

Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
GRADUATION THESIS

Анализ и визуализация армянских манускриптов

Обучающийся / Student Петросян Анна Мнацакановна

Факультет/институт/клластер/ Faculty/Institute/Cluster факультет
инфокоммуникационных технологий

Группа/Group K34422

Направление подготовки/ Subject area 45.03.04 Интеллектуальные системы в
гуманитарной сфере

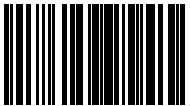
Образовательная программа / Educational program Интеллектуальные системы в
гуманитарной сфере 2020

Язык реализации ОП / Language of the educational program Русский

Квалификация/ Degree level Бакалавр

Руководитель ВКР/ Thesis supervisor Коцюба Игорь Юрьевич, кандидат технических
наук, Университет ИТМО, факультет инфокоммуникационных технологий, доцент
(квалификационная категория "ординарный доцент")

Обучающийся/Student

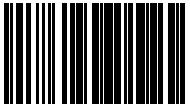
Документ подписан	
Петросян Анна Мнацакановна	
15.05.2024	

(эл. подпись/ signature)

Петросян Анна
Мнацакановна

(Фамилия И.О./ name
and surname)

Руководитель ВКР/
Thesis supervisor

Документ подписан	
Коцюба Игорь Юрьевич	
15.05.2024	

(эл. подпись/ signature)

Коцюба Игорь
Юрьевич

(Фамилия И.О./ name
and surname)

Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University

**ЗАДАНИЕ НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ /
OBJECTIVES FOR A GRADUATION THESIS**

Обучающийся / Student Петросян Анна Мнацакановна

Факультет/институт/клластер/ Faculty/Institute/Cluster факультет
инфокоммуникационных технологий

Группа/Group K34422

Направление подготовки/ Subject area 45.03.04 Интеллектуальные системы в
гуманитарной сфере

Образовательная программа / Educational program Интеллектуальные системы в
гуманитарной сфере 2020

Язык реализации ОП / Language of the educational program Русский

Квалификация/ Degree level Бакалавр

Тема ВКР/ Thesis topic Анализ и визуализация армянских манускриптов

Руководитель ВКР/ Thesis supervisor Коцюба Игорь Юрьевич, кандидат технических
наук, Университет ИТМО, факультет инфокоммуникационных технологий, доцент
(квалификационная категория "ординарный доцент")

Характеристика темы ВКР / Description of thesis subject (topic)

Тема в области фундаментальных исследований / Subject of fundamental research: нет /
not

Тема в области прикладных исследований / Subject of applied research: да / yes

Основные вопросы, подлежащие разработке / Key issues to be analyzed

Техническое задание:

Провести распознавание, анализ и визуализацию оцифрованных армянских рукописей.

Исходные данные к работе:

Данные Бодлианской библиотеки об армянских манускриптах, а также их цифровые копии.

Содержание работы:

В выпускной квалификационной работе рассматривается актуальность анализа
культурологических и исторических данных с помощью методов машинного обучения.
Также работа относится к исследованиям в области цифровой гуманистики. Таким
образом, конечным результатом работы является региональная карта с указанием наиболее
характерного вида культурного наследия, определённого в ходе моделирования текстовых
данных.

Цель работы:

Реализовать обработку и моделирование текстовых данных армянских манускриптов.

Задачи работы:

1. Изучение предметной области и источников по теме исследования,
2. Сбор и систематизация данных, формирование датасета,
3. Распознавание текста: изучение, выбор и применение методов машинного зрения,
4. Анализ текста: изучение, выбор и применение методов анализа текстов,
5. Визуализация текстов.

Рекомендуемые материалы и пособия для выполнения работы:

1. Геворгян Л.П., Пилипосян А.С. Храм Звартноц: о состоянии сохранности объекта всемирного наследия ЮНЕСКО // Журнал «Наследие и современность». – 2019. – Том2, №2. – С. 74–91.
2. Armenian Manuscripts and Printed Books of Digital Bodleian // Открытый электронный источник Бодлианской библиотеки – 2023. URL: <https://digital.bodleian.ox.ac.uk/collections/armenian/>.
3. Universal Dependencies - URL: <https://universaldependencies.org/>.
4. Sample Archival Documents on the Armenian Genocide - URL: <https://www.armenian-genocide.org/sampledocs.html>.

Форма представления материалов ВКР / Format(s) of thesis materials:

Письменный отчёт, презентация

Дата выдачи задания / Assignment issued on: 23.01.2024

Срок представления готовой ВКР / Deadline for final edition of the thesis 24.05.2024

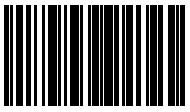
СОГЛАСОВАНО / AGREED:

Руководитель ВКР/
Thesis supervisor

Документ подписан	
Коцюба Игорь Юрьевич	
19.02.2024	

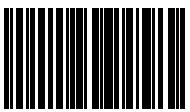
Коцюба Игорь
Юрьевич

Задание принял к
исполнению/ Objectives
assumed BY

Документ подписан	
Петросян Анна Мнацакановна	
28.02.2024	

Петросян Анна
Мнацакановна

Руководитель ОП/ Head
of educational program

Документ подписан	
Хлопотов Максим Валерьевич	
28.05.2024	

Хлопотов
Максим
Валерьевич

**Министерство науки и высшего образования Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО
ITMO University**

**АННОТАЦИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
SUMMARY OF A GRADUATION THESIS**

Обучающийся / Student Петросян Анна Мнацакановна

Факультет/институт/кластер/ Faculty/Institute/Cluster факультет
инфокоммуникационных технологий

Группа/Group K34422

Направление подготовки/ Subject area 45.03.04 Интеллектуальные системы в
гуманитарной сфере

Образовательная программа / Educational program Интеллектуальные системы в
гуманитарной сфере 2020

Язык реализации ОП / Language of the educational program Русский

Квалификация/ Degree level Бакалавр

Тема ВКР/ Thesis topic Анализ и визуализация армянских манускриптов

Руководитель ВКР/ Thesis supervisor Коцюба Игорь Юрьевич, кандидат технических
наук, Университет ИТМО, факультет инфокоммуникационных технологий, доцент
(квалификационная категория "ординарный доцент")

**ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
DESCRIPTION OF THE GRADUATION THESIS**

Цель исследования / Research goal

Реализация обработки и моделирования текстовых данных армянских манускриптов

Задачи, решаемые в ВКР / Research tasks

Изучение предметной области и источников по теме исследования, сбор и систематизация
данных, формирование датасета, распознавание текста: изучение, выбор и применение
методов машинного зрения, анализ текста: изучение, выбор и применение методов анализа
текстов, визуализация текстов.

Краткая характеристика полученных результатов / Short summary of results/findings

В ходе выпускной квалификационной работы достигнута цель, то есть реализовано
распознавание и моделирование текстовых данных армянских рукописей, и выполнены
задачи, поставленные в начале исследования: изучена предметная область, сформирован
датасет, распознан, проанализирован и визуализирован текст.

Наличие публикаций по теме выпускной работы / Publications on the topic of the thesis

1. Петросян А.М. Электронный сборник тезисов докладов XIII КМУ - 2024 (Тезисы)

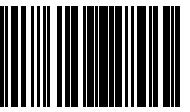
**Наличие выступлений на конференциях по теме выпускной работы / Conference
reports on the topic of the thesis**

1. XXII Международная конференция молодых ученых «Векторы», 18.04.2024 - 21.04.2024
(Конференция, статус - международный)

2. Гуманитарные проблемы актуальных наук: цифровая дисциплина и проект, 15.04.2024 -
17.04.2024 (Конференция, статус - всероссийский)

3. XIII Конгресс молодых ученых ИТМО, 08.04.2024 - 11.04.2024 (Конгресс, статус -
всероссийский)

Обучающийся/Student

Документ подписан	
Петросян Анна Мнацакановна	
15.05.2024	

(эл. подпись/ signature)

Петросян Анна
Мнацакановна

(Фамилия И.О./ name
and surname)

Руководитель ВКР/
Thesis supervisor

Документ подписан	
Коцюба Игорь Юрьевич	
15.05.2024	

(эл. подпись/ signature)

Коцюба Игорь
Юрьевич

(Фамилия И.О./ name
and surname)

СОДЕРЖАНИЕ

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ.....	6
ВВЕДЕНИЕ.....	7
1 Обзор предметной области.....	9
1.1 Культурное наследие Армении.....	9
1.2 Обзор проблематики.....	12
1.3 Объект и предмет исследования, детализация задач и гипотезы.....	15
1.4 Итоги раздела 1.....	18
2 Моделированные данные.....	19
2.1 Формирование датасета.....	19
2.1.1 Выбор источников по теме исследования.....	22
2.1.2 Сбор данных.....	22
2.1.3 Описание данных.....	23
2.2 Распознавание текстовых данных.....	24
2.2.1 Выбор метода распознавания текстов.....	24
2.2.2 Предобработка изображений и сегментация.....	24
2.2.3 Реализация распознавания и оценка результатов.....	27
2.3 Итоги раздела 2.....	29
3 Анализ и визуализация текстовых данных.....	30
3.1 Лингвистическая специфика.....	31
3.1.1 Проработка гипотез о появлении новых символов и диалектах....	31
3.1.2 Проработка гипотезы о стратификации языковых конструкций...	33
3.2 Географическо-тематическая сегментация.....	34
3.3 Итоги раздела 3.....	37
ЗАКЛЮЧЕНИЕ.....	38
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....	41

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

РА – Республика Армения

ЮНЕСКО – United Nations Educational, Scientific and Cultural Organization
(специализированное учреждение Организации Объединённых Наций по вопросам образования, науки и культуры)

API – Application Programming Interface (способ взаимодействия одной компьютерной программы с другими)

ATR – Automatic Text Recognition (автоматическое распознавание текста)

BI – Business intelligence (компьютерные методы и инструменты для организаций)

CER – Character Error Rate (метрика точности распознавания символов)

CRISP-DM – Cross-Industry Standard Process for Data Mining (методология по исследованию данных)

DH – Digital Humanities (цифровая гуманитаристика)

HTR – Handwritten Text Recognition (распознавание рукописного текста)

LSTM – Long short-term memory (долгая краткосрочная память)

OCR – Optical Character Recognition (оптическое распознавание символов)

OEM – Optical Engine Mode (режим оптического движка)

PCM – Page Composition Mode (режим компоновки страницы)

PDF – Portable Document Format (межплатформенный открытый формат электронных документов)

ВВЕДЕНИЕ

Анализ культурных данных с точки зрения цифровой гуманитаристики является одной из основных её предметных областей и способствует исследованию и сохранению культурных и исторических цивилизаций. Несмотря на то что на первый взгляд может показаться, что культурные данные тяжело поддаются формализации, объединение гуманитарного и технического знаний позволяет реализовывать проработанные аналитические исследования, строить математические модели и выявлять закономерности и особенности.

Выбранная тема актуальна с точки зрения как цифровой гуманитаристики, так и выбранного региона:

– изобилие материала для анализа, ведь Армения, расположенная в колыбели цивилизации, может похвастаться разнообразным и древним культурным наследием, который необходимо проанализировать исходя из неоднозначной культурной и географической сегментацией, что требует количественного анализа,

– специфичность и хрупкость объектов культурного наследия ввиду их древности, что также выдвигает особые требования к анализу изображений, распознаванию текстов и их анализу,

– совсем недавно Бодлианская библиотека [1] при поддержке Нью-Йоркской Корпорации Карнеги оцифровала больше сотни армянских манускриптов, добавив к многим из которых метаданные. Это подтверждает интерес мировых научных институций к армянскому региону в рамках цифровой гуманитаристики. Именно эти данные служат основой исследования.

Цель работы – реализовать обработку и моделирование текстовых данных армянских манускриптов. Для этого необходимо провести распознавание, анализ и визуализацию оцифрованных рукописей.

Задачи, решаемые в работе:

- изучение предметной области и источников по теме исследования,
- сбор и систематизация данных, формирование датасета,
- распознавание текста: изучение, выбор и применение методов машинного зрения, оценка предложенных методов распознавания,
- анализ текста: изучение, выбор и применение методов анализа текстов, в том числе, построение количественных гипотез, data mining,
- визуализация текстов.

Таким образом, данная предметная область имеет свою специфику и требует углублённого анализа, моделирования, построения гипотез, что и обуславливает актуальность тематики.

1 Обзор предметной области

Цифровая гуманитаристика занимается исследованием социальных, исторических, лингвистических, философских, искусствоведческих вопросов с помощью математических и компьютерных наук. Одно из основных назначений этой междисциплинарной области – обеспечение сохранности культурного наследия. В задачи входит и трансформация данных в цифровой вариант, и сбор уже оцифрованных данных, но наиболее популярное направление на данный момент – анализ данных. Таким образом, исследования цифровой гуманитаристики позволяют расширить использование технологий, которые обычно применяются на очевидно поддающихся математической интерпретации данных (например, экономических показателях или любых других легко категоризуемых данных).

Культурное наследие является показателем духовного и материального достоинства человечества. Как в генетике есть понятие о наследственности и изменчивости, так и культура сегодняшнего дня, несмотря на перманентные трансформации, базируется на культурном опыте предыдущих поколений. Именно по причине своего влияния на формирование поколений культурное наследие является предметом регулирования во многих государствах. Соответственно, встаёт вопрос о его сохранении. Формы сохранения могут быть разными – от охранительно-запретительных до созидательно-репродуктивных. Анализ данных как раз можно считать инструментом последней.

1.1 Культурное наследие Армении

Будучи одним из древнейших центров мировой цивилизации, Армения является плодотворной почвой для анализа данных. История культуры армянского народа берёт начало с VI–V веков до нашей эры и является продолжением ещё более древней культуры Урарту. Во многом

определяющим моментом стало принятие христианства на государственном уровне в 301 году нашей эры, что не могло не повлиять на материальную и нематериальную культуру.

Что касается материальной культуры, например, в части архитектуры традиционной считаются храмы, базилики и монастыри [2]. В настоящее время по всей Армении в разных ее частях находятся уникальные памятники культуры, многие из которых включены в список Всемирного наследия ЮНЕСКО [3], например:

– Эчмиадзинский кафедральный собор (рисунок 1), основанный в 303 году нашей эры, являющийся одним из древнейших христианских храмов мира. Кафедральный собор Эчмиадзина (основан 1700 лет назад) вместе с тремя древними церквями, также как и руины храма в Звартноце, являются выдающимися памятниками армянской церковной архитектуры. Они наглядно иллюстрируют развитие и совершенствование армянских церквей крестово-купольного типа, оказавших основополагающее влияние на архитектуру и искусство этого региона [3],

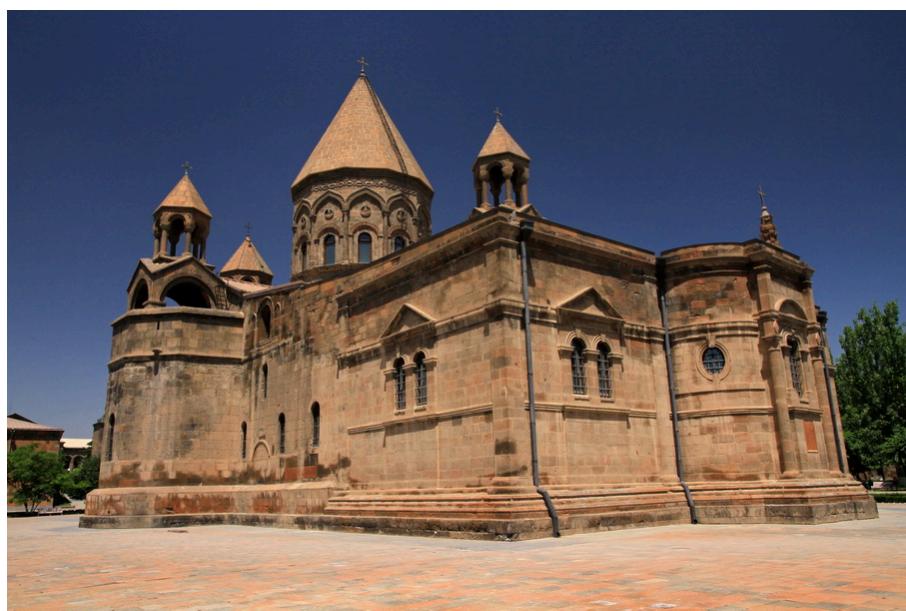


Рисунок 1 – Эчмиадзинский кафедральный собор [4]

– Монастырь Гегард (рисунок 2), который в переводе с армянского носит название “Монастырь Копья”, так как некогда здесь хранилась одна из

ключевых для христианства реликвий, а именно копьё Лонгина. Древние церкви и могилы монастыря Гегард, часть которых высечена прямо в скалах, представляют собой шедевры средневековой армянской архитектуры. Ансамбль монастырских построек органично вписан в великолепный природный ландшафт верховьев реки Азат, и окружен скалами, напоминающими своими формами башни [3].

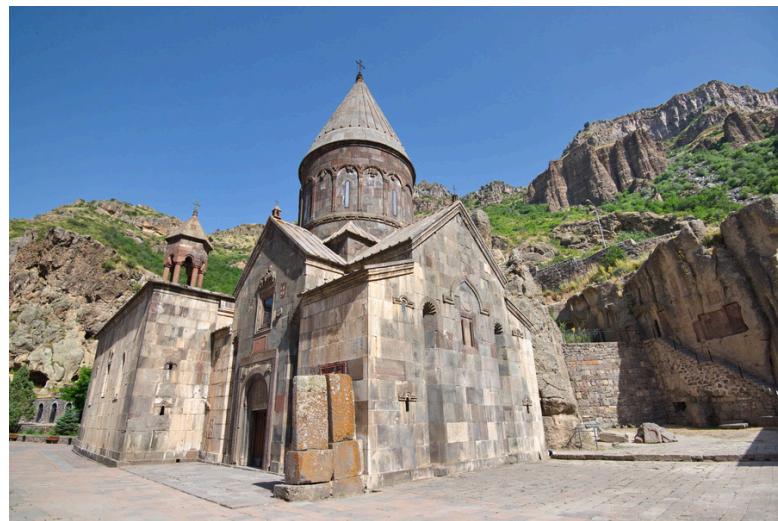


Рисунок 2 – Монастырь Гегард [4]

– Монастырь Санаин (рисунок 3), который в своё время являлся культурно-образовательным центром, так как здесь располагалась мастерская по переписи рукописей. Имеет уникальный стиль которой сформировался при смешении византийских и местных кавказских традиций.



Рисунок 3 – Монастырь Санаин [4]

Что касается нематериальной культуры, то в список [5] входят как бытовые элементы жизни человека, например, приготовление лаваша, кузнечное мастерство, так и связанные с искусством, например, музыка дудука или танец кочари. Само по себе армянское письмо, созданное в 405 году нашей эры, также было внесено в список нематериального культурного наследия в 2019 году, а несколькими годами ранее создавался корпус для восточного диалекта [6].

1.2 Обзор проблематики

Памятниками армянской письменности являются в том числе рукописи. В оцифрованном виде (хорошего качества) находится более 100 книжных объектов общим объёмом более 30000 страниц. В связи с этим возникает необходимость упорядочивания, систематизации и анализа накопленных данных. Армянские рукописи действительно являются объектом интереса учёных, о чём говорят исследования, связанные с ними:

- учёные Оксфордского и Эдинбургского университетов в статье “Сравнение эффективности гиперспектральной визуализации и рамановской спектроскопии: тематическое исследование на армянских рукописях” [7] отмечают необходимость изучения физических характеристик (материала и цвета) рукописей как маркеров времени и источника знаний о применяемых в то время технологиях графики. Это исследование помогло рассмотреть особенности цветопередачи и цветокоррекции,
- также исследуются колофоны [8], то есть тексты на последней странице рукописи, в которых сообщаются данные об авторе, времени и месте создания этого произведения, которые являются первоначальной информацией о манускриптах, а в терминологии анализа данных, по сути, – метаданными,
- если точечно касаться распознавания текстов, то в рамках этого аспекта важно упомянуть Международную конференцию по анализу и

распознаванию документов, в рамках которой представлено множество работ по распознаванию рукописей, в том числе и армянских – в статьях “Модульная и автоматизированная платформа аннотаций для рукописного текста: оценка языков с ограниченными ресурсами” [9] и “Пост-коррекция армянских текстов с использованием нейронных сетей” [10] описываются различные этапы работы с текстом и компьютерной лингвистикой в целом. Например, учёные утверждают, что ручное аннотирование рукописных документов является трудоёмкой задачей, особенно когда данные специфичны. В работах рассматривается новый модульный и готовый к использованию онлайн-интерфейс для многоуровневого аннотирования и быстрого просмотра рукописных и печатных документов, в том числе для языков с письмом справа налево. Этот интерфейс позволяет создавать индивидуальные проекты, а также управлять, преобразовывать и экспортить данные в различные форматы и современные стандарты,

– объединив историческое, лексикографическое и техническое знания, учёные (в основном) французских образовательных организаций разрабатывают сервис, который ориентирован преимущественно на восточные языки с собственной (не латинской) системой письменности (в том числе, армянский, грузинский, греческий, но не ограничиваясь ими). В статье “От рукописи к размеченной корпоре: автоматизированный процесс для древних армянских и других малообеспеченных языков христианского Востока” [11] они описывают основные механизмы работы с древними историческими источниками. В статье также отмечается, что в настоящее время системы оптического распознавания символов были адаптированы для распознавания рукописного текста (HTR) и используются в ряде цифровых гуманитарных проектов, а для исторических документов обычно достигают точности 95% или выше, даже при использовании сценариев с ограниченными ресурсами и сложных макетов,

– самой заинтересованной в анализе рукописей институцией в Армении является Музей-институт древних рукописей имени Месропа Маштоца – “Матенадаран” (последнее слово переводится с армянского как “хранилище рукописей”). Как отмечается на сайте музея [12], одна из его главных миссий – это описание, подробное изучение и издание каталога хранящихся в нём рукописей, которых больше 17 тысяч на разных языках и больше 10 тысяч на армянском. Проверкой, вычиткой и редактурой занимаются с 2007 года около 30 специалистов. Создание такого генерального каталога приведёт к тому, что рукопись предстанет перед читателем как памятник материальной и духовной культуры.

Сотрудники Матенадарана, отмечают, что применение интеллектуальных информационных технологий помогло бы не только ускорить работу по созданию генерального каталога, но и создать корпус армянского языка по манускриптам, таким образом, корпус получился бы узкоспециализированным. Также они отметили, что географическая сегментация помогла бы в вопросе изучения неочевидных аспектов в качестве структурированной визуализации в просветительских или образовательных целях, особенно с учётом того, что армянская письменность включена в список нематериального культурного наследия ЮНЕСКО.

Принимая во внимание историчность темы, обширность пula объектов культурного наследия и количество исследуемых данных, проблематику работы можно сформулировать как культурную стратификацию на основе древних рукописей по таким направлениям как: выявление различных типов материальной и нематериальной культуры, типологизация культурной деятельности людей (например, выявление различных типов профессиональной, религиозной, языковой, региональной культуры или культуры, связанной с историческими этапами развития общества).

1.3 Объект и предмет исследования, детализация задач и гипотезы

Таким образом, объектом исследования являются рукописи как источник знаний о регионально-культурных особенностях, а предметом – применение методов анализа культурных данных для культурной стратификации.

Если говорить об опыте анализа текстов рукописей, то такие проекты существуют:

– Факультет вычислительных наук и данных Парижского университета (Сорбонны) [13], где занимаются цифровым анализом иконографических корпусов, а именно разработкой инструментов анализа и новых инструментов компьютеризированной обработки и интеграцией этих инструментов в средах цифровых изданий и платформах, которые могут служить, например, историческим словарём, потому как реализован поиск по слову по оцифрованным рукописным материалам (рисунок 4). Основная миссия проекта – сформулировать два взаимодополняющих исследовательских подхода к данным из корпусов текстов и изображений: локальный анализ и детальное описание этих корпусов (например, критических изданий) и использовать цифровые исследовательские инструменты, более ориентированные на количественный анализ, а также разработать методы обучения, применимые к более крупным наборам корпусов,

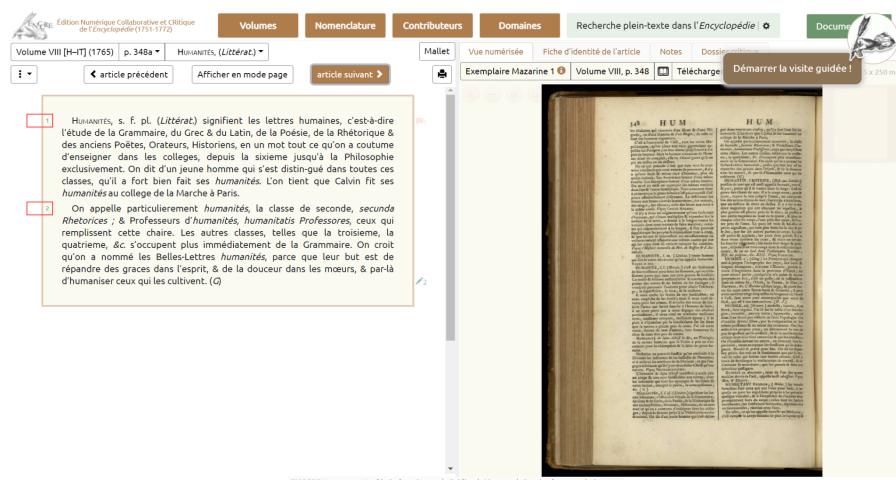


Рисунок 4 – Поиск в историческом словаре [14]

– Университет Индианы [15], чьи учёные придали песеннику Франческо Петrarки цифровой формат (рисунок 5) и добавили различные метаданные, благодаря которым студенты могут быстрее изучать творчество итальянского поэта. Исследователи руководствовались принципами простоты и удобства использования, ведь проект изначально реализован для студентов. Команда подчёркивает, что ею создан инновационный подход в цифровой работе с историческими документами, который четко проводит различие между историей и повествованием, между материальными ресурсами и ненадежными анекдотическими рассказами.

80 **C**hi è fermato di menar sua vita,
 Su per l'onde fallaci et per li scogli
 Scervo da morte con un picciol legno
 Non pò molto lontan esser dal fine;
5 Però sarrebbe da ritrarsi in porto
 Mentre al governo anchor crede la vela.
§ L'aura soave, a cui governo et vela,
 Commisi entrando a l'amorosa vita
 Et sperando venire a miglior porto,
10 Poi mi condusse in più di mille scogli;
 Et le cagion' del mio doglioso fine
 Non pur d'intorno avea, ma dentro al legno.
§ Chiuso gran tempo in questo cieco legno 35
 Errai senza levar occhio a la vela
 Ch'anzi al mio dì mi trasportava al fine;
 Poi piacque a lui che mi produsse in vita
 Chiamarme tanto indietro da li scogli
 Ch'al men da lunga m'apparisce il porto.
§ Come lume di notte in alcun porto

Vid'io le 'nsegne di quell'altra vita,
 Et allor spirrai verso 'l mio fine.
§ Non perch'io sia secolo anchor del fine,
 Ché volendo col giorno esser a porto
 È gran viaggio in così poca vita;
 Poi temo, ché mi veggio in fraile legno,
 Et più che non vorrei piena la vela
 Del vento che mi pinse in questi scogli.
§ S'io esca vivo de' dubbiosi scogli,
 Et arrive il mio exilio ad un bel fine,
 Ch'i' sarei vago di voltar la vela,
 Et l'anchore gitar in qualche porto!
 Se non ch'i' ardo come acceso legno,
 Si m'è duro a lassar l'usata vita.
§ Signor de la mia fine et de la vita,
 Prima ch'i' fiacchi il legno tra li scogli
 Drizza a buon porto l'affannata vela.

Рисунок 5 – Оцифрованные песенники Франческо Петrarки [16]

Однако даже по рисункам 4-5 видно, что такие проекты представляют размеченный распознанный текст, однако анализ культурных особенностей в них отсутствует.

Одним из подходов в типологизации культур является её разделение на материальную и нематериальную [17]. Объектами первой являются предметы труда и материального производства, быта, топоса, то есть материальная культура является совокупностью овеществленных результатов человеческой деятельности. Объекты нематериальной, или духовной, культуры – идеи,

мысли, язык и культура речи, религия, искусство. Другая типологизация разделяет культуру на экономическую, правовую, экологическую, эстетическую, поведенческую, и всё это является атрибутами в задачах классификации.

Культуры бывают разные, но в специфике Армении мы берём географическую сегментацию и культуру материальную (как рукописи как частный случай общей типологии).

Соответственно, если детализировать задачу по анализу данных, то из поставлены две подзадачи, касающиеся лингвистического и географико-тематического моделирования:

- анализ манускриптов с точки зрения лингвистической специфики в части выявления символных и диалектных особенностей,
- географико-тематическая сегментация: анализ таких манускриптов как письменных источников с точки зрения количественного анализа (классификации) в целях рассмотрения плотности скопления армянских манускриптов на различных территориях и изучения неочевидных по причине современных границ Армении регионов создания рукописей.

Для первой подзадачи (связанной с компьютерной лингвистикой), поставлены следующие гипотезы:

- потери в точности распознавания связаны с заменой букв,
- можно выделить разные диалекты, которые отличаются написанием,
- определённые виды языковых конструкций появляются в контексте определённого вида деятельности, то есть социально стратифицированы.

Для второй подзадачи возникает необходимость проверить регионально-видовую зависимость культурного наследия, а также какого вида культурного наследия (материального или нематериального) представлено больше. Следовательно, нужно проверить следующие гипотезы:

- в регионах создания, которые не относятся к современным границам Армении, доминирует религиозная тематика,

- виды занятий были сугубо сельскохозяйственного или промышленного характера в соответствии с этапом исторических обществ,
- на протяжении исследуемого периода была широко представлена фольклорная тематика.

1.4 Итоги раздела 1

Раздел 1 посвящён обзору предметной области, в рамках которого:

- установлено, что армянский регион действительно представляет интерес с точки зрения анализа данных благодаря богатой истории, накопленному объёму культурных данных и их разнообразию, о чём свидетельствуют уже существующие научные работы,
- сформировано понимание об объектах культурного наследия, признанных мировой общественностью, для дальнейшего анализа,
- сформулированы первичные гипотезы, проблематика и конкретизированы требования к дальнейшему анализу.

2 Моделированные данные

2.1 Формирование датасета

Датасет – это набор данных, чаще всего собранный для какого-то определённого исследования, имеющий общую тематику и являющийся репрезентативным для заданной темы. При этом не стоит путать данные и датасет – именно структурированность присуща датасету, а хаотичность – данным (рисунок 6).



Рисунок 6 – Разница между данными и датасетом [18]

Выделяют разные методы формирования датасетов [19]:

- ручной с помощью различных методов сбора данных, например, наблюдения (в реальной жизни или социальных сетях), интервью или тесты (с закрытыми или открытыми вопросами),
- автоматизированный сбор из открытых электронных источников с

использованием открытых API, собственных скриптов, написанных специально под страницу, или готовых инструментов парсинга,

– краудсорсинг, который совмещает методы сбора данных по типу наблюдения или интервьюирования, но автоматизирован в плане технической части. Например, окно с просьбой оставить оценку о работе сайта или оценить качество распознавания автоответчика, которое, по сути, является разметкой данных.

С технической точки зрения и в рамках машинного обучения датасет должен обладать: репрезентативностью, симметричностью, полнотой, однородностью и качеством данных. Репрезентативность заключается в отражении разнообразия характеристик и условий, присущих целевой генеральной совокупности, симметричность – в одинаковом распределении относительно среднего значения, полнота – в достаточном объёме данных, однородность – в согласованности и структурированности, а качество – в надёжности, точности и актуальности. Также данные должны быть этичными, то есть не нарушать законодательство.

В статье “Сбор и обработка исторических данных в автоматизированных информационных системах” [20] описываются методы работы с датасетами, например, предобработка (избавление от аномальных данных, пропусков, исправление ошибок, нормализация), исследовательский анализ на распределения и корреляции для понимания структуры данных и специфики работы с ними.

Так как данные собирают для их анализа, то важным этапом является оценка точности результатов анализа, проведённого с данными. Метрики могут быть разными, например, среднеквадратичная ошибка и коэффициент детерминации для регрессии; средняя абсолютная ошибка для сравнения настоящих и прогнозируемых данных; полнота и точность для задач классификации. После этого можно переходить к визуализации данных, которая позволяет преподнести результаты широкой публике.

Как заявляют авторы статьи “Модификация алгоритма CRISP для визуализации лабораторных данных на платформах класса ВI” [21], в целом, такой подход является основополагающим в методологии CRISP-DM [22] (рисунок 7).



Рисунок 7 – Схема работы с датасетами [21]

В доработанной авторами схеме предлагается в этап «Анализ данных» добавить шаг «Форматирование данных» для их разметки, разделения и преобразования, что полезно в работе с таблицами для их загрузки в аналитический контур модели (рисунок 8).

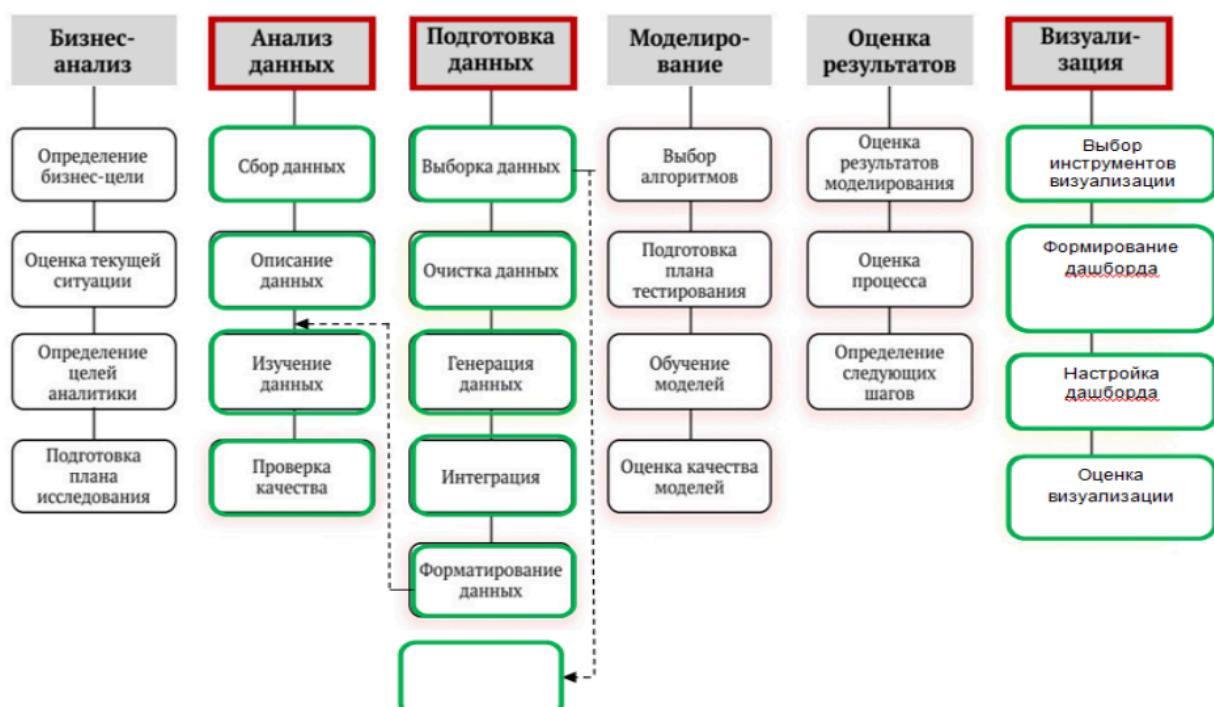


Рисунок 8 – Доработанная схема работы с датасетами [21]

2.1.1 Выбор источников по теме исследования

В качестве источника данных выбрана Бодлианская библиотека Оксфордского университета, уже упомянутая во введении, которая имеет в своей коллекции больше нескольких сотен рукописей от западных и мезоамериканских до северо-африканских и южно-азиатских [23]. В текущие проекты этой институции входит и работа с армянскими рукописями (рисунок 9). Именно в этой коллекции рукописи имеют хорошее качество оцифровки.

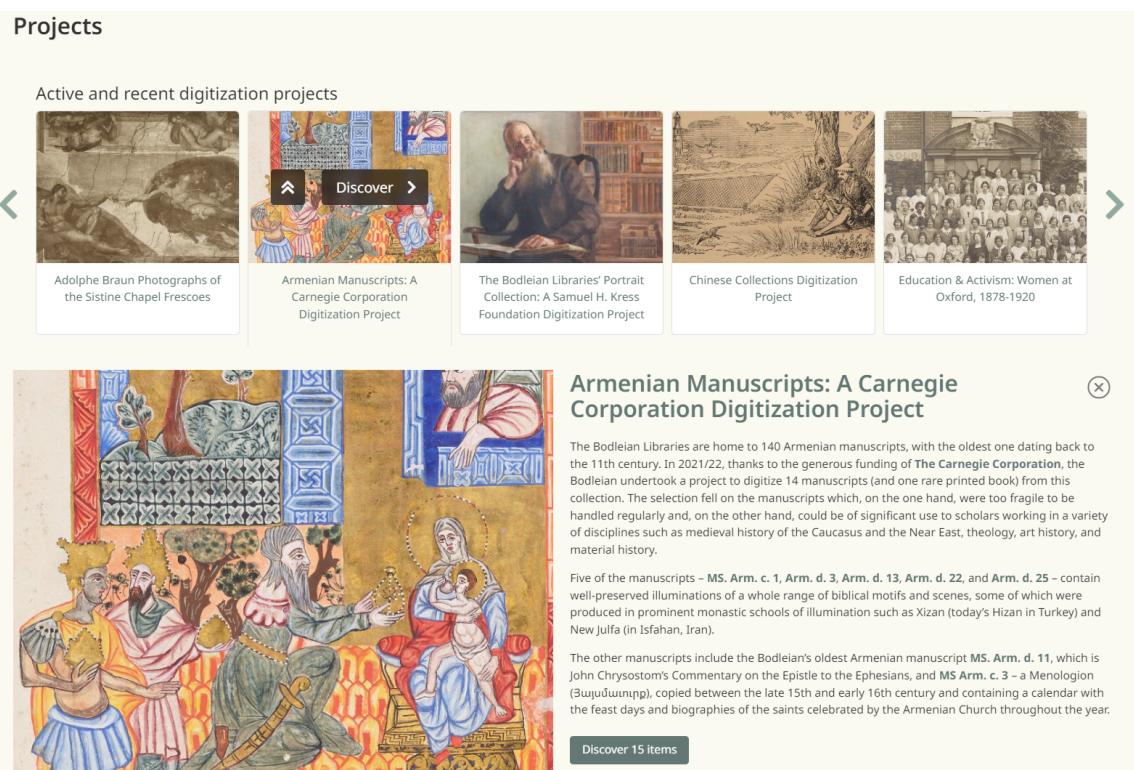


Рисунок 9 – Текущие проекты Бодлианской библиотеки [23]

В качестве дополнительных данных выбрана электронная коллекция Матенадарана, в которой сканы рукописей хоть и не выложены, но имеются некоторые статьи по теме. Эти данные используются как справочно-пояснительные, в том числе для оценки качества распознавания.

2.1.2 Сбор данных

Данные собраны с помощью браузерного расширения для автоматизированного сбора данных WebScrapper [24]. Он обладает

достаточным для этого сбора функционалом и позволяет экспортить данные в табличном формате.

Для этого необходимо настроить схему так называемых “селекторов”, то есть индикаторов для выбора элемента с веб-страницы. Внутри инструмента эти элементы разделяют по типу данных, основные из которых – текстовый, изображенческий, табличный, html-селектор, а также селектор ссылки. Эти селекторы в большинстве своём совпадают с CSS-кодом страницы. Благодаря этому для страницы, на которой представлены все рукописи, и для страницы самой рукописи нет необходимости создавать разные деревья карты сайта, а можно собирать данные во вложенных страницах, что упрощает формирование датасета. В ином случае пришлось бы совмещать два датасета.

Сбор данных о 95 рукописях с 7 признаками занял 3 минуты. Пример выбора селекторов представлен на рисунке 10.

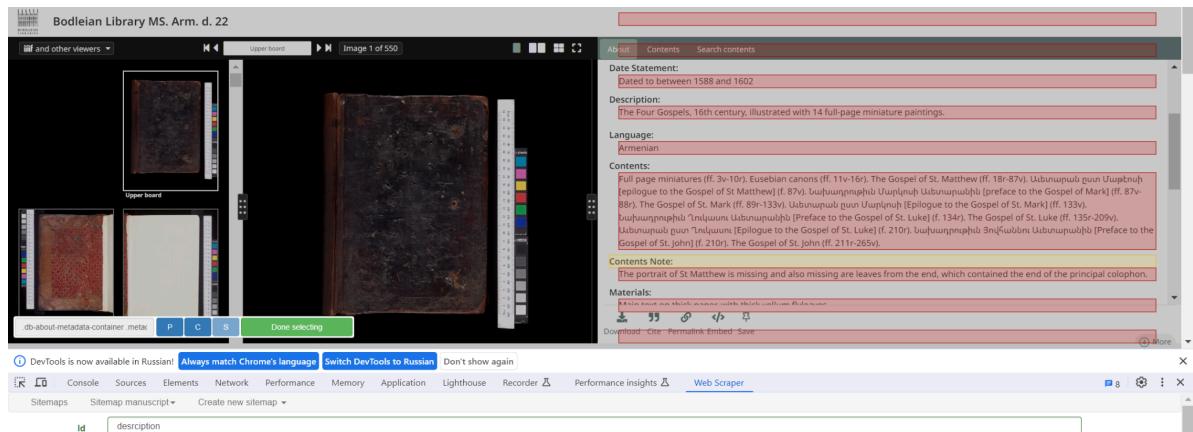


Рисунок 10 – Выбор селекторов во вложенной странице

2.1.3 Описание данных

Так, собран датасет, который содержит данные о названии, где и месте создания, примерном описании содержания (около 3 предложений), языке и ссылке на источник. Также собраны цифровые копии рукописей в формате pdf.

2.2 Распознавание текстовых данных

2.2.1 Выбор метода распознавания текстов

Существуют сервисы по распознаванию рукописных текстов с собственной (не латинской) системой письменности. Однако в данной работе предполагается применение библиотек, а не использование готовых сервисов.

Наибольшее признание [25-29] имеют две библиотеки, а именно Tesseract OCR и Google Cloud Vision. Они регулярно упоминаются в прикладных исследованиях в рамках Международной конференция по анализу и распознаванию документов [28-29]. На основании этих статей и документации инструментов выявлены особенности библиотек, которые прежде всего позволяет оценить применимость технологий в нашем случае, когда инструментов, поддерживающих армянский язык не так много, а именно:

- языковая поддержка: Google Cloud Vision и Tesseract OCR поддерживают указанный язык,
- методы распознавания: Google Cloud Vision применяет свёрточные нейронные сети для классификации объектов на изображениях. В Tesseract OCR используется Optical Character Recognition, которое возможно комбинировать с LSTM-моделями и скрытыми марковскими моделями.

Однако для работы с Google Cloud Vision требуется платное переменное окружение. К тому же Tesseract OCR показывает лучшие результаты [28-29], поэтому для распознавания будет применена эта библиотека.

2.2.2 Предобработка изображений и сегментация

Предобработка является важным этапом, так как упрощает работу алгоритма и способствует его эффективности. Шагами предобработки являются:

- изменение размера изображения с помощью функции “cv2.resize()”,

- конвертация цветового пространства в оттенки серого с помощью функции “cv2.cvtColor()”,
- уменьшение шума с помощью функции “cv2.fastNlMeansDenoising()”,
- улучшение контрастности с помощью аддитивной гистограммной эквализации с ограничением контраста с использованием функции “cv2.createCLAHE()” и метода “CLAHE.apply()”.

На рисунке 11 представлено поэтапное графическое отображение этих шагов.

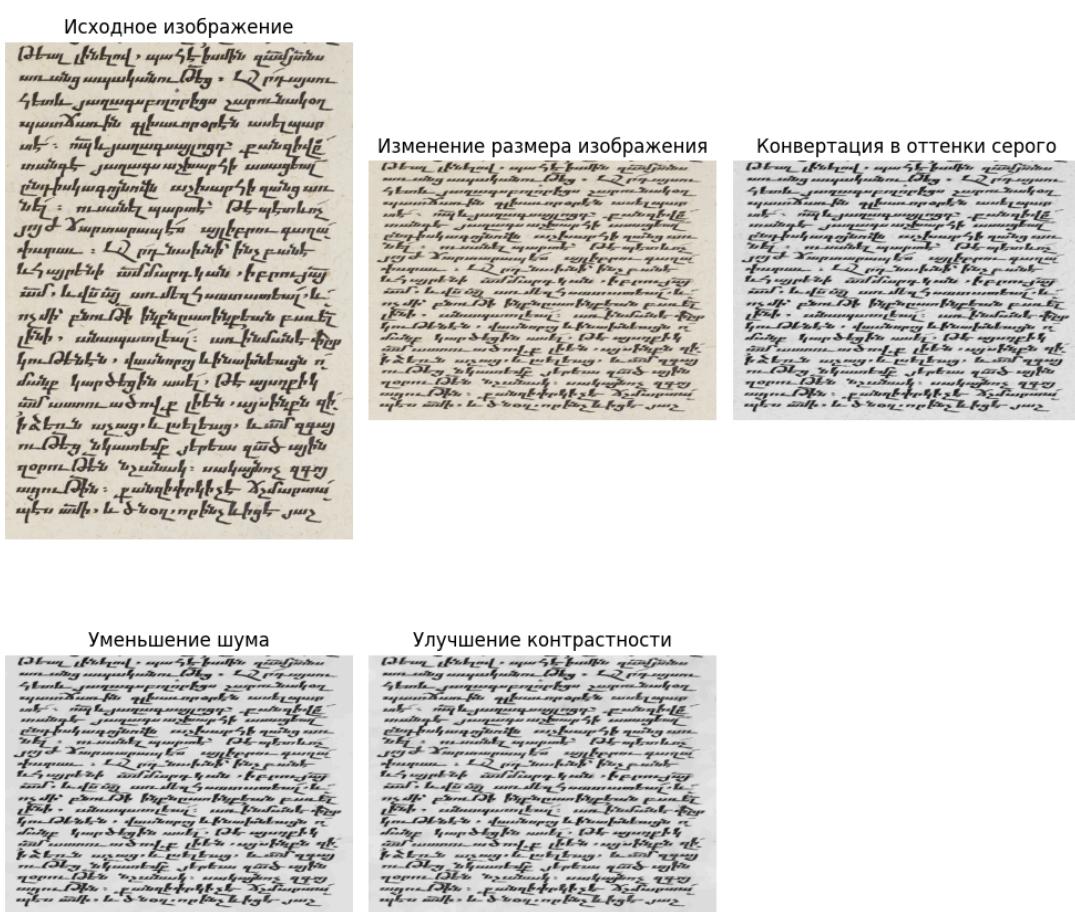


Рисунок 11 – Шаги предобработки

Учтена специфика исторических документов и удалены изображения с низкой степенью информативности (рисунок 12). Таких страниц оказалось меньше 5% и их удаление не повлияло на репрезентативность выборки по географическому признаку. Для этого выведены размеры “потерь” и выбрано пороговое значение с помощью экспертного метода (просмотра страниц с

потерями) выбрано значение в 60 пикселей, что составляет около 2 сантиметров.



Рисунок 12 – Пример страницы с потерями

Что касается сегментации (рисунок 13), то она является частным случаем задач компьютерного зрения. Основной функцией в сегментации строк служит “cv2.findContours()”, которая ищет области, ограниченные кривой, которая образует границу объекта, в нашем случае, строки.

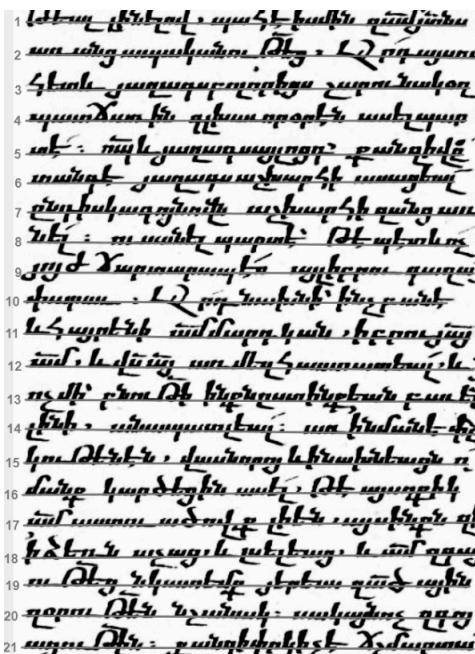


Рисунок 13 – Сегментация строк

2.2.3 Реализация распознавания и оценка результатов

Для распознавания создаётся пользовательская конфигурация, в которой задается язык распознавания (-l hye для армянского языка). Также необходимо задать параметры “--oem 3” и “--psm 6”, которые определяют режимы работы движка Tesseract OCR (OEM 3 – для выбора LSTM-модели OCR и PSM 6 – один текстовый блок без выравнивания). Параметр --oem определяет используемый движок распознавания текста. Возможные значения от обычного посимвольного распознавания до гибридных моделей. Этот режим может использовать как оригинальный движок Tesseract, так и LSTM для распознавания текста. 3 – Default, основанный на модели LSTM (Long Short-Term Memory), это комбинированный режим, использующий LSTM и другие методы для распознавания текста. В большинстве случаев этот режим дает лучшие результаты. “Одним текстовым блоком” является непрерывный набор строк, как на рисунке 7. Если же блоков несколько, как на рисунке 14, то используется режим “--psm 4”.

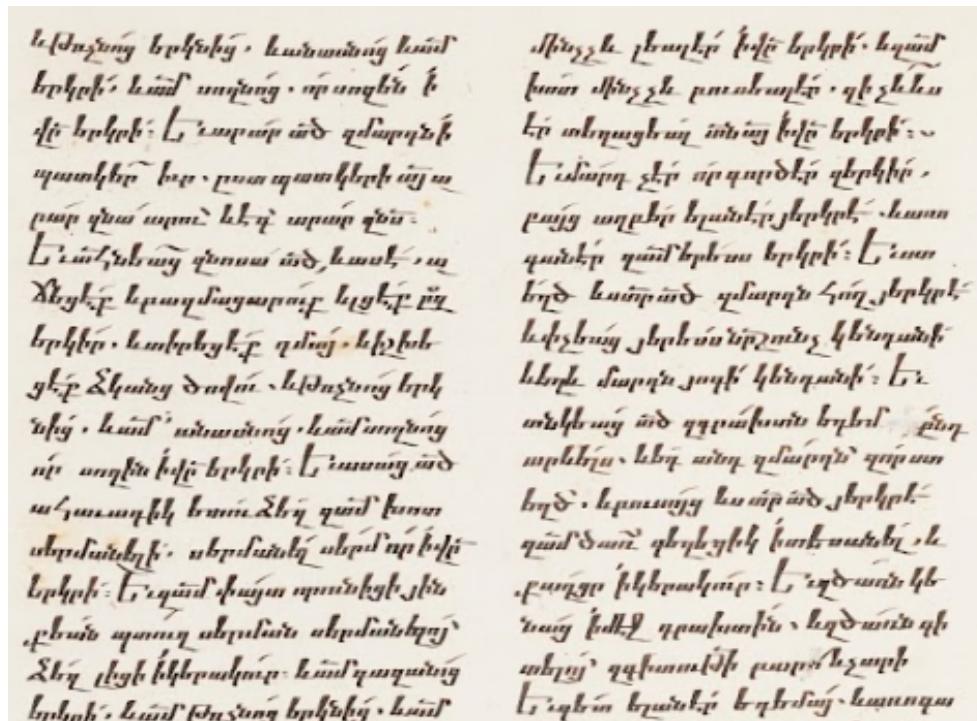


Рисунок 14 – Два текстовых блока

С помощью функции “pytesseract.image_to_string()” из библиотеки pytesseract производится распознавание текста на изображении, используя указанную пользовательскую конфигурацию.

На рисунке 15 представлен пример распознавания.

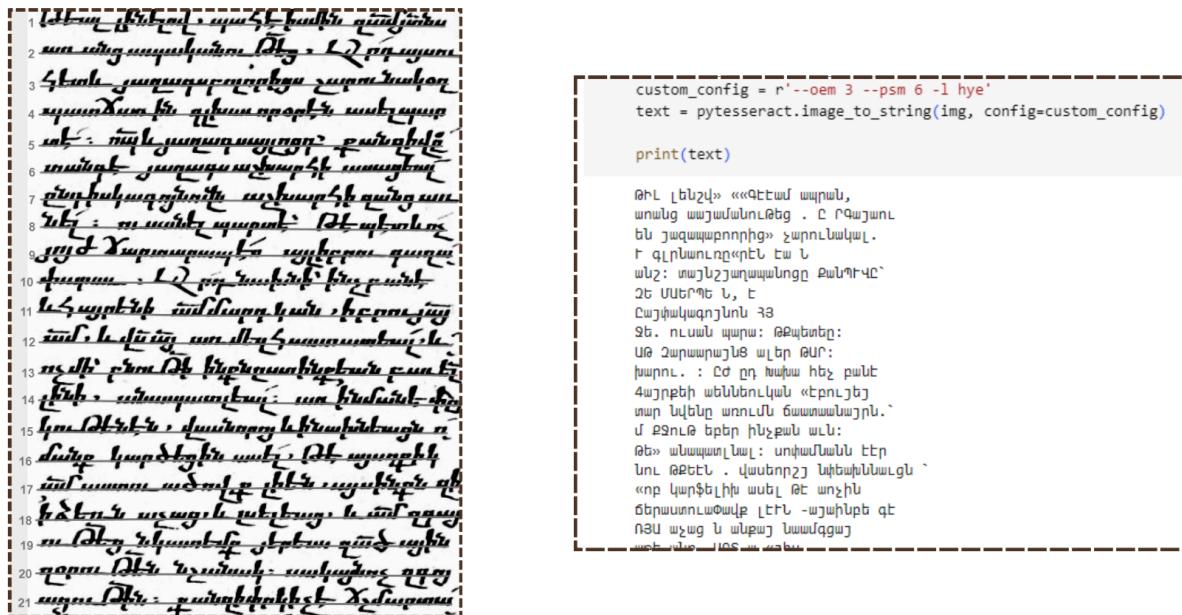


Рисунок 15 – Пример распознавания

В общей сложности распознано около 3000 страниц. В хранилище рукописей Матенадаран есть расширенные специалистами издания (назовём их эталонными), поэтому появилась возможность оценить точность распознавания. С помощью поиска пересечений слов эталонного и распознанного текста рассчитывается отношение количества совпадающих слов на общее количество слов в эталонном тексте и умножает результат на 100 для получения процентного значения. Такая тестовая выборка репрезентативна и составляет 1024 страницы (выбирались из рукописей, которые распознавались и имелись в коллекции Матенадарана), или чуть больше 30%.

Таким образом, точность при этом составляла около 84%, что считается хорошим результатом, так как с точки зрения развития письменности следует учитывать некоторые изменения в части начертания символов.

Поэтому было принято решение проверить, действительно ли это связано с различиями в наборе символов рукописи и инструмента распознавания, и рассчитать метрику Character Error Rate (CER), которая измеряет долю символов, которые были неправильно распознаны моделью, по сравнению с эталонным текстом. Это можно реализовать с помощью расстояния Левенштейна, которое сравнивает распознанную и эталонную строки, а далее делится на общее количество символов в эталонном тексте, чтобы получить нормализованное значение CER в процентах.

Действительно, CER принял значение, близкое к 16%. Такие символы можно наблюдать на страницах рукописей, а лингвисты отмечают, что они были заменены на другие буквы. В следующей главе будет приведена статистика буквенных замен (заменённых символьных употреблений).

2.3 Итоги раздела 2

Раздел 2 посвящён моделированию данных, в рамках которого:

- рассмотрены источники данных,
- собран эмпирический материал для исследования,
- осуществлена предобработка данных,
- изучены и выбраны методы распознавания, предобработки и сегментации изображений,
- проведено распознавание изображений, выбрана метрика оценки и сама оценка его точности. Она реализовано с помощью сравнения пересечений между имеющимися расшифрованными исследователями текстами и результатами автоматического распознавания с помощью Tesseract OCR. Точность распознавания в данном случае удовлетворяет поставленным задачам распознавания Матенадарана.

3 Анализ и визуализация текстовых данных

Визуализация является необходимой в данном случае не только для удобного представления материалов широкой публике, но и структурирования данных, потому что как в коллекции Бодлианской библиотеки, так и в электронном каталоге Матенадаран рукописи выложены просто сборниками (рисунок 16).

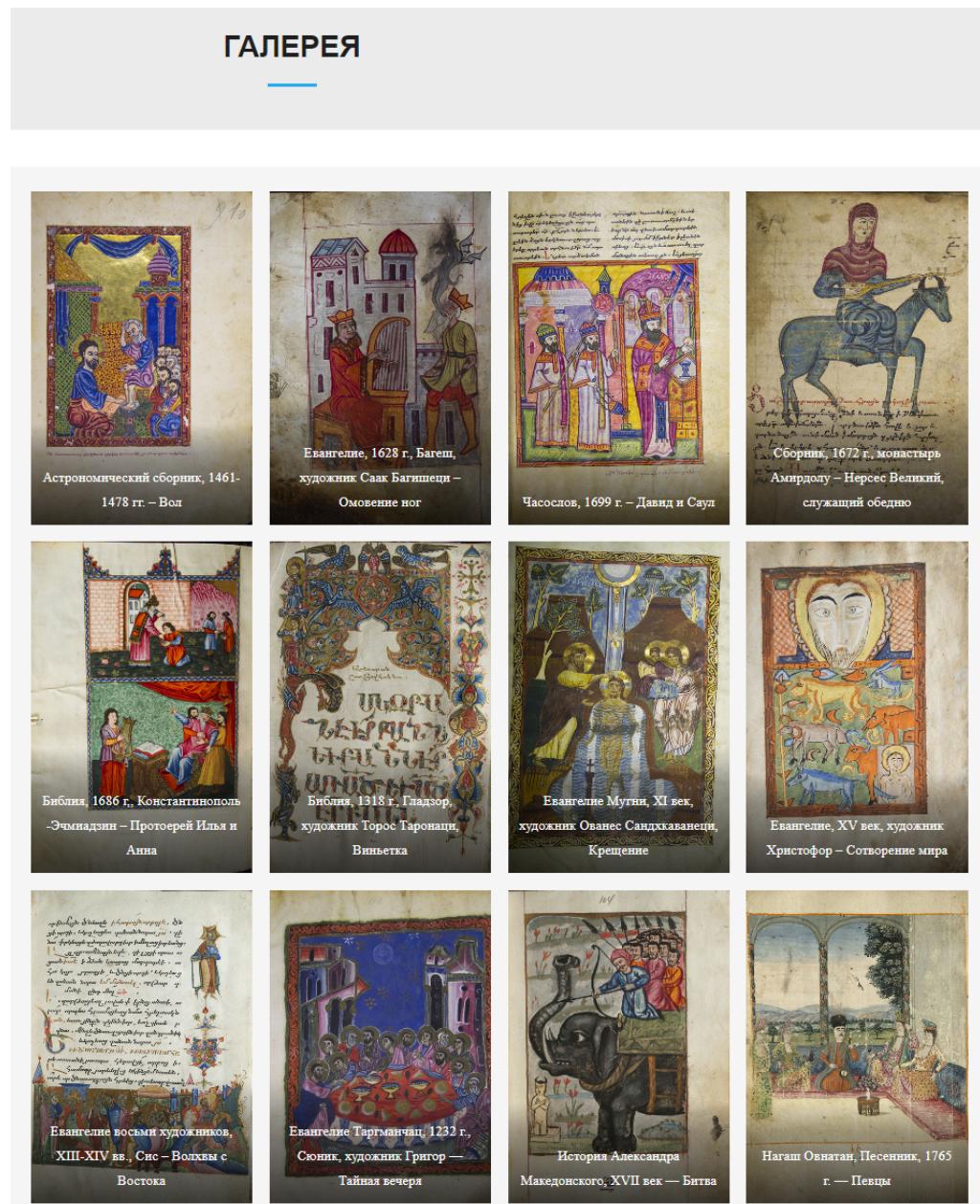


Рисунок 16 – Электронный каталог Матенадарана [30]

Итак, анализ разделён на два направления: лингвистическо-морфологическое и географическо-тематическое. Такой

подход позволяет исследовать центры книгописания, что является одной из ключевых миссий Матенадарана, а также изучить особенности для распознавания текстов.

3.1 Лингвистическая специфика

3.1.1 Проработка гипотез о появлении новых символов и диалектах

Что касается первой гипотезы (потери в точности распознавания связаны с заменой букв), то она заключается в проверке несоответствия символов в связи с их появлением. Данная гипотеза является отсылкой к оценке качества распознавания. Здесь речь идёт о метрике CER (Character Error Rate), которая была рассчитана с помощью нахождения пересечения эталонного и распознанных текстов.

Исследование символов также приведено в книге “Апокрифы, псевдоэпиграфы и арменоведение: армянские рукописи, текстология и Святая Земля” [8], где учёные вручную исследовали особенности написания строчных и заглавных букв, а далее подошли к задаче творчески-автоматизировано и сгенерировали армянские буквы (рисунок 17).

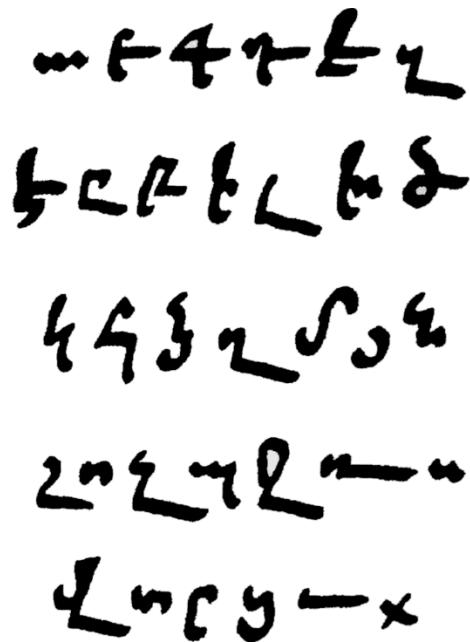


Рисунок 17 – Сгенерированные компьютером символы [8]

Армянский алфавит дошёл до наших дней почти в неизменном виде, однако было установлено, что некоторые символы в определённый рубеж времени стали употребляться реже. Таких найдено 2 и динамика их употребления ухудшается, а динамика употребления тех, на которые они были заменены, улучшается (рисунок 18), что подтверждает и саму гипотезу, и вывод, сделанный при анализе причин результатов оценки точности.

Система письменности такова, что один звук обозначается одним символом - исключением является звук "у", однако он заменён не был, поэтому употребление слова "буква" вместо "символа" уместно.

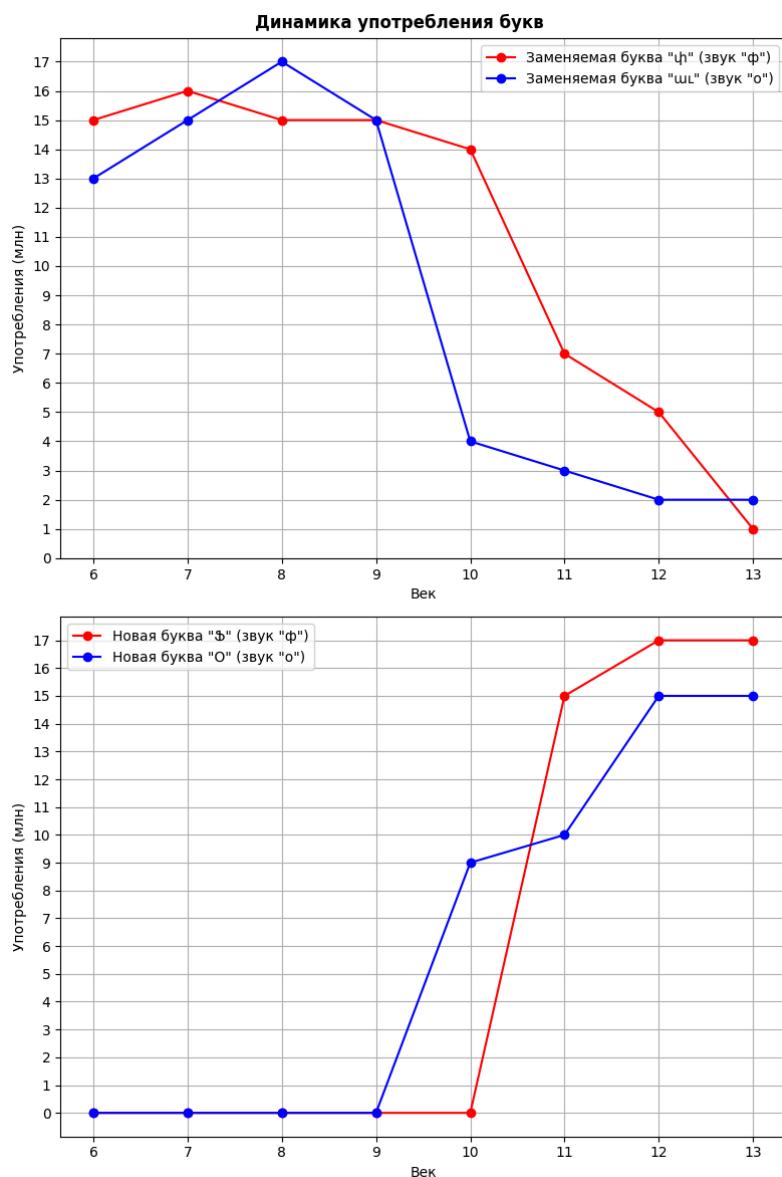


Рисунок 18 – Динамика употребления букв

Ещё одна гипотеза для символического анализа – это проверка того, что можно выделить разные диалекты, которые отличаются написанием. Гипотеза подтверждается: в результате получились две группы и их главное отличие – это оглушение и озвончение согласных для восточного и западного диалектов, соответственно. Особенно это показательно и понятно на именах собственных (транслитерация приведена в таблице 1).

Таблица 1 – Примеры оглушения и озвончения в восточном и западном диалектах

Восточный диалект	Западный диалект
Крикор	Григор
Сурп	Сурб
Гехард	Гаград
Эрсrum	Эрзрум
Артвин	Ардин
Ашот	Ашод

3.1.2 Проработка гипотезы о стратификации языковых конструкций

Третья гипотеза подтверждается: определённые виды языковых конструкций появляются в контексте определённого вида деятельности, то есть социально стратифицированы (рисунок 19): для неё был составлен корпус и проведена классификация с помощью метода опорных векторов.

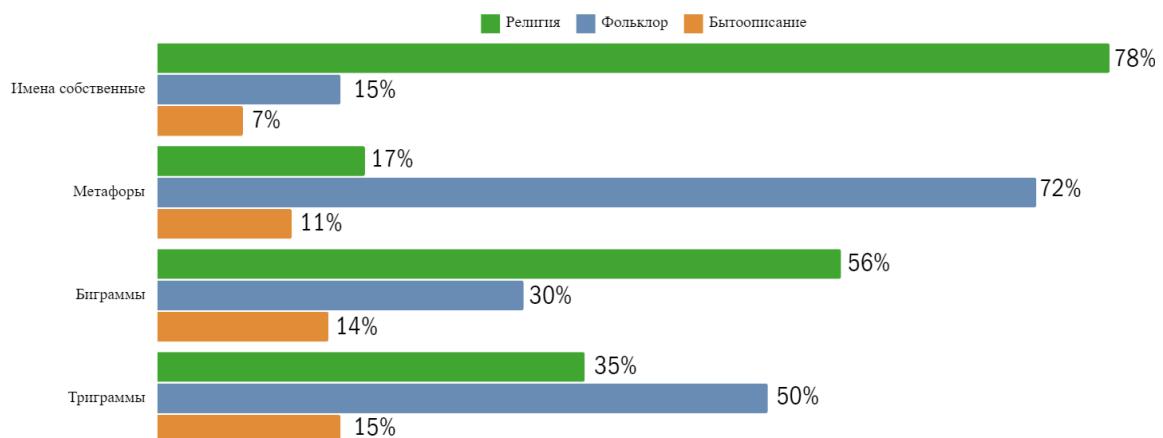


Рисунок 19 – Языковые конструкции и виды деятельности

Сделаны следующие выводы:

- имена собственные наиболее характерны для религиозной сферы, что объясняется спецификой литературы, связанной с религией,
- метафоры же более типичны для описание фольклорной деятельности. Это связано с тем, что приведены тексты традиционных песен, которые имеют художественную направленность,
- далее проведён анализ биграмм и триграмм для выявления выражений-паттернов. Для этого проведена токенизация, лемматизация, удаление стоп-слов.

Таким образом, употребление определённых конструкций речи социально стратифицировано.

3.2 Географическо-тематическая сегментация

Географическо-тематическая сегментация покрывает миссию Матенадарана по изучению центров книгописания, что необходимо для литературоведения и анализа самих рукописей и истории. Географическая сегментация уже применялась для анализа нематериального культурного наследия, например, карта костюма [31] (рисунок 20).



Рисунок 20 – Карта народного костюма [31]

Она обусловлена культурным и профессиональным разделением общества, внутри которого выбраны направления, позволяющие проводить социальную стратификацию (например, по видам деятельности) методом классификации. Как метод, географическо-тематическая сегментация позволяет рассматривать данные и разделять их на два направления одновременно.

Поставленные гипотезы:

- в регионах создания, которые не относятся к современным границам Армении, доминирует религиозная тематика,
- виды занятий были сугубо сельскохозяйственного или промышленного характера в соответствии с этапом исторических обществ,
- на протяжении исследуемого периода была широко представлена фольклорная тематика.

В качестве инструмента выбрано латентное преобразование Дирихле, реализованное с помощью python, позволяющее разделять текстовый корпус на темы и извлекать основные из них (рисунок 21).

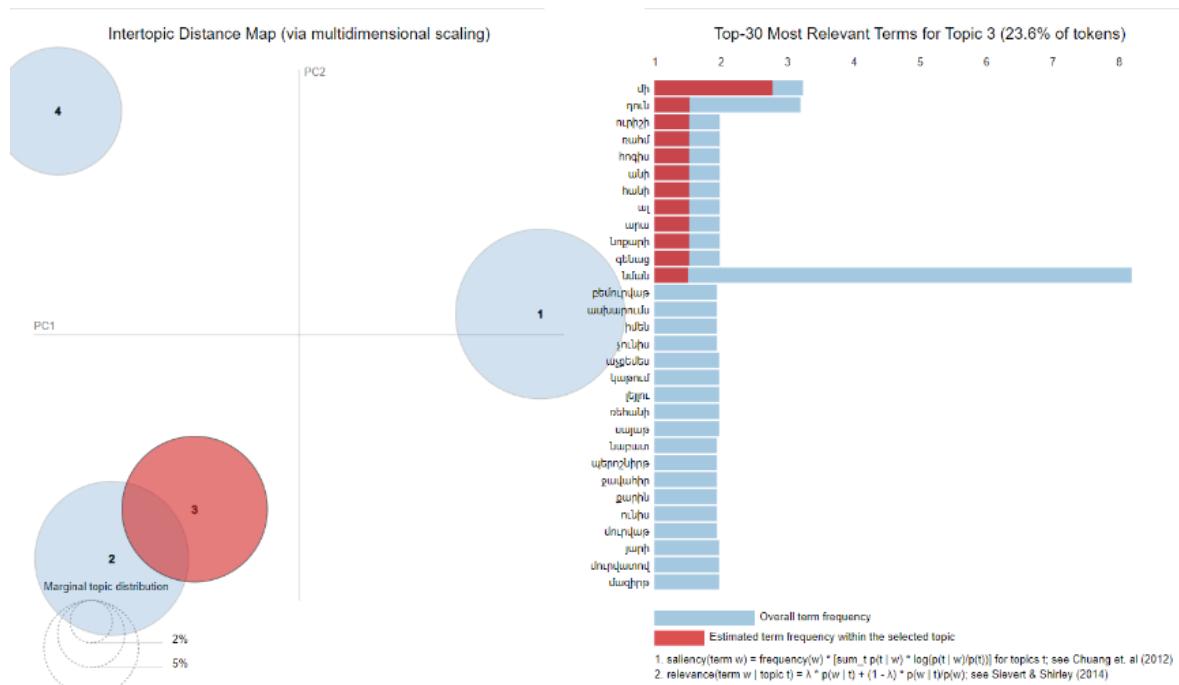


Рисунок 21 – Тематическая карта, созданная с помощью латентного преобразования Дирихле

Применив этот метод для имеющихся в данных регионов, можно моделировать карту с указанием наиболее типичного для определённого региона вида культурного наследия (рисунок 22). В качестве инструмента выбран инструмент Tableau, который позволяет проводить аналитику и визуализацию данных.

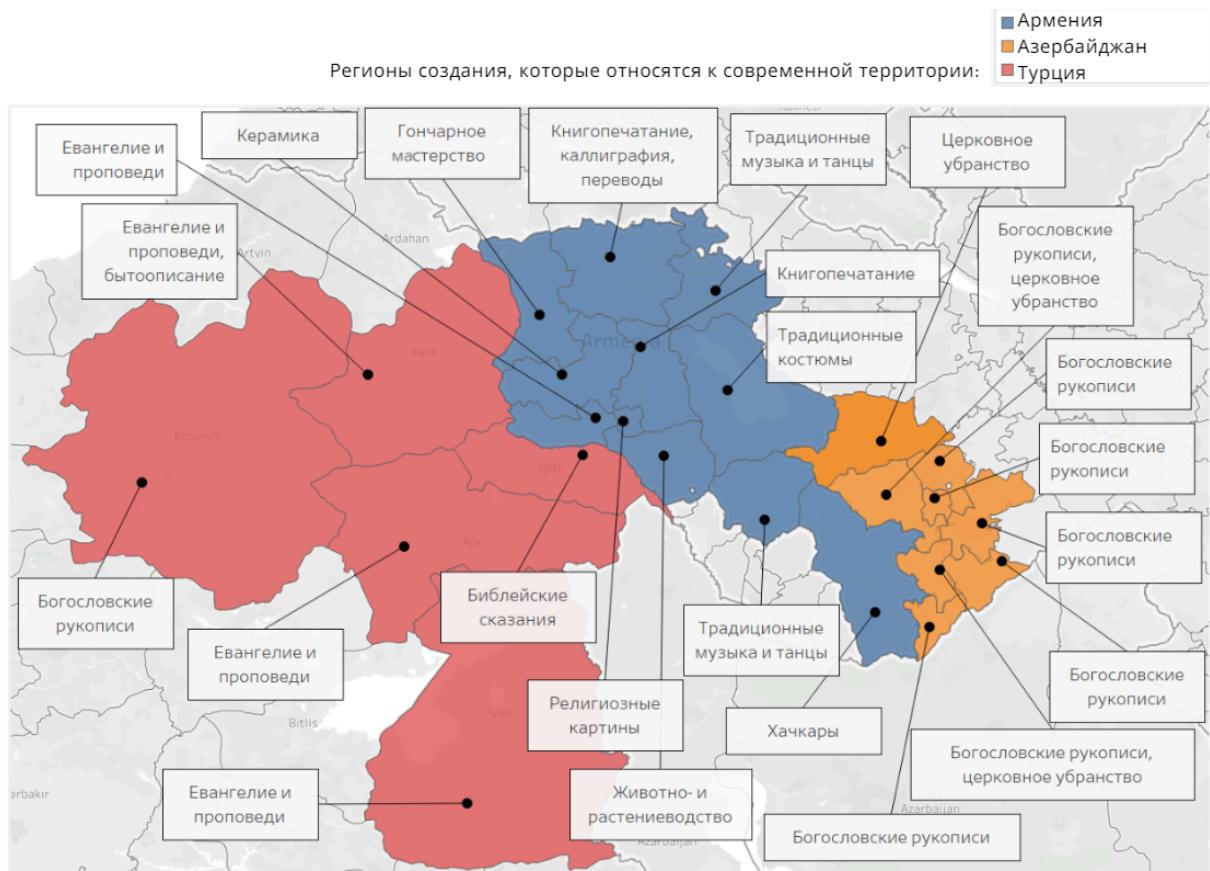


Рисунок 22 – Карта для проверки гипотез

Создание такой карты отвечает одной из миссий Матенадарана, а именно изучению центров книгописания.

Что касается поставленных гипотез, то:

– действительно, в регионах создания, которые не относятся к современным границам Армении, доминирует религиозная тематика, например, переводы священных книг, проповеди и Евангелие, богословские рукописи и убранство церквей, а также хачкаров – вид армянских архитектурных памятников и святынь, представляющий собой каменную стелу с резным изображением креста (рисунок 23),



Рисунок 23 – Мемориальный хачкар [32]

- действительно, на протяжении исследуемого периода была широко представлена фольклорная тематика, в том числе, традиционные костюмы, музыка и танцы,
- однако виды занятий не были сугубо сельскохозяйственного или промышленного характера, так как в бытоописании хоть и можно найти упоминания растениеводства и животноводства, но оно не доминирует в видах деятельности.

3.3 Итоги раздела 3

Раздел 3 посвящён анализу и визуализации текстов. В рамках раздела:

- изучены методы распознавания списка.
- применены методы классификации и кластеризация текстов, которые являются данными для визуализации гео-тематической и лингвистических направлений,
- проведена визуализация, позволяющая проверить поставленные гипотезы, а также оценить результаты проверки точности распознавания.

Данные результаты позволили провести содержательную интерпретацию и оценить качество полученных количественных результатов.

ЗАКЛЮЧЕНИЕ

В ходе выпускной квалификационной работы достигнута цель, то есть реализовано распознавание и моделирование текстовых данных армянских рукописей, и выполнены задачи, поставленные в начале исследования, а именно:

- изучена предметная область: около 30 источников, связанных как с цифровой, так и с культурно-историческим составляющей (в том числе не только теоретический материал, но и готовые смежные проекты). Соединение этих аспектов является необходимым в работах, связанных с цифровой гуманитаристикой,
- сбор данных: сформирован датасет благодаря применению эффективного инструмента по сбору данных, который снимает необходимость написания собственного парсера,
- распознавание текста: рассмотрены подходы и наиболее популярные библиотеки для обработки текстовых данных. Также проведено оценивание качества распознавания и посимвольная метрика распознавания с исследованием дополнительных данных, исследованных специалистами вручную,
- анализ и визуализация: проведён гео-тематический и лингвистический анализ с применением методов классификации и кластеризации. Предварительно поставлены гипотезы. Будучи инструментом представления результатов анализа, визуализация необходима для их структурирования, что позволяет дать ответы, в частности, на миссии хранилища рукописей.

Что касается содержательной интерпретации, то сделаны следующие выводы:

- в регионах создания, которые не относятся к современным границам Армении, доминирует религиозная тематика,

- виды занятий не были сугубо сельскохозяйственного или промышленного характера в соответствии с этапом развития исторических обществ,
- на протяжении исследуемого периода была широко представлена фольклорная тематика,
- определённые виды языковых конструкций появляются в контексте определённого вида деятельности, то есть социально стратифицированы,
- потери в точности распознавания связаны с заменой букв,
- можно выделить разные диалекты, которые отличаются написанием (озвончением или оглушением согласных).

Применённые инструменты:

- WebScraper [18], позволяющий автоматизировать процесс сбора данных с веб-страниц,
- Tesseract OCR для распознавания текста,
- python для предобработки данных и построения графиков,
- tableau для моделирования и карт.

В результате, применив интеллектуальные технологии работы с данными, удалось определить историко-культурную природу исследуемых явлений, выявить основные виды культурного наследия по регионам и выявить некоторые языковые особенности, что особенно важно с учётом региона и его специфики, заключающейся в древности истории.

Работа представляет интерес, так как развивает проекты по оцифровке рукописей, которые пока ещё не полностью отсканированы. К тому же такой подход может быть полезным и для уже готовых проектов, которые, в основном, представляют из себя скан страницы из рукописи и расшифрованный текст рядом. Но именно анализ и визуализация (особенно текстовых) данных позволяет манускриптам “не пылиться” пусть даже на электронных полках. Также задачи позволяют сохранять память о культурном наследии, что благоприятно влияет на межкультурную коммуникацию.

Работа была апробирована на 3 научных конференциях – “XIII Конгрессе Молодых Учёных” [33] (опубликован тезис [34]), конференции “Гуманитарные проблемы актуальных наук: цифровая дисциплина и проект” DH-центра Университета ИТМО [35] и “XXII Международной конференции молодых ученых «Векторы»” Московской Высшей Школы Социальных и Гуманитарных наук [36], где был задан вопрос “Как это исследование улучшит жизнь людей в целом и армян (как представителей выбранного региона), в частности?”. Данный вопрос обуславливает актуальность количественных исследований и стратификации культурных атрибутов, что позволяет ставить некоторые гипотезы, проверка которых помогает сохранить историческую память, которая, в свою очередь, является одним из двигателей непрерывного развития цивилизаций, ведь накопленный опыт всегда является фундаментом чего-то нового.

Таким образом, в рамках данной ВКР была установлена специфика географической и культурной сегментации в результате работы над проблематикой культурной стратификации на основе древних рукописей по двум направлениям раскрыта полностью в соответствии с поставленными задачами и гипотезами.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Armenian Manuscripts and Printed Books of Digital Bodleian – URL: <https://digital.bodleian.ox.ac.uk/collections/armenian/> (дата обращения 16.03.2024).
2. Пилипосян А.С., Геворгян Л.П. Изучение и проект развития археологического комплекса – музея – заповедника «Мецамор» // Наследие и современность. – 2019. – N 3. – С.74 – 91.
3. World Heritage Sites – Armenia – URL: https://www.unesco.org/en/world-heritage/grid?f%5B0%5D=countries%3A03c705d8-880e-55bf-b4b5-c023086a7e75&f%5B1%5D=dataset_filters%3A6b092825-9770-47e0-92c6-084caebeca46#toggle-facets (дата обращения 23.04.2024).
4. Объекты Всемирного наследия ЮНЕСКО в Армении – URL: <https://www.advantour.com/rus/armenia/unesco-world-heritage-sites.htm> (дата обращения 23.04.2024).
5. Armenia, Elements on the Lists of Intangible Cultural Heritage – URL: <https://ich.unesco.org/en/state/armenia-AM?info=elements-on-the-lists> (дата обращения 23.04.2024).
6. UD for Armenian (Eastern) – URL: <https://universaldependencies.org/hy/index.html> (дата обращения 23.04.2024).
7. Comparing the effectiveness of hyperspectral imaging and Raman Maybury I., Howell D., Terras M., Viles H.A. Comparing the effectiveness of hyperspectral imaging and Raman spectroscopy: a case study on Armenian manuscripts // Heritage Science – N 6 – 2018 – P. 42-57.
8. Stone M. E. Apocrypha, Pseudepigrapha, and Armenian Studies: Armenian manuscripts, textual studies, and Holy Land // Peeters – 2006 – P. 489-492.

9. A Modular and Automated Annotation Platform for Handwritings: Evaluation on Under-Resourced Languages – URL:
https://link.springer.com/chapter/10.1007/978-3-030-86334-0_33 (дата обращения 23.04.2024).
10. Post-OCR Correction of Armenian Texts Using Neural Networks – URL:
https://www.researchgate.net/publication/354067467_Post-OCR_Correction_of_Armenian_Texts_Using_Neural_Networks (дата обращения 23.04.2024).
11. From Manuscript to Tagged Corpora An Automated Process for Ancient Armenian or Other Under-Resourced Languages of the Christian East – URL:
https://edizionicafoscarini.unive.it/media/pdf/article/armeniaca/2022/1/art-10.30687-arm-2974-6051-2022-01-005_lMiAeVT.pdf (дата обращения 23.04.2024).
12. Генеральный каталог рукописей Матенадарана – URL:
<https://matenadaran.am/ru/> (дата обращения 23.04.2024).
13. Digital analyses and critical editions of textual and iconographic corpora – URL:
https://iscd.sorbonne-universite.fr/research/incubated-teams/digital-humanities/textual_iconographic_corpora/ (дата обращения 23.04.2024).
14. Edition numérique collaborative et critique de l'encyclopédie – URL:
<http://enccre.academie-sciences.fr/encyclopedie/article/v8-1274-2/> (дата обращения 23.04.2024).
15. An edition of Petrarch's songbook – URL:
<https://dcl.luddy.indiana.edu/petrarchive/> (дата обращения 23.04.2024).
16. Petrarch's Rerum vulgarium fragmenta – URL:
<https://dcl.luddy.indiana.edu/petrarchive/content/c019r.xml?facsl=active> (дата обращения 23.04.2024).

- 17.Фомина Н.Н., Джалиашвили З.О., Борисов О.С., Свечникова Н. О., Толстикова И. И., Филичева Н. В., Деньгинова И. Г. Дидактическое построение культурологического знания в контексте информационных технологий // Научно-технический вестник информационных технологий, механики и оптики. – 2003. – N 8. – C.160-173.
- 18.Для чего аналитику данных датасет и где его взять – URL:<https://practicum.yandex.ru/blog/dataset-dlya-mashinnogo-obucheniya-i-analiza/> (дата обращения 23.04.2024).
- 19.Creusier J. Le rôle de la collecte de données dans la qualité des productions scientifiques – URL:
https://www.researchgate.net/publication/320627366_Le_role_de_la_collecte_de_donnees_dans_la_qualite_des_productions_scientifiques (дата обращения 23.04.2024).
- 20.Котельников А.С., Якунина В.Н., Рысина А.Д., Красникова С.А., Атовмян И.О., Шувалов В.Б. Сбор и обработка исторических данных в автоматизированных информационных системах // Прикладная информатика. – 2012. – N 6 (42). – C.9– 14.
- 21.Палкина С.А., Ухлова В.В. Модификация алгоритма CRISP для визуализации лабораторных данных на платформах класса BI // IN SITU. – 2023. – N 2. – C.11-14.
- 22.CRISP-DM Help Overview – URL:
<https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview> (дата обращения 24.04.2024).
- 23.Digital Bodleian Collections – URL:
<https://digital.bodleian.ox.ac.uk/browse/#collections> (дата обращения 24.04.2024).
- 24.Powerful web scraper for regular and professional use – URL:
<https://webscraper.io/> (дата обращения 24.04.2024).

- 25.Бобров К.А., Шульман В.Д., Власов К.П. Анализ технологий распознавания текста из изображения // Международный журнал гуманитарных и естественных наук. – 2022. – N 3-2. – С.124 – 128.
- 26.Марков А.В. Проведение сравнительного анализа Attention OCR и Tesseract в задаче распознавания символов на изображениях прейскурантов // Евразийский Союз Ученых. – 2020. – N 5-3 (74). – С.65 – 67.
- 27.Золотарев О.В., Юрчак В.А. Инструменты решения проблем распознавания и кластеризации данных из документов методами машинного обучения // ИВД. – 2023. – N 2 (98). – С.156 – 164.
- 28.Karatzas D., Shafait F., Uchida S., Iwamura M., Gomez i Bigorda L., Robles Mestre S., Mas J., Fernandez Mota D., Almazan Almazan J., Pere de las Heras L., ICDAR 2013 Robust Reading Competition – URL: <http://refbase.cvc.uab.es/files/KSU2013.pdf> (дата обращения 05.05.2024).
- 29.Karatzas D., Gomez-Bigorda L., Nicolaou A., Ghosh S., Bagdanov A., Iwamura M., Matas J., Neumann L., Chandrasekhar V. R., Lu S., Shafait F., Uchida S., Valveny E., ICDAR 2015 competition on Robust Reading – URL: <https://ieeexplore.ieee.org/document/7333942/authors#authors> (дата обращения 06.05.2024).
- 30.Генеральный каталог рукописей Матенадарана – URL: <https://matenadaran.am/ru/> (дата обращения 23.04.2024)
- 31.Патрик А. Карта армянского национального костюма. XIX – первая четверть XX века – URL: <https://www.armmuseum.ru/clothes> (дата обращения 26.09.2023).
- 32.Murmansk, Russia, Armenian Genocide Memorial Khachkar, – URL: https://www.armenian-genocide.org/Memorial.325/current_category.269/memorials_detail.html#memorial_image (дата обращения 21.09.2023).
- 33.XIII Конгресс молодых учёных – URL: <https://kmu.itmo.ru/> (дата обращения 10.05.2024).

- 34.Петросян А.М., (науч. рук. Коцюба И.Ю., Пригодич Н.Д.), Анализ данных культурной сферы и объектов культурного наследия Армении – URL: <https://kmu.itmo.ru/digests/article/12592> (дата обращения 10.05.2024).
- 35.Конференция “Гуманитарные проблемы актуальных наук: цифровая дисциплина и проект”, – URL: <https://dh.itmo.ru/april-conference-2024> (дата обращения 10.05.2024).
- 36.Программа XXII Международной научной конференции «Векторы», – URL: <https://vectors2024.tilda.ws/program> (дата обращения 10.05.2024).