Министерство науки и высшего образования РФ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра математической теории
упругости и биомеханики

Создание рекомендательного алгоритма для музыкальных стриминговых сервисов

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студента 4 курса 442 группы

направления 09.03.03 – Прикладная информатика

механико-математического факультета

Арсениевича Данилы Милутиновича

Научный руководитель
к.ю.н., доцент _____ Р.В. Амелин
подпись, дата

Зав. кафедрой
д.ф.-м.н., профессор _____ Л.Ю. Коссович
подпись, дата

Саратов 2023

Введение. Музыкальные стриминговые сервисы, такие как Spotify, Apple Music, Deezer, Яндекс Музыка и другие, стали неотъемлемой частью жизни миллионов пользователей по всему миру. Они предлагают доступ к огромному количеству музыкальных композиций из разных уголков планеты, полностью изменяя представление о музыкальном пространстве и взаимодействии с музыкой. Однако, с ростом количества доступного контента, возникла проблема выбора подходящих треков, отвечающих вкусам и настроению каждого пользователя.

Для решения этой проблемы стриминговые сервисы используют системы рекомендаций, которые применяют алгоритмы машинного обучения и анализа данных для предсказания музыкальных предпочтений пользователей на основе их истории прослушивания. Эти системы стали не только важным компонентом успешных платформ, но и поспособствовали увеличению вовлеченности пользователей и удержанию их на платформе через предоставление персонализированного контента.

Цель и задачи. Целью данной работы является разработка рекомендательного алгоритма для музыкальных стриминговых сервисов, который будет учитывать неявные признаки, проявляющиеся через звуковые характеристики песен, для предоставления точных и персонализированных рекомендаций. Для достижения этой цели были поставлены следующие задачи:

1. Исследование и анализ существующих алгоритмов и подходов в рекомендательных системах для музыкальных стриминговых сервисов.
2. Определение требований к разрабатываемому рекомендательному алгоритму, включая ключевые метрики эффективности и учет пользовательских предпочтений и аудио характеристик песен.
3. Проектирование структуры и основных компонентов рекомендательного алгоритма, выбор моделей и методов машинного обучения, определение способов обработки и анализа данных.

4. Подготовка и предобработка данных для обучения и тестирования алгоритма, включая сбор, очистку, анализ и предобработку метаданных о песнях и их аудио характеристик.
5. Реализация и тестирование разработанного рекомендательного алгоритма, обучение и тестирование на подготовленных данных, анализ результатов.

Данная работа позволит провести исчерпывающий анализ применения алгоритма K-средних в контексте рекомендательной системы и определить его эффективность при выборе оптимального числа кластеров.

Структура и объем работы. Выпускная квалификационная работа состоит из введения, 3 глав, заключения, списка использованных источников, включающего 20 наименований, работа изложена на 43 листах машинописного текста, содержит 8 рисунков и код. Далее приведены наименования глав:

Структура работы:

1. Анализ методологии и практики рекомендательных систем в музыкальных стриминговых сервисах
2. Реализация
3. Предложения по дальнейшим улучшениям

Во введении дипломной работы акцентируется внимание на актуальности исследования. В результате были сформулированы цель и задачи.

В первой главе производится анализ методологии и практики использования рекомендательных систем в сервисах потокового воспроизведения музыки. Глава включает 5 разделов:

1. Обзор музыкальных стриминговых сервисов: концепция и ключевые характеристики. В данном разделе рассматривается концепция музыкальных стриминговых сервисов и их важные характеристики, такие как доступность, масштабируемость и персонализация. Описывается, что эти сервисы позволяют пользователям слушать музыку через интернет в любом месте и в любое время, предлагая широкий выбор треков и артистов. Также отмечается возможность персонализации, позволяющая сервисам адаптировать предложение под индивидуальные предпочтения пользователей. Обсуждаются также вызовы, с которыми сталкиваются музыкальные стриминговые сервисы, включая управление правами на контент, обеспечение качества звука и предоставление эффективных рекомендаций для удержания внимания слушателей и продвижения новой музыки.

2. Анализ методов и алгоритмов рекомендательных систем: преимущества и ограничения. Раздел посвящен исследованию различных методов и алгоритмов, применяемых в рекомендательных системах. В данном разделе осуществляется анализ нескольких подходов, включая коллаборативную фильтрацию, контент-ориентированный подход и гибридные системы.

Для каждого подхода проводится обсуждение его преимуществ, ограничений и особенностей. В частности, рассматривается использование методов, основанных на памяти и моделировании в коллаборативной фильтрации. Это включает анализ преимуществ и ограничений подходов, основанных на пользователе и элементе, а также методов вычисления сходства между пользователями или элементами.

Также в разделе обсуждается использование метаданных и аудио характеристик в контент-ориентированном подходе. Здесь рассматривается анализ музыкальных характеристик, таких как метаданные и аудио свойства, и их влияние на формирование рекомендаций. Обсуждаются преимущества использования музыкальных характеристик в контент-ориентированном подходе и их способность предлагать рекомендации на основе сходства музыкальных свойств.

Наконец, в разделе рассматривается преимущество гибридных систем, которые комбинируют различные подходы для улучшения рекомендаций. Гибридные системы объединяют преимущества и компенсируют ограничения различных методов и алгоритмов, что позволяет достигать более точных и персонализированных рекомендаций.

3. Использование моделей машинного обучения в рекомендательных алгоритмах. В данном разделе идёт речь об исследовании применения различных моделей машинного обучения в контексте систем рекомендаций. В данном разделе проводится анализ нескольких моделей, включая K-средних, матричную факторизацию, метод ближайших соседей, нейронные сети и градиентный бустинг.

Основная цель анализа заключается в выявлении преимуществ и ограничений каждой модели, а также определении их применимости в контексте рекомендательных систем. Каждая модель имеет свои особенности и подходы к обработке данных, что может оказывать влияние на точность и эффективность рекомендаций.

Преимущества и ограничения моделей машинного обучения в рекомендательных алгоритмах могут включать следующие аспекты: способность модели учесть скрытые факторы и предсказывать предпочтения пользователей, возможность работы с большими объемами данных, устойчивость к разреженности данных, сложность и время обучения моделей, адаптируемость к изменяющимся предпочтениям пользователей и новым элементам.

Анализ этих моделей и их применимости в рекомендательных системах позволяет лучше понять и выбрать наиболее подходящую модель для конкретного контекста и требований системы рекомендаций. Комбинирование различных моделей и подходов может привести к созданию гибридных систем, которые объединяют преимущества нескольких моделей и компенсируют их ограничения для достижения более точных и персонализированных рекомендаций.

4. Применение рекомендательных алгоритмов в музыкальных стриминговых сервисах. В данном разделе был проведен анализ применения рекомендательных алгоритмов в различных музыкальных стриминговых сервисах. Были рассмотрены такие глобальные платформы, как Spotify, Apple Music, Pandora, Deezer, SoundCloud и Яндекс.Музыка.

В результате анализа были выявлены различные подходы и методы, используемые в этих сервисах. Spotify и Apple Music оба применяют гибридные системы, комбинируя различные алгоритмы и источники информации для формирования персонализированных рекомендаций. Pandora использует уникальный контент-ориентированный подход с помощью Music Genome Project, анализируя сотни атрибутов каждой песни.

Другие платформы, такие как Deezer, SoundCloud и Яндекс.Музыка, активно применяют методы коллаборативной фильтрации. Например, Deezer представила функцию Flow, которая использует коллаборативную фильтрацию для предсказания следующих треков, которые будут интересны пользователю.

Анализ применения рекомендательных алгоритмов в этих сервисах позволяет выделить преимущества и ограничения каждого подхода. Гибридные системы могут предоставлять более точные рекомендации, комбинируя различные методы. Однако контент-ориентированный подход, хотя и обладает потенциалом, может быть более затратным в реализации из-за необходимости анализа множества атрибутов каждой песни.

Таким образом, проведенный анализ показывает, что музыкальные стриминговые сервисы используют разнообразные алгоритмы и подходы для формирования рекомендаций, с учетом предпочтений пользователей и анализа музыкального контента. Это позволяет улучшить пользовательский опыт и предложить более релевантную музыку каждому отдельному пользователю.

5. Обоснование необходимости создания рекомендательного алгоритма. В данном разделе проводится обоснование необходимости создания рекомендательного алгоритма для музыкальных стриминговых сервисов. Рассматривается значимость рекомендаций для улучшения музыкального опыта пользователей и эффективности индустрии. Сравниваются различные музыкальные платформы, такие как Spotify, Apple Music, Pandora и другие, их подходы к рекомендациям и использование различных алгоритмов. Особое внимание уделяется контексту России и необходимости разработки алгоритма, учитывающего отечественную музыкальную продукцию. Также обсуждаются преимущества и ограничения алгоритма K-средних в контексте музыкальной рекомендательной системы и его способность формировать кластеры схожих композиций на основе аудио характеристик треков. В целом, рассматривается значение рекомендательных алгоритмов и их роль в улучшении музыкального опыта пользователей.

Вторая глава посвящена созданию рекомендательного алгоритма и включает 6 разделов:

1. Описание набора данных для рекомендательного алгоритма. В данном разделе описывается использование набора данных "Million Song Dataset" в качестве основы для создания рекомендательного алгоритма. Описывается формат данных, включающий метаинформацию и аудио характеристики треков. Проводится выбор подвыборки из около 2000 песен для дальнейшего анализа. Обсуждается формат H5 для представления аудио характеристик и использование библиотеки pytables для работы с данными в формате H5. Указывается решение сосредоточиться на аудио характеристиках и исключить метаинформацию. Описывается преобразование характеристик

сегментов песни в их средние значения для сокращения сложности вычислений. Упоминается сохранение модифицированного набора данных в формате CSV и описываются некоторые столбцы, такие как идентификатор песни, название, продолжительность и год выпуска. Также отмечается наличие атрибута time_signature, который указывает на количество ударов в такте. После анализа столбцов в датасете проводится оценка значимости функций и используются различные техники визуализации данных.

2. Визуализация данных для набора данных Million Song. В данном разделе проводится анализ распределения данных и визуализация связей между признаками. График гистограммы демонстрирует распределение значений для каждого признака (рисунок 1). Некоторые характеристики, такие как продолжительность, год, танцевальность и энергия, имеют постоянные значения для всех песен, и решено исключить их из модели. Затем используется тепловая карта для визуализации корреляций между признаками (рисунок 2). Светлые области на карте указывают на высокую корреляцию, а темные области - на низкую или отсутствие корреляции. Анализ данных подтверждает наличие ожидаемых линейных связей между большинством характеристик, а темные области указывают на независимость некоторых признаков. После завершения анализа данных проводится предобработка данных перед их использованием в модели.

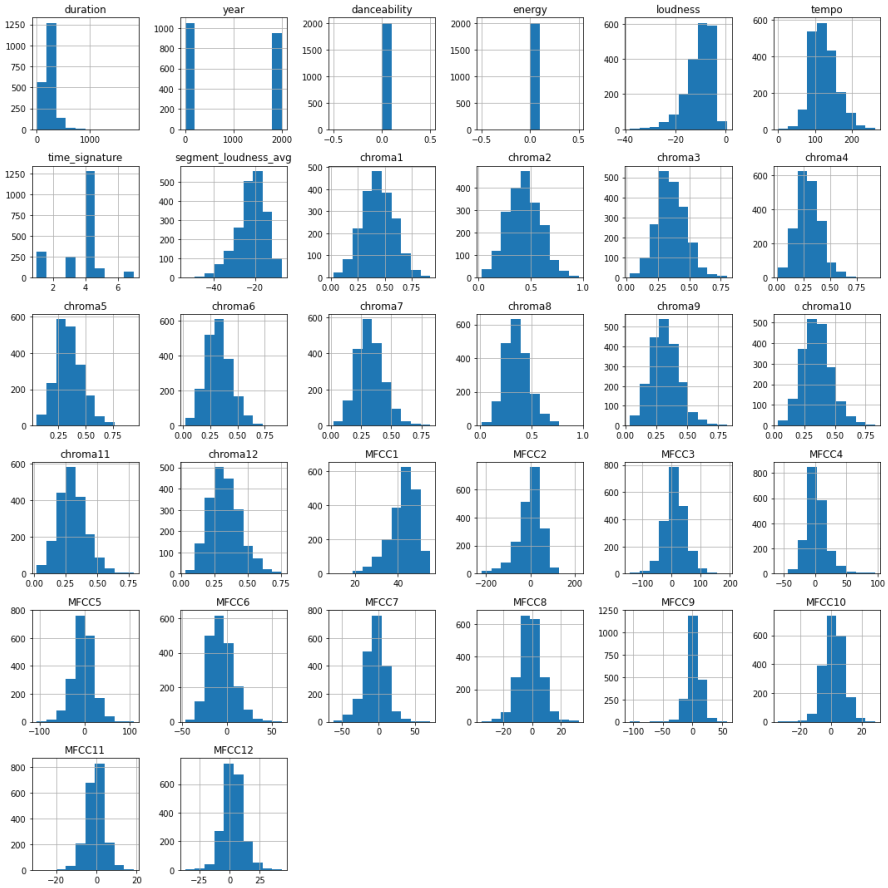

Рисунок 1- Гистограммы

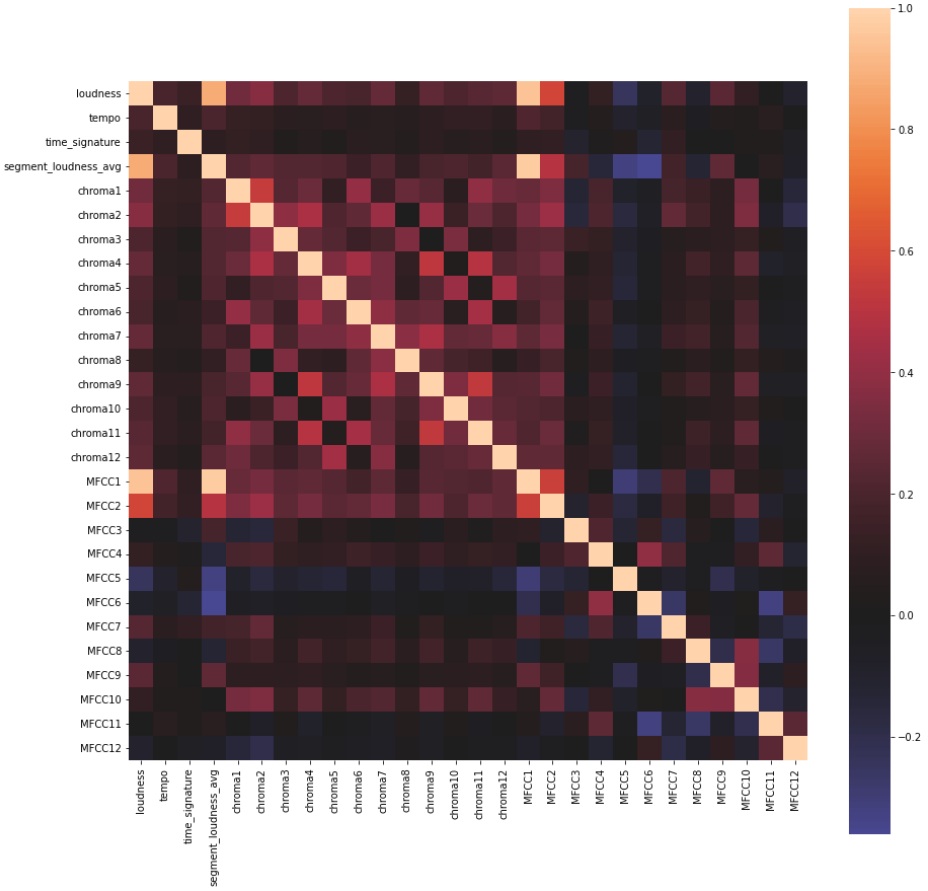

Рисунок 2 – Тепловая карта

3. Предварительная обработка и кластеризация данных. В данном разделе проводится предварительная обработка и кластеризация данных. Основное внимание уделяется нормализации признаков с использованием инструмента MinMaxScaler из библиотеки sklearn.preprocessing. Нормализация позволяет привести значения признаков к определенному диапазону, обычно от 0 до 1. Это необходимо для обеспечения правильной работы алгоритмов машинного обучения, основанных на расстоянии, таких как K-средних. Масштабирование признаков помогает уравнять их важность и делает алгоритм более чувствительным к небольшим изменениям в любом признаке. После масштабирования и преобразования данных в формат DataFrame с помощью библиотеки pandas, данные готовы для использования в модели K-средних.

4. Разработка модели. В данном разделе основное внимание уделяется применению алгоритма K-средних в разработке модели. Алгоритм K-средних является неконтролируемым методом машинного обучения, который позволяет выявить структуру и зависимости в данных без заданного выходного значения. В данном случае, используя базу из 2000 песен, целью является разделение их на 10 категорий (кластеров) на основе их музыкальных характеристик. В разделе также упоминается альтернативный метод кластеризации - иерархическая кластеризация. Однако, из-за требований к большим объемам памяти, иерархическая кластеризация становится непрактичной для больших наборов данных, поэтому в данном случае выбран метод K-средних вместо нее.

5. Кластерное распределение. В данном разделе применяется алгоритм K-средних из библиотеки scikit-learn для кластеризации данных. Предварительно обработанный и нормализованный набор данных подается на вход алгоритму, с указанием количества кластеров равного 10. После обучения модели на данных с помощью метода fit(), получаются метки кластеров для каждой песни, которые сохраняются в переменной predictions. Затем эти метки добавляются в исходный набор данных в виде нового столбца

clusters, что позволяет идентифицировать кластер, к которому относится каждая песня. Применение этого подхода позволяет сегментировать песни на 10 уникальных групп, основываясь на их музыкальных характеристиках. Это создает структуру для рекомендательной системы, где можно предложить пользователю песни из того же кластера, к которому относятся песни, которые он ранее прослушал и оценил высоко. Такой подход повышает точность рекомендаций и уровень удовлетворения пользователей.

6. Результаты для системы рекомендаций на основе k-средних. В данном разделе после уменьшения размерности конечных кластеров с использованием метода главных компонент (PCA) производится их визуализация. Отмечается, что кластеры не имеют четкой границы разделения, но большинство песен все равно группируются внутри своих кластеров. Для улучшения результатов возможно увеличение размера вектора признаков путем анализа звуковых характеристик каждого сегмента вместо вычисления среднего значения для всей песни. Также отмечается, что удаление выбросов может улучшить результаты, особенно при анализе больших наборов данных.

В последней главе рассматриваются перспективы для улучшения системы музыкальных рекомендаций. Предлагается упорядочить песни внутри каждого кластера на основе их близости к исходной песне с использованием метода KD Tree и других данных, таких как популярность и коэффициент кликабельности (CTR). Также предлагается создание аналитического центра, который использует различные функции, включая метаданные и аудио характеристики песен. Для улучшения рекомендаций предлагается использовать модель Word2Vec для анализа истории прослушивания пользователя и преобразования каждой песни в вектор, учитывающий ее "семантическое" содержание. Затем предлагается объединить векторы песен, полученные с помощью Word2Vec, с данными аудио характеристик и метаданными. Для сокращения размерности разреженных данных предлагается использовать автоэнкодеры. Наконец, предлагается применить алгоритм K-средних для кластеризации песен и

ранжирования их внутри каждого кластера на основе близости к центру. Этот гибридный подход объединяет традиционную систему музыкальных рекомендаций с методами машинного обучения и может привести к более точным рекомендациям, учитывающим интересы каждого пользователя. В целом, данная система музыкальных рекомендаций имеет потенциал для дальнейшего улучшения и расширения.

Заключение. В ходе выполнения данной дипломной работы было произведено исследование и анализ существующих подходов к созданию рекомендательных систем в области музыкального контента. Было рассмотрено множество методов, включая коллаборативную фильтрацию, контент-ориентированный подход, гибридные системы и др. На основании проведенного анализа было принято решение о выборе контент-ориентированного подхода с использованием алгоритма K-средних на основе аудио характеристик песен.

Причины выбора данного подхода заключаются в следующем:

1. Контент-ориентированный подход позволяет решить проблему холодного старта, которая встречается в других методах. Это означает, что даже новым или низкоактивным пользователям, у которых нет достаточной истории прослушивания, можно предложить качественные рекомендации на основе анализа аудио характеристик треков.
2. Алгоритм K-средних помогает выявить кластеры треков с похожими аудио характеристиками. Это обеспечивает гибкость системы и возможность предлагать пользователям треки, которые максимально соответствуют их предпочтениям, учитывая их музыкальные вкусы.

В результате проведенного анализа и выбора метода подтверждается необходимость создания нового алгоритма рекомендаций, основанного на аудио характеристиках и кластеризации. Этот подход улучшит качество предлагаемых рекомендаций, повысит удержание пользователей и способствует их превращению в постоянных пользователей платформы.

Внедрение данного подхода ожидается привести к сокращению затрат на удержание пользователей и обеспечению более глубокой персонализации музыкального контента, улучшая общий опыт пользователей на платформе.