МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

# КЛАСТЕРИЗАЦИЯ ПЕРВИЧНЫХ ТЕКСТОВЫХ ОПИСАНИЙ ТРАВМАТИЧЕСКИХ ПОВРЕЖДЕНИЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Чеснокова Павла Андреевича

Научный руководитель:
профессор кафедры
информатики и
программирования, д.т.н _____ Фалькович А. С.
подпись, дата

Зав. кафедрой:
к.ф.- м.н, доцент _____ Огнева М. В.
подпись, дата

Саратов 2025

# ВВЕДЕНИЕ

**Актуальность темы.** Задача обработки текстов медицинской направленности имеет большое значение для здравоохранения. Особую ценность представляет возможность получить новую информацию из результатов первичной диагностики, так как неполноценная догоспитальная травматологическая помощь является существенной причиной предотвратимых смертей [1]. Предполагается, что результаты исследования возможно применить в экстремальных условиях, являющихся предметом исследования теории катастроф. Актуальность работы подтверждается наличием большого числа новых исследований в данной области за последние 5 лет.

**Цель бакалаврской работы** – Провести кластеризацию развёрнутых описаний диагнозов травматических повреждений.

Поставленная цель определила **следующие задачи**:

1. Рассмотреть набор исследовательских работ из области обработки естественных языков в направлении биомедицинских и клинических текстов.
2. Определить подходы, используемые при обработке текстовых данных.
3. Определить особенности обработки клинических текстов.
4. Рассмотреть теоретические основы алгоритмов кластеризации.
5. Провести предварительную обработку данных.
6. Провести кластеризацию на тестовой и валидационной выборках.
7. Провести анализ полученных результатов и сравнить с результатами исследований специалистов-травматологов.

**Методологические основы** кластеризации первичных текстовых описаний травматических повреждений представлены в работах: Beilman G, Finzel R., Gipson J.C., Knoll B.C., Lindemann E.A., Lyng J.W., McEwan R.,

Melton G.B., Silverman G.M., Pakhomov S., Tignanelli C.J., Trembley A.L [1]. B. Clay, H.I. Bergman, S. Salim, G. Pergola, J. Shalhoub, A. H. Davies [2]. Radu R.G. [3]. Прошиной М.В. [4]. Сердюк Ю. П., Власовой Н. А. [5].

**Теоретическая и/или практическая значимость бакалаврской работы.** Практически полученные параметры обработки и кластеризации предполагается возможным использовать в системах, относящихся к медицине катастроф. Полученные кластеры диагнозов могут быть проанализированы экспертами из области медицины.

**Структура и объём работы**. Бакалаврская работа состоит из введения, четырёх разделов, заключения, списка использованных источников и одиннадцати приложений. Общий объем работы – 79 страниц, из них 46 страниц – основное содержание, включая 10 рисунков и 17 таблиц, список использованных источников информации – 25 наименований.

## КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

**Первый раздел «Обзор источников»** посвящен рассмотрению существующих работ, относящихся к обработке естественных языков в области медицины. Приведено описание тринадцати работ, опубликованных в последние 5 лет, непосредственно посвящённых решению некоторых из задач обработки естественных языков, в частности из области медицины.

Таким образом, в первом разделе приведён анализ существующих исследований. Были сделаны выводы об особенностях предметной области, с которыми работали в предшествующих работах.

**Второй раздел «Математическая модель текста»** посвящен теоретическим основам предварительной векторизации текстов. Рассмотрена теория статистических и нейросетевых методов моделирования текста. Сделан вывод об актуальности применения статистических методов векторизации текстов.

**Третий раздел «Существующие алгоритмы кластеризации»** посвящен теории алгоритмов кластеризации. Определена общая цель применения кластеризации. В разделе были рассмотрены: алгоритм плоской кластеризации $k$ средних, внутренние и внешние меры оценки качества кластеризации, а также методы понижения размерности, необходимые при создании визуальных представлений объектов кластеров.

**Четвёртый раздел «Полученная реализация и результаты экспериментов»** посвящен практической реализации кластеризации первичных текстовых описаний травматических повреждений. Полученный в работе алгоритм состоит из четырёх шагов:

1. Подготовки данных.
2. Построения векторной модели и анализа свойств текста.
3. Кластеризация данных.

4. Описания проведённых экспериментов и оценки качества кластеризации.

В задачах машинного обучения эффективность применения алгоритмов зависит от качества проведённой предварительной работы с данными. Сложностью подготовки данных на естественных языках является отсутствие единственно верного подхода к предварительной обработке текста. Существует общий рекомендованный набор действий с исходным текстом: приведение к единому регистру, удаление «стоп-слов», пунктуации, лемматизация и прочие. Однако, необходимая последовательность действий по подготовке данных зависит от особенностей исходных текстов, выясняемых на этапе анализа набора данных и корректируется на основании эмпирических наблюдений.

Поэтому четвёртый раздел работы посвящен подходам, использовавшимся для понижения размерности признакового описания без потери существенных особенностей исходных данных. Так, в результате реализованных методов предварительной обработки удалось сократить размерность признакового описания с 2625 слов в исходных текстах до 1526 слов в обработанных текстах. Данный результат был получен с применением следующих шагов:

1. Удаления дат из текстов диагнозов.
2. Удаления анатомических обозначений позвонков.
3. Удаления цифр и чисел.
4. Удаления знаков пунктуации, лишних пробельных символов, знаков табуляции и перевода строки.
5. Приведения встречающиеся в текстах синонимов к нормальной форме.
6. Удаления «стоп-слов» – слов, имеющих высокую частоту употребления в тексте, но при этом не имеющих большой семантической значимости.

Для данных шагов предварительной обработки были реализованы два программных модуля. Также опытным путём было установлено, что лемматизация, либо стемминг снижали качество последующей кластеризации.

Далее, в разделе четыре, приведено описание действий, выполнявшихся для построения векторной модели текстов диагнозов. В работе описан анализ качества векторизации на основе изображений, полученных при помощи реализованных функций визуализации. Приведён результат анализа корреляции слов в векторах, а также описан опыт применения векторизации триграмм.

Следуя приведённому в начале четвёртого раздела алгоритму работы с набором данных, после описания векторизации приведён результат подбора оптимального числа кластеров алгоритма кластеризации с помощью реализованной функции визуализации. В остальной части четвёртого раздела работы приведено описание проведённых экспериментов. Описаны параметры экспериментов, также приведены результаты анализа полученных результатов. В данной части работы приведены результаты трёх экспериментов с различными параметрами векторизации и кластеризации. При векторизации, основными, корректирующимися параметрами были: ограничение максимального числа признаков и параметр размерности N-грамм. В таблице 1 приведены параметры векторизации, использовавшиеся в каждом из трёх проведённых экспериментов.

Таблица 1 — Параметры векторизации в трёх экспериментах

| Параметр | Значение в эксперименте №1 | Значение в эксперименте №2 | Значение в эксперименте №3 |
|---|---|---|---|
| Analyzer | word | word | word |
| max_df | 0,95 | 0,95 | 0,95 |

Продолжение таблицы 1

| | | | |
|---|---|---|---|
| min_df | 1 | 1 | 1 |
| max_features | 200 | 20 | 25 |
| ngram_range | 1,1 | 1,1 | 3,3 |

После проведённой кластеризации результаты оценивались на основе мер точности и полноты. Наилучший результат был получен в третьем эксперименте при кластеризации с числом кластеров k = 4 и применением триграмм с векторами размерности 25. Лучшие полученные с данными параметрами результаты – точность 0.425 и полнота 0,818.

В приложениях к основной части был приведён код вошедших в работу функций и классов, также приведены технические характеристики компьютера, использовавшегося для проведения экспериментов.

Полученный алгоритм был испытан с применением тестовой и валидационной выборок. Исходные данные были разделены в пропорции 1 / 3. Результаты показали, что значение полноты не уменьшилось, следовательно, при разбиении на 4 кластера и параметрах векторизации, подобранных в третьем эксперименте, действительно выделяются кластеры диагнозов с наибольшей летальностью. В таблице 2 приведены результаты испытания полученного алгоритма кластеризации с параметрами третьего эксперимента.

Таблица 2 — Результаты точности (precision) и полноты (recall) полученные на тестовой и валидационной выборках

| | Тестовая выборка | Валидационная выборка |
|---|---|---|
| Precision | 0,492 | 0,357 |
| Recall | 0,838 | 0,833 |

Таким образом, в работе был реализован алгоритм кластеризации первичных описаний текстов травматических повреждений. В результате кластеризации были получены значения точности и полноты, схожие со значениями прогнозов точности и полноты, рассчитанными для прогнозов врачей-исследователей. Тестирование показало, что проведённая кластеризация позволяет выделять кластеры диагнозов с наибольшей летальностью на основе статистических характеристик текстов диагнозов.

# ЗАКЛЮЧЕНИЕ

Рассмотренные в работе данные ранее были использованы для построения прогноза летальности врачами-исследователями с помощью расчёта баллов, присваиваемых каждому виду повреждений каждого органа или групп органов. При этом точность (precision) их прогноза составила 0.477, а полнота (recall) 0.895. В данной работе была проанализирована возможность построения аналогичного прогноза с помощью кластеризации текстовых описаний. При этом был получен сопоставимый по качеству результат — точность 0.425 и полнота 0.818.

Для кластеризации развёрнутых описаний диагнозов травматических повреждений были решены следующие задачи: рассмотрены исследовательские работы в области обработки текстов в направлении биомедицинских и клинических текстов на естественных языках, определены подходы, используемые при обработке текстовых данных, определены особенности обработки клинических текстов, изучены теоретические основы алгоритмов кластеризации, проведена предварительная обработка данных, проведена кластеризация на тестовой и валидационной выборках, выполнен анализ полученных результатов.

Таким образом, поставленная цель выпускной квалификационной работы была достигнута.

## ОСНОВНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ


1 Beilman G, Finzel R., Gipson J.C., , Knoll B.C., Lindemann E.A., Lyng J.W., McEwan R., Melton G.B., Silverman G.M., Pakhomov S., Tignanelli C.J., Trembley A.L. Natural language processing of prehospital emergency medical services trauma records allows for automated characterization of treatment appropriateness. // J Trauma Acute Care Surg. 2020 May;88(5):607-614. doi: 10.1097/TA.0000000000002598.

2 Clay B. Natural language processing techniques applied to the electronic health record in clinical research and practice - an introduction to methodologies / B. Clay, H. I. Bergman, S. Salim, G. Pergola, J. Shalhoub, A. H. Davies // Computers in Biology and Medicine. 2025. V. 188. 109808. 15 p. [Электронный ресурс]. URL: https://www.sciencedirect.com/science/article/pii/S0010482525001581

3 Radu R.-G. Clustering Documents using the Document to Vector Model for Dimensionality Reduction / R.-G. Radu, I.-M. Radulescu, C.-O. Truica, E.-S. Apostol, M. Mocanu // Conference: 2020 IEEE Intern. Conf. on Automation? Quality and Testing, Robotics (AQTR). 7 p. [Электронный ресурс] URL: https://www.researchgate.net/publication/342613401_Clustering_Documents_using_the_Document_to_Vector_Model_for_Dimensionality_Reduction/

4 Прошина М. В. Разработка рекомендательной системы научных публикаций в области медицины на основе методов машинного обучения. // Научно-аналитический журнал Инновации и инвестиции №5. 2022 — [С. 142-148]

5 Сердюк Ю. П., Власова Н. А., Момот С. Р. Система извлечения упоминаний симптомов из текстов на естественном языке с помощью нейронных сетей // Программные системы: теория и приложения. 2023. Т. 14. № 1(56). С. 95–123. [Электронный ресурс] URL: https://psta.psiras.ru/read/psta2023_1_95- 123.pdf