнейших этапов для любых методов автоматический обработки и анализа тестов, качество которой напрямую влияет на результат. Цель предварительной обработки – преобразовать неструктурированные тексты в структурированный набор данных. В наиболее общем виде предварительная обработка текстов включает такие процедуры как: 1. Токенизация – разбиение текста на фрагменты (абзацы, предложения, слова, N-граммы). 2. Очистка текста от лишних пробелов и пустых строк, типографских знаков, чисел, знаков препинания, перевод букв в нижний регистр. 3. Удаление стоп-слов – малозначимых и низкоинформативных (служебные части речи, местоимения и числительные). 4. Лемматизация – приведение слова к нормально (словарной) форме. Для русского языка нормальными формами считаются: для существительных – именительный падеж, единственное число; для прилагательных – именительный падеж, единственное число, мужской род; для глаголов, причастий, деепричастий – глагол в инфинитиве. 5. Морфологический анализ слов. 6. Синтаксический анализ предложений [37, с. 37–44, 38, pp. 37–44].

Для анализа был использован текст дневника с сайта проекта «Прожито» [4], выверенный по скану оригинала. Каждая запись из дневника была сохранена в виде отдельного файла, поименованного в соответствии с датой записи. Под датой 9 ноября 1860 года имеются две записи, сохраненные в отдельные файлы. Чтобы используемые программы адекватно работали с дневником в его тексте было изменено написание ряда слов в соответствии с современными правилами написания: «разсказал» – «рассказал», «счастие» – «счастье», «извощик» – «извозчик», «воксал» – «вокзал», «прикащик» – «приказчик» и т.п. По той же причине используемые автором для указания родителей «мамашенька» и «папашенька» заменены на «мама» и «папа».

Предварительная обработка и анализ текста проводился на языке программирования R. Текст скрипта на языке R и все ресурсы доступны в репозитории сервиса GitHub [39]. В качестве модели для тематического моделирования была избрана модель латентного размещения Дирихле (англ. Latent Dirichlet Allocation, LDA), наиболее популярная и хорошо документированная модель для тематического моделирования в гуманитарных исследованиях.

Из текстовых файлов с помощью пакета quanteda [40] был сформирован текстовый корпус, который был аннотирован с использованием языковой модели, обученной на основе размеченного корпуса СинТагРус [42], в составе пакета UDPipe [42].

Пакет UDPipe обладает широкими возможностями для обработки текстов: позволяет разбить тексты корпуса на предложения и токены, провести лемматизацию и синтаксический парсинг. В результате была сформирована таблица данных (data frame), в которой для каждого слова в тексте приведена информация о начальной форме слова (лемме), морфологической форме слова и синтаксической роли в предложении.

Тематические модели строятся, как правило, на основе матриц документ-термин (document-term matrix) [43, p. 70]. В такой матрице каждая строка соответствует одному документу, а столбцы – терминам. Количество строк равно количеству документов, а количество столбцов – размеру словаря. В нашем случае мы строили тематическую модель на основе словарных форм (лемм) существительных и прилагательных. Попытки строить модель с использованием глаголов дала на наш взгляд худший результат в распределении терминов по темам. Из матрицы документ-термин были удалены термины, встречающиеся менее чем 4 раза, а также некоторые, встречающиеся излишне часто. За основу списка стоп-слов были взяты стоп-слова, входящие в состав пакет quanteda, дополненный специфическими для данного источника, в первую очередь, названия месяцев и дней недели.

**Подбор оптимальных параметров моделирования**

Помимо предварительной обработки текстов, другим важнейшим составляющим качества тематического моделирования является выбор оптимальных параметров моделирования – гиперпараметров модели. Главный из них – количества тем, которые будут извлечены из корпуса текстов, а также параметры α-альфа, управляющий вероятностным распределением тем по документам, и парметр β-бета, определяющий вероятностное распределение слов по темам [44, 45].

К определению оптимального количества тем при тематическом моделировании существует множество подходов [25, с. 245-249].единого стандарта не выработано. В нашем случае мы ориентировались на результаты применения к матрице документ-термин пакета ldatuning [46]. Данный пакет вычисляет оптимальное количество тем сразу по четырем метрикам. В нашем случае были получены следующие результаты (см. рис. 1).

{Рисунок1}

Рисунок 1. Оптимально количество тем, рассчитанное при помощи пакета ldatuning.

Судя по графику, оптимальным количеством тем является 6 или 10. Именно на этих значениях совокупность всех четырех метрик показывают наилучший результат. Мы исходим из того, что чем меньше число тем, тем лучше. С возрастанием числа тем интерпретируемость большинства их падает, слова излишне часто повторяются в темах. При значении 10 темы нам показались не интерпретируемыми, с излишним повторением терминов, поэтому для построения тематической модели было выбрано число тем равное 6.

Определение оптимального значения параметров альфа и бета, является отдельной дискутируемой проблемой. Высокое значение альфа, приводит к тому, что документ соотносится с несколькими темами, и больше документов будет классифицироваться как похожие друг на друга, и наоборот. Высокое значение бета приводит к тому, что слова будут часто повторяться в разных темах, и темы будут казаться излишне похожими.

Используемый нами для тематического моделирования пакет topicmodels [47] позволяет явно задать значение только параметра альфа. Поскольку мы стремились избежать смешения тем в документах выбирались минимальные значения параметра альфа. В конечном итоге остановились на значении.0,3. Для параметра бета при моделировании была задействована опция оптимизации.

В итоге процесса моделирования была сформирована тематическая модель, содержащая два распределения: распределение слов по темам – матрица 𝛷 в общепринятом обозначении(матрица beta в пакете topicmodels), и распределение тем по документам – матрица 𝛩 (матрица gamma в пакете topicmodels).

**Распределение слов то темам**

Одной из важнейших характеристикой качества тематического моделирования (наряду с другими) является интерпретируемость тем. Предполагается, что в хорошей тематической модели эксперт может легко понять, чему посвящена та или иная тема на основе наиболее вероятных для неё слов. На практике так получается далеко не всегда, список слов, составляющих тему, не всегда дает возможность точно судить, о чем эта тема. В нашем случае было получено следующее распределение вероятности принадлежности слов к темам (см. рис. 2).

{Рисунок2}

Рисунок 2. Распределение слов по темам.

**Распределение тем по документам**

При определении смысла выявленных тем мы опирались не только на составляющие тему слова, а на содержание конкретных дневниковых записей, в которых эта тема преобладает. У дневнике оказалось достаточно много записей, в которых абсолютно превалирует какая-либо одна тема.

Рассмотрим темы в порядке их присутствия в тексте дневника, начиная с более всего представленных в записях дневника (см. табл. 1, табл. 2).

|  |  |
| --- | --- |
| Тема | % |
| класс экзамен гимназия язык история | 0.2954918 |
| мама рыба папа снег погода | 0.2016063 |
| любовь веселый общество театр глупый | 0.1408911 |
| гимназия воля сердце польза папа | 0.1227732 |
| собака золотой картина скамейка замок | 0.1206863 |
| студент университет академия попечитель профессор | 0.1185512 |

Таблица 1. Распределение тем по корпусу в процентном отношении.

|  |  |
| --- | --- |
| Тема | Документы |
| класс экзамен гимназия язык история | 40 |
| мама рыба папа снег погода | 26 |
| любовь веселый общество театр глупый | 17 |
| студент университет академия попечитель профессор | 14 |
| гимназия воля сердце польза папа | 13 |
| собака золотой картина скамейка замок | 13 |

Таблица 2. Количество документов для каждой темы, в которых тема имеет наибольший вес.

Во всем тексте дневника превалирует тема номер 4. Она охватывает 29,5 % текста и преобладает в 40 дневниковых записях. Смысл этой темы легко угадывается по составляющим её словам: класс, экзамен, гимназия, язык, история, билет, французский и т.д. Тему можно именовать как «Учеба в гимназии и сдача экзаменов». Как мы уже отмечали выше, основная часть дневника написана в годы обучения в гимназии, естественно, что автор в первую очередь пишет о том, как он готовится и отвечает на уроках, сдает экзамены, ищет необходимые книги, пытается попасть в университетскую библиотеку. Дневник полон интересных подробностей о жизни гимназиста середины XIX века, гимназии и педагогике. В числе преподавателей В. Чемезова были выдающиеся педагоги того времени: В.Я. Стоюнин, Ф.И. Буссе, В.Х. Лемониус, К.В. Кедров.

В образования Володя Чемезов видит единственный путь для занятия достойного места в обществе. Пределом его мечтаний являются получение степени доктора, женитьба и «безбедное состояние». «Все бы мои желания окончились со степенью доктора медицины и хоршенькой и умной и доброй и пр. и пр. женой. Более для меня ничего не нужно. Конечно к этому надо прибавить безбедное состояние, чтобы не мучить жены» [4, запись от 21 апреля (9 апреля) 1862]. Но успехи гимназиста в учебе весьма скромные, успеваемость не на высоте. За первые шесть лет обучения средняя оценка не дотягивает до 4 баллов. Второй класс он окончил восьмым учеником, третий пятнадцатым, четвертый десятым, пятый седьмым. В седьмой класс он переходит лишь десятым в классе, хотя предполагал седьмым и надеялся получить похвальный лист [4, запись от 26 июня (14 июня) 1861]. Выйти из гимназии в числе лучших значит не только получить аттестат, медаль или похвальный лист, но также и право на чин XIV класса – коллежского регистратора при поступлении на службу. Неплохое начало карьеры. В.И. Чемезов часто сетует, что плохо учился все предыдущие годы, и основную проблему этого видит в собственной лености и отсутствии воли, дабы её преодолеть [4, запись от 5 октября (23 сентября) 1860]. Обучение в гимназии – источник постоянных переживаний, начиная с первого класса он испытывает боязнь не перейти в следующий [4, запись от 24 июня (12 июня) 1861].

Дальнейшее своё образование В.И Чемезов планирует продолжить в Медико-хирургической академии. Один из мотивов поступления в Академию – простые экзамены [4, запись от 22 сентября (10 сентября) 1861]. В 1861 году в Академию поступил его друг Фридрих Рибо, от него В.И. Чемезов узнает об экзаменах в Академию, несколько раз они вместе посещают её и даже заходят в морг.

В конце марта 1862 года гимназистов отпустили на подготовку к экзамену. В это время в дневнике появляется всего несколько записей. Видимо подготовка к экзаменам отнимает много времени и сил. Усилия на подготовку были вознаграждены. За седьмой класс средний балл гимназиста – 4, на экзаменах получено 20 пятёрок, 25 четверок и 10 троек. Из гимназии В.И. Чемезов вышел 7-м, получив похвальный аттестат с правом на чин XIV класса при поступлении на службу. Но впереди были ещё экзамены в Медико-хирургическую академию. Судя по дневнику, бывший гимназист особенно не усердствует в подготовке к очередным экзаменам: «16 августа, в четверг подал прошение в Академию, в сентябре, тоже в четверг намерен держать вступительные экзамены. Здесь должно надеяться на одно только счастье, потому что я не повторял ни математики, ни географии, ни русского, латинского, немецкого языков, ни всеобщей средней и новой истории. Не знаю, что будет, а повторять лень. Ничего, сойдет. Сошло же выпускные экзамены. Еще труднее этих были» [4, запись от 6 сентября (25 августа) 1862]. 4 и 5 сентября 1862 года В.И. Чемезов успешно выдерживает вступительные экзамены. Вместе с ним в Медико-хирургическую академию поступили учившиеся с ним в гимназии Людвиг Барро и Александр Пасвик (отец советской ученой-радиохимика, участницы ядерной программы СССР М.А. Пасвик). Среди выпускников 3-й Санкт-Петербургской гимназии, продолжающих обучение, Медико-хирургическая академия находится на втором месте по популярности после Санкт-Петербургского университета. В 1860 году в Академию поступили 2 выпускника гимназии, 1861 – также 2, 1862 – 3, 1863 – 4, 1864 – 6 [22, с. 70 -76].

Чуть позже В.И. Чемезов записывает в дневнике: «Я намерен как можно больше и подробнее писать об Медицинской Академии, чтобы потом иметь возможность теперешний быт Академии сравнить с будущим. Не знаю с чего и начать. С того разве, что студенты носят офицерскую форму и отличаются от них только красным шнуром вместо серебряного или золотого, смотря по пуговицам. Однако лень стало писать. Надо заниматься и – и того не хочется. Так лучше упражняться в силе. Все так на свете. Что хочешь сделать – никогда того не сделаешь» [4, запись от 20 ноября (8 ноября) 1862]. Последняя фраза оказывается пророческой. На несколько месяцев студент Академии забывает про дневник. Новые записи появятся только в конце весны 1863 года и отметят новый этап в его жизни.

Вторая по значимости тема дневника – тема номер 1. Она охватывает 20 % текста и преобладает в 26 дневниковых записях. Смысл этой темы не столь ясно проступает из составляющих её слов, как у предыдущей: мама, рыба, папа, снег, погода, рубль, небо и т.д. По записям, в которых она доминирует однако становится ясно, что эту тему можно условно назвать «Летний отдых, рыбалка, поездки». В основном тема локализуется в записях лета 1861 года. Это по сути последнее беззаботное лето детства нашего героя. В тот год 23 мая В. Чемезов узнает, что его отпускают без экзаменов. Уже 28 мая он едет на поезде в Царское Село, где живет на семейной квартире вместе со слугой Кириллой.

Петербургские дачи и дачная жизнь это уникальное социально-культурное явление [48, 49]. Уже в середине XIX века сложилась определенная топографическая «специализация» дачных пригородов Петербурга. Каменный остров, Царское Село и Павловск были местом, куда на лето переезжали высшие слои петербургского общества [50]. «Дачи в Павловском и Царском Селе, т.е. квартиры, нанимаемые на лето в селении и в городе, принадлежат к особому разряду. Тут живут семейства, любящие городской шум, городскую жизнь и городской туалет, ищущие многолюдных гульбищ, виста, преферанса, словом рассеяния. Здесь самая приятная городская жизнь среди великолепной природы, изукрашенной искусством. Удобство сообщения с Петербургом, по железной дороге, возвышает приятности этого местопребывания» [51].

Прогулки и рыбалка в Царском Селе очень по душе нашему герою, но его тяготит соседство с семьей Леонтьевых, а также то, что Чемезовы не могут провести лето в Грузино – «маленьком Петербурге», бывшем имении А.А. Аракчеева. Знакомство и тесное общение Чемезовых и Леонтьевых весьма примечательно. Вероятно, именно дачное соседство и свело семью статского советника, получившего потомственное дворянство в 1840 году с одним из знатных родов империи. «Дамы разговаривают, потому что на дачах легко знакомятся и по соседству часто сходятся. Зимой можно и не продолжать летнего знакомства, ибо два города, летний и зимний, имеют особые нравы и обычаи», писал в 1837 году о дачном отдыхе Ф. Булгарин [52]. Мария Павловна Леонтьева – статс-дама, директор Смольного института неоднократно приезжает к Чемезовым, принимает их в Смольном институте и даже организует у них обеды. В Москве Володя Чемезов почти каждое воскресенье ходит к ним в гости. Отец Володи, однако, не просто дачный знакомый. Он ведет с М.П. Леонтьевой какие-то дела и выполняет её поручения по строительству неких «служб» [4, запись от 4 июля (22 июня) 1861], приобретению стройматериалов [4, запись от 2 июля (20 июня) 1861] и оплате труда работников [4, запись 26 июня (14 июня) 1861]. Сын Марии Павловны – Владимир Николаевич Леонтьев – действительный статский советник, камергер двора, и его жена Софья Константиновна – фрейлина двора воспитывают двух детей. В Царском Селе В. Чемезов часто без особого желания играет со старшим мальчиком – Володенькой и всё больше беспокоится, что не может поехать в Грузино.

Чемезовы стеснены в средствах, поэтому поездка в Грузино зависит от того удастся ли сдать в аренду квартиру в Царском Селе и на эти деньги арендовать дом. В тот год им повезло: царскосельскую квартиру взял в аренду Э.Э. Эйхвальд, профессор Медико-хирургической академии, который впоследствии сыграет существенную роль в судьбе молодого врача Владимира Чемезова, а одну из квартир в Санкт-Петербурге – Л.Н. Модзелевский, выдающийся русский педагог XIX века, в описываемое время преподаватель Смольного института [4, запись от 25 июня (13 июня) 1861].

27 июня он наконец уехал в Грузино. На лето за 35 рублей серебром семья арендовала «молочный домик» [4, запись от 6 августа (25 июля) 1861]. Возможно имеется ввиду перестроенная под сдачу в наем аракчеевская молочня – деревянный красиво украшенный внутри домик с подвалом-колодцем для хранения молочных продуктов, упоминаемая современниками А.А. Аракчеева [53, с. 52, 54, с. 39].

Грузино – полная противоположность Царскому Селу. Сравнительно удаленный населенный пункт, не имеющий прямого железнодорожного сообщения с Петербургом и даже не относящийся к Санкт-Петербургской губернии. Н.Г. Богословский, c 1854 по 1863 год бывший старшим священником в Андреевском соборе в Грузино, описывал село как тихое безлюдное место [55, с. 45]. Но в Грузино в тот период ещё сохранялась живая память о графе А.А. Аракчееве и военных поселениях, были живы люди знавшие графа лично. Собор Лето проходило для В. Чемезова весьма плодотворно: «Я в это лето многому научился: стрелять, играть в шахматы, гресть веслами, править рулем и даже немного держаться на воде. Хотел заниматься Саллюстием, но Костя не привез его. Теперь я хочу по возможности отовсюду собирать сведения о Графе Аракчееве, чтобы потом подать Стоюнину или просто написать для себя два сочинения. Предметом первого будет разсмотрение характера Аракчеева, второго описание важнейших фактов его жизни. Сколько я могу судить о нем по отзывам одного старика по имени Ивана Васильевича, жившего еще при нем, а потом и некоторых писателей, как нпр. Искандера, Пушкина, Дашковой, то можно сказать, что он был нехороший человек» [4, запись от 6 августа (25 июля) 1861]. Далее в дневнике приводится пересказ воспоминания Ивана Васильевича о суровости графа А.А. Аракчеева, наказавшего его однажды ночью, проведенной в покойницкой во время вспышки холеры.

9 августа В.И. Чемезов возвращается в Петербург и ожидает начала занятий, но каникулы продолжены до 21 августа, и он еще раз едет в Грузино. Уже приступив к занятиям гимназист не раз вспоминает это лето.

Третья по значимости тема дневника – тема номер 5. Она охватывает 14 % текста и преобладает в 17 дневниковых записях. Наиболее значимые для этой темы слова – любовь, веселый, общество, театр, глупый. Эта тема объединяет записи, рассказывающие о любовных переживаниях автора. Дневник сохранил истории двух влюбленностей студента Медико-хирургической академии. Первая случилась с ним летом 1863 года. Любовные переживания заставляют В.И. Чемезова вернуться к заброшенному дневнику. Им собственно и посвящены почти все записи в дневнике за этот год и начало следующего.

Предмет страсти студента в данном случае легко идентифицируется – Елизавета Ивановна Джустиниани – дочь Джованни (Ивана) Джустиниани. Д. Джустиниани весьма примечательная личность – поэт-импровизатор, путешественник, ученый, библиофил [56]. Он родился в 1810 г. в городе Имола в Папской области, изучал юриспруденцию в Риме, много путешествовал, выступая с поэтическими импровизациями, участвовал в борьбе за объединение Италии. В 1839 впервые прибыл в Россию, где познакомил публику с жанром поэтических импровизаций. Выступал с импровизациями в Одессе, Москве и Санкт-Петербурге. C 1851 по 1859 годы преподавал итальянский язык в Петербургском университете.

Первая дневниковая запись за 1863 год сообщает: «Вчера я получил от Джустиниани Екат[ерины] Ивановны атлас хирургической анатомии в память ее сына Юлия Ивановича, умершего 6-го января нынешнего года»[4, запись от 8 июня (27 мая) 1863]. Далее мы узнаем, что он присутствовал на похоронах, и там впервые увидел Елизавету Ивановну, сестру покойного. Предполагаем, что Юлий Джустиниани учился с ним вместе в Академии, отсюда и его участие в похоронах, и такой памятный подарок в виде атласа анатомии.

4 июля 1863 г. в дневнике записано: «Удивительно, что со мною делается. Я. кажется, влюбился в Дж. … Что я влюблен, в этом теперь нет сомнения. Для чего я всегда хорошенько одеваюсь и украшаюсь, когда иду к Дж.? … Я сам прежде не верил в любовь, смеялся над влюбленными, а теперь самому пришлось испытать это. Я бы впрочем очень рад был, если б освободился от этой глупой любви. Что в ней толку? Я еще молокосос, решиться не могу для чего ж развивать в себе это чувство? … Если она останется ко мне совершенно равнодушно, не выйдет за меня замуж и любовь моя разовьется еще более – я кажется застрелюсь. До сих пор я не понимал, что значат душевные страдания, а теперь к несчастию коротко познакомился с ними. … Да чем я могу заслужить ее расположение? Никаких достоинств, отличающих меня от других, по крайней мере недюжинных — не имею, ничего порядком не знаю, физиономией – урод уродом» [4, запись от 16 июля (4 июля) 1863].

Летом семья Джустиниани живет на даче в Павловске. Владимир Чемезов уже не вспоминает Грузино, но стремится в Павловск, чтобы ещё раз увидеть предмет своей страсти. Впрочем, уже 21 июля он записывает: «Сегодня, к великому моему прискорбию, я окончательно узнал, что я не пара Ел.Ив. и по какому глупому обстоятельству: потому что не говорю по французски, а ее отец умеет говорить только по французски и итальянски». Взявшись за изучение французского языка, наш герой далеко не продвинулся и 26 июля записывает: «Моя любовь к Дж. кажется начинает охлаждаться; по крайней мере я не так часто об ней мечтаю; но я все ее люблю».

Вторая любовная история изложена в четырех записях за январь 1867 года. В.И. Чемезов продолжает обучение в Академии. Дневник был надолго заброшен, и свое возвращение к дневниковым записям он объясняет такими словами: «Я думаю снова начать записывать, чтобы легче впоследствии обсудить, что я за человек, куда гожусь и как себя держать в обществе, чтобы быть хоть сколько нибудь для него нужным. Под словом общество я разумею простое светское общество. Я, вообще, хочу записывать свое умственное и если можно выразиться так, чувственное развитие. Это разумеется не [нрзб]можно сделать, как строго записывать своим мысли и чувства, и чрез несколько времени прочитать, чтобы увидеть, что я пережил. В наст. время меня интересует, что за чувства я питаю к О.М.Т. Мне кажется, подобные записки принесут ту пользу, что, прочитав их, я составлю себе правила к[a]к держаться в обществе; но об этом рассуждение впереди»[4, запись от 23 января (11 января) 1867]. Далее следуют обширные, многословные и весьма сумбурные описания развития чувств к некоей О.М.Т., чье имя так и не будет названо. Здесь меняется даже вид дневника. Если ранее записи располагались в две ровные колонки, то теперь колонок нет, запись занимает весь лист. Текст полон сокращений, помарок и исправлений.

Итогом размышлений и самоанализа стали следующие умозаключения: «Мне кажется, что все надо предоставить течению судеб, держать себя к[а]к я решил прежде, [нрзб] приводить в исполнение мой девиз: Вежливость и благородство, кроме того к[а]к можно меньше говорить о себе, быть естественным, веселым. … Мне кажется, что в уме я себя никогда не сделаю мужчиной, т.е. человеком, к[ото]рый привык себя держать известным образом, к[ото]рый на дне ежесекундно помнит, что он всегда д-н быть естеств., вежлив и благороден, одним словом я умом не выделаю из себя человека установившегося, но что таким я могу сделаться только часто обращаясь в обществе, т.к. что [нрзб] я сделаюсь окончательно мужчиной не ранее, к[а]к чрез год. когда буду иметь более свободного времени, чтоб завести знакомства» [4, запись от 23 января 5 февраля (24 января) 1867].

Оставшиеся три темы весьма близки между собой по представленности в тексте дневника.

Тема номер 3 охватывает 12,2% текста, преобладает в 13 дневниковых записях. Как и четвертая тема, она сравнительно равномерно распределена по тексту. Наиболее значимые для этой темы слова – гимназия, воля, сердце, польза, папа. Записи, в которых эта тема преобладает, содержат самоанализ автора, его психологические рассуждения и переживания. Ярким примером может быть запись от 4 апреля 1862 года, где тема представляет 99% текста. «Пятый день пришел с тех пор как нас отпустили, а я повторил только половину арифметики. Мне кажется, что ко 2му Мая я кроме математики не повторю ничего другого. Это будет худо. Надо торопиться. Сегодня я исповедовался, да успел столько нагрешить, что хоть сейчас идти снова исповедоваться. Завтра день моего рождения и причастия. Завтра минет мне 17 лет. Довольно, чтобы сделать что-нибудь полезное, научиться хоть чему-нибудь, а как посмотришь так выходит, что ничего не сделано хорошего, ничему не научился. Грустно, право грустно, да что станешь делать. Такая уж натура. Понимаю, что худо делаю и все-таки делаю, а отчего? От недостатка воли, а может быть и оттого, что разучился заниматься т.е. не то, чтобы разучился, а и совсем не научился. Как я теперь раскаиваюсь, что не занимался в меньших классах! Раскаяние никогда не поздно — это правда, но оно принесет пользу только тогда, когда будет соединено с исправлением, а в противном случае только лишняя трата времени. Странный право у меня характер. Я не зол и не вспыльчив, а серединка на половинку. Впрочем самому трудно верно разобрать свой характер. Непременно крайность: или слишком беспристрастен, или пристрастен. Середины не бывает. Хочу с сегодняшнего дня вставать часов в 6 и заниматься арифметикой, вообще математикой» [4, запись от 16 апреля (4 апреля) 1862].

Тема номер 6 охватывает 12% текста, как и предыдущая преобладает в 13 дневниковых записях. Наиболее значимые для этой темы слова – собака, золотой, картина, скамейка, замок. Записи, в которых эта тема преобладает, содержат бытовые зарисовки, описания походов на выставки, в театр и даже похорон. Присутствует она в основном в начале дневника. Самым примечательным в их числе оказываются записи с описанием кражи белья из чулана Чемезовых, случившейся в сентябре 1860 года. « Вчера у нас случилось замечательное происшествие. Только что я прихожу из гимназии как мне объявляют что нас обокрали. Признаться я испугался этого. Оказалось что украдены несколько пар юбок новых и три платка из которых два были совершенно новые голландские. Кража произведена утром часу в 10м со взломом замков. В это время в чердаке смежном с нашим, стирали белье чернорабочая и кухарка Таубе. Спрашивается кто же сделал это как не они?» [4, запись от 7 октября (25 сентября)1860]. Далее следуют некоторые интересные подробности составления сыскного объявления и подачи его частному приставу. Пропажа, впрочем, так и не будет обнаружена.

Тема номер 2 охватывает 11% текста, преобладает в 14 дневниковых записях. Наиболее значимые для этой темы слова – студент, университет, академия, попечитель, профессор. Записи, в которых эта тема преобладает, рассказывают о студенческих волнениях 1861 года в Санкт-Петербурге. Именно эти свидетельства дневника более всего привлекли внимание Н.Я. Эйдельмана и подробно им рассмотрены [5]. В январе 1862 года В.И. Чемезов пишет в дневнике: «Я решился записывать решительно все, что говорят об университете; конечно большая часть из этого неправда или преувеличение, но это важно с той стороны, что изо всех рассказов, ходящих по городу, я могу вывести взгляд общества на дело студенчества и понятие его об университете. По крайней мере лет через десяток, на досуге, я прочитаю это и обсужу, правы ли были студенты в своих требованиях или нет и наконец самый мой образ мыслей. Может быть мне даже захочется написать историю этого происшествия, конечно для себя и знакомых, а не для публики, потому что печатать историю, в которой высказано много правды касательно правительства не позволяют [4, запись от 13 января (1 января) 1862]». В дневнике описаны события, которые видел сам автор, о которых узнал с чужих слов, слухи и мнения. Будущий студент и брат студента он всегда в своих рассуждениях встает на сторону студентов: «Студенты отстаивают свои права. То то и худо, что в народе не совсем понимают Университет, и поэтому и сходки студентов, и их требования перетолковать в другую сторону» [4, запись от 11 октября (29 сентября) 1862].

**Визуализация**

Визуализация данных давно перестала быть способом проиллюстрировать какую-либо идею или наглядно представить информацию. Скотт Беринато в книге «Сделай наглядно. Как визуализировать данные понятно и убедительно» выделяет четыре типа визуализаций. В их числе – визуальное исследование, в ходе которого с помощью представления информации в графической форме проверяются гипотезы и выбирается лучший вид визуализации [57, с. 71-74]. Наглядное представление тематической модели позволяет лучше понять структуру текстовой коллекции, распределение тем по ней. В настоящее время существует множество инструментов визуализации тематических моделей [58, 59, с. 64-66]. Мы для визуализации использовали возможности библиотеки ggplot2 [60]. Опробовали несколько вариантов визуализации, с которыми можно познакомится в материалах проекта [39], и в итоге остановились на горизонтальной гистограмме, где для каждой темы отображается насколько она представлена в каждой записи дневника (cм. рис. 3). Малиновая линия – регрессионная сглаженная кривая отображающая тенденцию присутствия темы в дневнике.

{Рисунок3}

Рисунок 3. Гистограмма распределения тем по тексту дневника.

**Заключение**

Подводя итог, отметим, что в построенной нами тематической модели дневника В.И. Чемезова мы выделили 6 тем. Преобладает во всем тексте дневника тема номер 4 – «Учеба в гимназии и сдача экзаменов». Эта тема, а также тема номер 3 – «Самоанализ автора, его психологические рассуждения и переживания» более или менее равномерно распределяются по тексту дневника. Четвертая тема фактически исчезает из записей с началом обучения в Медико-хирургической академии. Остальные темы номер 1, номер 2 и номер 5 довольно явственно локализуются, что не удивительно. Темы номер 1 и номер 2 описывают вполне конкретные события в жизни автора: летний отдых 1861 года и студенческие волнения, случившиеся в том же году. Бытовые зарисовки и многословные описания походов в театр и на выставки присутствуют в начале дневника и исчезают во второй половине записей. Тема номер 5 – «Любовные переживания» напротив преобладает в конце дневника, что нагляднее всего отражает изменение интересов и переживаний автора.