**1 Слайд**

Здравствуйте уважаемый председатель и члены комиссии, в данном докладе по ВКР мной будут рассмотрены основные понятия, принципы и определения алгоритмов кластеризации данных.

Актуальность данной темы характеризуется наличием ряда важных прикладных задач, которые требуют наличие больших наборов данных, необходимы для исследования. Огромные объёмы и сложность массивов сильно снижают эффективность различных методов статистического анализа. В результате появляется необходимость в средствах предварительной фильтрации и группировки данных по признакам, которая осуществляется при помощи кластеризации.

**2 Слайд**

Так как мой проект комплексный, то моей основной целью являлась разработка алгоритма кластеризации текстовых данных, хранимых в электронной библиотеке сайта «Литература»

**3 Слайд**

Среди задач данного проекта можно выделить:

– анализ предметной области

– обзор инструментария разработки

– разработка алгоритма кластеризации

– обработка текстовых данных

– оценка эффективности

**4 Слайд**

Центральным элементом в проводимых мной исследованиях и разработке был веб-сайт «Литература», который представляет собой некую электронную библиотеку для хранения, чтения и загрузки произведений русской литературы.

Применение кластеризации обуславливается, тем что на данном ресурсе любой превилигилированный пользователь может загрузить то или иное произведение и при забыть указать автора или жанр. Подобные характеристики необходимы для навигации и поиска по сайту.

Кластеризация позволяет быстро провести группировку книг и выявить недостающие параметры.

**5 Слайд**

C# – это объектно-ориентированный высокоуровневый язык программирования с широким спектром возможностей и задач.

Данный язык применяется для:

– веб-программирования;

– создания многофункциональных десктопных приложений;

– тестирования и отладки;

– написания скриптов;

– разработки игр с использованием среды Unity.

Visual Studio 2019 – это IDE, включающая обширный спектр средств для работы с различными языками программирования.

– рефакторинг

– разнообразие поддерживаемых языков программирования

– IntelliSense

– NuGet

– интуитивный стиль кодирования

ASP.NET Core – это платформа для разработки веб приложений, которая была разработана компанией Microsoft как полностью обновлённая версия более старого фреймворка ASP.NET.

– встроенный сервер (ISS, Kestrel)

– Razor Pages

– Blazor

– MVC

– Entity

– Identity

ML.NET – это библиотека, предоставляющая обширный функционал для проведения машинного обучения на данных.

ML.NET позволяет использовать широкий спектр алгоритмов прогнозирования, от нейронных сетей Кохонена до k-means++. При помощи этой технологии можно производить следующие прогнозы:

– классификация, разбивающая информация по классам;

– регрессия, благодаря которой можно оценить, например, цены на мороженное отталкиваясь от места жительства и времени года;

– обнаружения аномалий, позволяющее выявить ошибки в расчётах;

– рекомендации, выявление такой информации, которую покупатель захочет увидеть прежде всего;

– кластеризация, которая разбивает данные на группы без каких-либо изначальных установок.

**6 Слайд**

Перед тем как перейти к непосредственному разбору принципов и методов кластеризации мне бы хотелось развеять одно из самых больших заблуждений в сфере анализа данных.

Кластеризация – это процесс разделения наборов данных на кластеры.

Существует схожий процесс – классификация, которая группирует данные по классам.

Исходя из определений данных процессов можно подумать, что они представляют собой одно и тоже. Однако это не так.

В теории машинного обучения это два совершенно разных процесса.

Классификация – обучение с учителем, следовательно, требует затрат времени и ресурсов на обучение, требует, чтобы наборы данных уже были разделены на классы, обладает очень высокой точностью после обучения

Кластеризация – обучение без учителя, работает сразу, группирует любые наборы данных без какой-либо информации о них, обладает средней точностью, которая сильно зависит от самих наборов данных.

**7 Слайд**

Суть кластеризации состоит в работе с точками на координатной оси. Каждый объект в исследуемом наборе данных преобразуется в вектор или же точку в многомерном Евклидовом пространстве. На данном слайде вы можете видеть простейший пример, когда все данные представлены точками в двухмерном пространстве, конечно на практике такого практически никогда не происходит ведь обычно исследуемые объекты очень сложны и обладают большим набором характеристик, а чем больше характеристик тем более многомерным является пространство.

Для того, чтобы разбить объекты на группы, алгоритм просто оценивает расстояния между этими точками.

**8 Слайд**

Сердцем алгоритма k-means является циклический процесс, в ходе которого оцениваются все точки из набора данных, а затем каждая из них соотносится с ближайшим к ней кластерным центром – центроидом. Стоит отметить, что центроиды каждого кластера постоянно перемещаются с каждым новым циклом алгоритма, точно также, как и точки постоянно переходят от одного кластера к другому. Когда все перемещения прекращаются или становятся пренебрежительно незначимыми алгоритм останавливается и выдаёт результат в виде данных разделённых на k кластеров.

**9 Слайд**

Однако алгоритм k-means практически никогда не используется в своём изначальном виде. На сегодняшний день абсолютно во всех областях науки, использующих плоские алгоритмы кластеризации, применяется модификация – k-means++. Данная модификация позволяет полностью избавиться от погрешностей, вызванных неопределённостью в выборе центров изначальных кластеров.

**10 Слайд**

Книга, загруженная из любого текстового файла помимо обычного русского текста, также содержит нежелательные особые символы, знаки препинания и слова на других языках, которые могу значительно замедлить работу алгоритма, а также привести к снижению точности конечных результатов.

Поэтому первым этапом обработки текста является избавление от всех нежелательных символов. В листинге 7.1 приведён код программы, который приводит все символы текста к нижнему регистру, а также заменяет все символы кроме русских слов и пробелов на пустую строку.

Второй этап – токенизация текста. В качестве токенов в данном случае выступают слова.

Третий этап подразумевает использование Алгоритма стемминга, разработанного Мартином Портером, который позволяет сократить количество слов путём отсекания их окончаний и суффиксов.

**11 Слайд**

Всё на слайде

**12 Слайд**

Пример TF-IDF

**13 Слайд**

Далее мной была произведена оценка эффективности алгоритма на различных наборах данных, которые отличаются количеством книг и авторов.

Зелёным цветом выделена таблица для алгоритма, разработанного с нуля, а красным – реализация с использованием библиотеки ML.NET.

Отсюда можно увидеть, что реализация, представленная библиотекой ML.Net (красный цвет) работает быстрее, но значительно теряет в точности с увеличением количества книг и авторов. Алгоритм из зелёной таблицы гораздо более точный, но требует больших затрат времени.

Можно сделать вывод, что при анализе небольшого количества данных выгоднее использовать ML.NET. Если же есть необходимость в кластеризации больших наборов данных, то лучше использовать мою реализацию.

**14 Слайд**

У разработанного мной приложение можно выделить следующие достоинтсва:

**15 Слайд**

В результате было

**Видео**