

Министерство образования и науки Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Институт кибернетики

Направление подготовки 09.04.02 Информационные системы и технологии

Кафедра вычислительной техники

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Применение методов и средств технологии Business Intelligence для анализа данных о судебной практике в Томской области

УДК _____

Студент

Группа	ФИО	Подпись	Дата
8ИМ4А	Хлопонин А.А.		

Руководитель

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент каф. ВТ	Кудинов А.В.	К.Т.Н		

КОНСУЛЬТАНТЫ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент каф. МЕН	Антонова И.С.	К.Э.Н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент каф. ЭБЖ	Пустовойтова М.И.	К.Х.Н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Зав. кафедрой	ФИО	Ученая степень, звание	Подпись	Дата
ВТ	Марков Н.Г.	Д.Т.Н., профессор		

Томск – 2016 г.

Планируемые результаты обучения по ООП

Код результата	Результат обучения (выпускник должен быть готов)
<i>Общепрофессиональные компетенции</i>	
P1	Воспринимать и самостоятельно приобретать, развивать и применять математические, естественнонаучные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте.
P2	Владеть и применять методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе в глобальных компьютерных сетях.
P3	Демонстрировать культуру мышления, способность выстраивать логику рассуждений и высказываний, основанных на интерпретации данных, интегрированных из разных областей науки и техники, выносить суждения на основании неполных данных, анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями.
P4	Анализировать и оценивать уровни своих компетенций в сочетании со способностью и готовностью к саморегулированию дальнейшего образования и профессиональной мобильности. Владеть, по крайней мере, одним из иностранных языков на уровне социального и профессионального общения, применять специальную лексику и профессиональную терминологию языка.
<i>Профессиональные компетенции</i>	
P5	Разрабатывать стратегии и цели проектирования, критерии эффективности и ограничения применимости, новые методы, средства и технологии проектирования ИС или промышленного программного обеспечения.
P6	Планировать и проводить теоретические и экспериментальные исследования в области создания интеллектуальных ИС и ИС технологии или промышленного программного обеспечения с использованием методов системной инженерии.
P7	Осуществлять авторское сопровождение процессов проектирования, внедрения и сопровождения ИС и ИС технологий или промышленного программного обеспечения с использованием методов и средств системной инженерии, осуществлять подготовку и обучение персонала.
P8	Формировать новые конкурентоспособные идеи в области теории и практики ИС и ИС технологий или системной инженерии программного обеспечения. Разрабатывать методы решения

	нестандартных задач и новые методы решения традиционных задач. Организовывать взаимодействие коллективов, принимать управленческие решения, находить компромисс между различными требованиями как при долгосрочном, так и при краткосрочным планировании.
<i>Общекультурные компетенции</i>	
P9	Использовать на практике умения и навыки в организации исследовательских, проектных работ и профессиональной эксплуатации современного оборудования и приборов, в управлении коллективом.
P10	Свободно пользоваться русским и иностранным языками как средством делового общения.
P11	Совершенствовать и развивать свой интеллектуальный и общекультурный уровень. Проявлять инициативу, в том числе в ситуациях риска, брать на себя всю полноту ответственности.
P12	Демонстрировать способность к самостоятельному обучению новым методам исследования, к изменению научного и научно-производственного профиля своей профессиональной деятельности, способность самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности, способность к педагогической деятельности.

Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Институт

Кибернетики

Направление подготовки 09.04.02 «Информационные системы и технологии»

Кафедра Вычислительной техники

УТВЕРЖДАЮ:

Зав. кафедрой ВТ

(Подпись) _____ (Дата) Марков Н.Г.

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации
(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	ФИО
8ИМ4А	Хлопонин Александр Александрович

Тема работы:

Применение методов и средств технологии Business Intelligence для анализа данных о судебной практике в Томской области	
Утверждена приказом директора (дата, номер)	05.02.2016 №776/с

Срок сдачи студентом выполненной работы:	01.06.2016
--	------------

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	Сайты судов Томской области, тексты судебных решений
Перечень подлежащих исследованию, проектированию и разработке вопросов	Анализ предметной области и существующих аналогов. Извлечение данных из текстов решений. Анализ данных с помощью технологий Data Mining.
Перечень графического материала	Мультимедийная презентация
Консультанты по разделам выпускной квалификационной работы	
Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Антонова И.С.
Социальная ответственность	Пустовойтова М.И.
Раздел на иностранном языке	Шепетовский Д.В.

	Мирошниченко Е.А.
Названия разделов, которые должны быть написаны на русском и иностранном языках:	
Data Mining технологии (Data Mining technologies)	

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	
---	--

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент каф. ВТ	Кудинов А.В.	К.Т.Н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ИМ4А	Хлопонин Александр Александрович		

Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Институт кибернетики

Направление подготовки 09.04.02 «Информационные системы и технологии»

Уровень образования магистр

Кафедра вычислительной техники

Период выполнения осенний / весенний семестр 2015/2016 учебного года

Форма представления работы:

Магистерская диссертация
(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

**КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы**

Срок сдачи студентом выполненной работы:	01.06.2016
--	------------

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
04.03.2016 г.	Анализ предметной области	15
25.04.2016 г.	Проектирование и разработка инструментов и базы данных	25
10.05.2016 г.	Анализ данных судебных дел с помощью технологии Data Mining	20
18.05.2016 г.	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
24.05.2016 г.	Социальная ответственность	10
28.05.2016 г.	Обязательное приложение на иностранном языке	10
01.06.2016 г.	Оформление пояснительной записки	10

Составил преподаватель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент каф. ВТ	Кудинов А.В.	к.т.н.		

СОГЛАСОВАНО:

Зав. кафедрой	ФИО	Ученая степень, звание	Подпись	Дата
Профессор	Марков Н.Г.	д.т.н., профессор		

Реферат

Выпускная квалификационная работа 126 с., 50 рис., 37 табл., 19 источников, 3 прил.

Ключевые слова: БАЗА ДАННЫХ, OLAP, DATA MINING, БОЛЬШИЕ ДАННЫЕ, BUSINESS INTELLIGENCE, СУДЫ.

Объектом исследования является задача нахождения скрытых закономерностей, а предметом исследования – решения по судебным делам.

Цель работы – анализ записей, полученных в результате обработки документов и решений судов Томской области с помощью методов и средств технологии Business Intelligence.

В процессе исследования проводился анализ предметной области судопроизводства Томской области, кроме того был выполнен обзор существующих аналогов.

В работе приведены результаты проектирования и реализации базы данных, а также проектирование и разработка инструментов для извлечения параметров, необходимых для проведения анализа, и наполнения базы данных судопроизводства. Приведены результаты OLAP анализа данных судебных решений Томской области.

В дальнейшем, разработанные инструменты могут позволить заинтересованным людям самостоятельно извлекать данные и проводить их анализ. Анализ можно расширить путем увеличения объема анализируемых решений и проводить анализ не только по судам Томской области, а по всем судам Российской Федерации.

Определения, сокращения

ФИО – фамилия имя отчество;

БД – база данных;

BI – Business Intelligence;

OLAP (Online Analytical Processing) – это оперативный анализ данных;

Data Mining – добыча данных, интеллектуальный анализ данных;

УК РФ – Уголовный Кодекс Российской Федерации;

РФ – Российская федерация;

Microsoft SQL Server — система управления реляционными базами данных (СУРБД), разработанная корпорацией Microsoft.

SQL Server Express LocalDB — это упрощенная версия SQL Server, которая имеет все функции программирования базы данных SQL Server.

Оглавление

Введение.....	10
1 Анализ предметной области судопроизводства	12
1.1 Актуальность анализа данных судопроизводства.....	12
1.2 Анализ структуры данных судебных решений	13
1.3 Обзор аналогов.....	15
1.3.1 «CaseLook»	15
1.3.2 «РосПравосудие»	17
1.4 Постановка задачи	18
2 Проектирование инструментов для обработки данных	20
2.1 Выбор инструментов разработки	20
2.2 Проектирование базы данных судопроизводства	21
2.3 Проектирование инструмента для извлечения атрибутов судебного решения	23
2.4 Проектирование инструмента для наполнения базы данных	30
3 Разработка инструментов для обработки данных	33
3.1 Разработка инструмента для извлечения атрибутов судебного дела..	33
3.2 Разработка инструмента для наполнения базы данных	34
4 Анализ данных с использованием технологии Data Mining.....	35
5 Финансовый менеджмент, ресурсоэффективность и ресурсосбережения	
Ошибка! Закладка не определена.	
6 Социальная ответственность	Ошибка! Закладка не определена.
Заключение	76
Список использованных источников	77
Приложение А	80
Приложение Б	81
Приложение В.....	82

Введение

В наше время происходит интенсивное накопление огромных объёмов данных разного типа в различных предметных областях измеряемые в петабайтах, это в свою очередь даёт возможность решать задачи получения новых фактов, зависимостей и скрытых корреляций, а также позволяет решать некоторые аналитические задачи, такие как прогнозирование, проверка статистических гипотез, расчёт агрегатных показателей и т.д.

Объектом исследования является задача нахождения скрытых закономерностей, а предметом исследования – решения по судебным делам.

Цель работы – анализ записей, полученных в результате обработки документов и решений судов РФ с помощью методов и средств технологии Business Intelligence.

Анализ судопроизводства довольно актуален, поскольку результаты анализа будут интересны как большим организациям, таким как правоохранительные органы и правозащитные организации, так и отдельным лицам, таким как независимые юристы и люди, находящиеся под следствием. Для больших организаций будет более интересен глубокий анализ для поиска скрытых зависимостей и отклонений, а отдельным лицам анализ по конкретным судьям и адвокатам, который может помочь предсказать какое решение может быть вынесено определенным судьей или выбрать наиболее эффективного адвоката.

Значимость данной работы состоит в том, что разработанные инструменты позволят заинтересованным людям самостоятельно извлекать и анализировать данные.

В первом разделе данной работы проведен анализ предметной области, обоснована актуальность исследования и изложены результаты обзора аналогов.

Во втором разделе представлены проектирование и реализация базы данных, а также приведены результаты проектирования инструментов для

извлечения атрибутов из текстов судебных дел и наполнения базы данных судопроизводства.

В третьем разделе представлены результаты разработки инструментов для извлечения атрибутов из текстов судебных дел и наполнения базы данных судопроизводства.

В четвертом разделе представлены результаты OLAP анализа данных, извлеченных из решений по судебным делам.

В пятом разделе рассмотрены вопросы финансового менеджмента, результаты расчетов и анализ ресурсоэффективности и ресурсосбережения.

В шестом разделе описаны основные требования социальной ответственности.

Извлечение данных выполнено в среде PyCharm, язык программирования – Python. Запись данных в БД выполнена в среде Visual Studio 2013, язык программирования – C#. Для создания БД использовалась среда Microsoft SQL Server 2014 Management Studio. OLAP и Data Mining анализ выполнен в среде Visual Studio 2013, с использованием Excel 2013. Пояснительная записка оформлена в Microsoft Word 2013.

1 Анализ предметной области судопроизводства

1.1 Актуальность анализа данных судопроизводства

Впервые термин «большие данные» использовал Клиффорд Линч в выпуске журнала Nature от 3 сентября 2008 года, с темой «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объёмами данных?». После этого направление обработки больших данных начало стремительно развиваться. Так, в 2010 году уже начали появляться первые программные продукты и решения для обработки больших данных.

1 июля 2010 года вступил в силу Федеральный закон «Об обеспечении доступа к информации о деятельности судов в Российской Федерации» [1], который предусматривает размещение решений по судебным делам на сайтах судов. Благодаря данному закону, появилась возможность открытого доступа к судебным решениям, и соответственно возможность выполнить анализ дел судопроизводства.

Таким образом, имеются не только инструменты необходимые для анализа, но и данные судопроизводства находятся в открытом доступе. Уже есть несколько проектов, занимающихся анализом судебных дел, однако, эти проекты либо не предоставляют возможности анализа вообще, либо набор функций довольно ограничен и доступ к ним платный.

Анализ судопроизводства довольно актуален, поскольку результаты анализа будут интересны как большим организациям, таким как правоохранительные органы и правозащитные организации, так и отдельным лицам, таким как независимые юристы и люди, находящиеся под следствием. Поскольку позволят не только собрать статистическую информацию, но и увидеть более глубокий анализ по отдельным судьям, адвокатам и другим атрибутам судебных дел для определения различных скрытых взаимосвязей между атрибутами дела и для поиска отклонений. Результат анализа может помочь предсказать какое решение может быть вынесено определенным судьей или помочь выбрать наиболее эффективного адвоката.

Задачей данной работы является анализ решений по судебным делам, с помощью методов и средств Business Intelligence, для выявления скрытых закономерностей.

1.2 Анализ структуры данных судебных решений

Поскольку дела судопроизводства появились в открытом доступе сравнительно недавно, структура официальных веб-сайтов судов, на которых размещаются судебные дела, претерпевала изменения. Так, на некоторое время на сайтах судов появились данные о движении дела, однако довольно быстро эти данные были убраны. Поэтому пришлось ориентироваться исключительно на содержимое судебных актов, это связано с крайней разобщённостью судов в отношении публикаций движения дел, а также с отсутствием таковой информации на некоторых сайтах судов вообще.

На сайтах судов дела выложены после некоторой обработки, которая выделила некоторые атрибуты: имя судьи, дата суда, вид дела, подвид дела, инстанция, регион, город. Однако, это лишь часть необходимых атрибутов. Поэтому, был проведен анализ текстов судебных дел для определения списка атрибутов, которые необходимо и возможно извлечь. Основная сложность заключается в том, что тексты судебных дел довольно не структурированы и некоторые данные, например, имена некоторых ответчиков, обезличены.

Текст судебного дела можно разделить на четыре части:

1. В вводной части решения суда должны быть указаны дата и место принятия решения суда, наименование суда, принявшего решение, состав суда, секретарь судебного заседания, стороны, другие лица, участвующие в деле, их представители, предмет спора или заявленное требование.

2. Описательная часть решения суда должна содержать указание на требование истца, возражения ответчика и объяснения других лиц, участвующих в деле.

3. В мотивировочной части решения суда должны быть указаны обстоятельства дела, установленные судом; доказательства, на которых основаны выводы суда об этих обстоятельствах; доводы, по которым суд отвергает те или иные доказательства; законы, которыми руководствовался суд.

4. Резолютивная часть решения суда должна содержать выводы суда об удовлетворении иска либо об отказе в удовлетворении иска полностью или в части, указание на распределение судебных расходов, срок и порядок обжалования решения суда. Резолютивная часть решения суда, принятого мировым судьей, также должна содержать указание на срок и порядок подачи заявления о составлении мотивированного решения суда [2].

Таким образом, необходимые атрибуты судебного дела в основном содержатся в основной и резолютивной частях дела.

В результате анализа текстов судебных дел и сайтов судов Томской области, было выявлено следующее. По типу судебные дела делятся на:

- административные;
- гражданские;
- уголовные;
- по материалам.

В зависимости от типа судебного дела можно выделить типы судебных решений:

- постановления;
- решения;
- определения;
- приговоры.

Каждое судебное решение имеет следующие общие атрибуты:

- номер дела;
- город;
- регион;
- название суда;

- судья (ФИО, пол);
- дата суда;
- нормативный акт (номер статьи, название кодекса);
- дата вступления в силу;
- решение по делу;
- физические лица, фигурирующие в делах.

Кроме основных атрибутов судебное решение может иметь дополнительные атрибуты, в зависимости от типа дела и вынесенного решения:

- наказание;
- размер наказания;
- мера наказания;
- ответчик;
- истец;
- ФИО адвоката;
- юридические лица, фигурирующие в делах.

У одного дела может быть несколько решений, если в деле фигурирует несколько ответчиков. В свою очередь одно решение может содержать несколько наказаний, в случае вынесения нескольких видов наказаний (арест, взыскание и т.д.). Кроме того, у каждого решения есть свой список статей, на основании которых было вынесено решение.

1.3 Обзор аналогов

1.3.1 «CaseLook»

Сервис «CaseLook» позиционируется как универсальный инструмент для поиска и анализа судебной практики. «CaseLook» - это коммерческий проект, который предусматривает повременную оплату начиная от месяца до года [3].

Сервис «CaseLook» имеет следующие элементы для фильтрации судебных решений из имеющейся в данном сервисе базы данных:

Судебный акт:

- текст;
- тип документа;
- дата принятия;
- дата поступления;
- норма права.

Судья и суд:

- суд;
- судья;
- тип суда;
- нижестоящий суд;
- субъект РФ.

Дело:

- вид спора;
- категория спора;
- номер дела;
- исход дела;
- участник;
- представитель.

В данное время аналитические инструменты в сервисе «CaseLook» находятся в экспериментальном режиме и не представлены в полном объёме, хотя заявлены на сайте и оглашаются во время переговоров со службой поддержки.

Основные свойства сервиса:

- Огромная база судебных решений.
- Широкие поисковые возможности документов.
- Анализ документов.
- Постоянное обновление судебных актов.
- Оценка вероятности исхода дела (экспериментальный режим).

Положительные стороны:

- Удобный интерфейс пользователя.
- Предоставление тестового режима на две недели для ознакомления.
- Служба поддержки клиентов.

Недостатки:

- Высокая цена использования сервиса.
- Закрытая база данных.
- Ограниченный набор задач.
- Некоторые инструменты работают в тестовом режиме.

1.3.2 «РосПравосудие»

«РосПравосудие» - проект, созданный в 2012 году для исследования общедоступной судебной практики.

«РосПравосудие» обеспечивает возможность изучения судебной практики с помощью анализа текстов решений, публикуемых судами на основании Федерального закона «Об обеспечении доступа к информации о деятельности судов в Российской Федерации» от 22 декабря 2008 г. за № 262-ФЗ. Тексты решений получаются с официальных сайтов судов общей юрисдикции, арбитражных и мировых судов и отражают только те решения, которые опубликованы судами за последние несколько лет.

Внесение изменений в тексты документов и принятие решений об их публикации является исключительной компетенцией работников суда.

Навигация на сайте «РосПравосудие» происходит при помощи фильтров, последовательно сужающих перечень документов.

В качестве фильтра может использоваться все, что угодно - регион, суд, категория дела, интервал времени, участник процесса или фраза.

Проект «РосПравосудие» позиционирует себя как общественный проект призванный сделать судебную систему прозрачной.

Финансирование проекта осуществляется на добровольное пожертвование людей.

Основные свойства сервиса:

- Огромная база судебных решений.
- Широкие поисковые возможности документов.
- Анализ документов.
- Постоянное обновление судебных актов.
- Содержание на сайте некоторой статистики по судебным решениям.

Положительные стороны:

- Удобный интерфейс пользователя.
- Бесплатный доступ к сервису.
- Очень широкие возможности по фильтрации судебных решений.

Недостатки:

- Отсутствие аналитических инструментов в режиме реального времени.
- Закрытая база данных.
- Медленное развитие проекта из-за отсутствия постоянной финансовой поддержки.

1.4 Постановка задачи

В результате анализа предметной области, включающий обзор аналогов, было выявлено, что базы данных по судопроизводству нет в свободном доступе. Поэтому появилась необходимость самостоятельно разработать БД, извлечь параметры судебного дела и наполнить БД. На рисунке 1 представлена общая схема анализа судопроизводства, которая демонстрирует этапы, через которые необходимо пройти для проведения анализа данных судопроизводства.



Рисунок 1 – Общая схема подготовки данных для анализа

Для того, чтобы выполнить анализ данных судопроизводства, необходимо выполнить следующие задачи:

1. Выбор инструментов проектирования, разработки и проведения анализа.
2. Проектирование и реализация базы данных.
3. Проектирование и разработка инструментов для извлечения параметров и наполнения базы данных.
4. Проведение OLAP анализа данных.

2 Проектирование инструментов для обработки данных

2.1 Выбор инструментов разработки

Для извлечения параметров судебных дел использовался язык программирования Python в среде разработки Pycharm. Основными достоинствами выбранной среды и языка является:

- Удобочитаемый синтаксис;
- Прост в обучении;
- Высокоуровневый язык;
- Объектно-Ориентированный язык программирования;
- Обширный набор подключаемых модулей.

Для создания БД использовался сервер Elastic Search который в то же самое время является NoSQL базой данных, благодаря следующим достоинствам:

- Удобный графический веб-интерфейс, позволяющий управлять данными через браузер;
- Богатые возможности, по манипуляции данными, а также наличие инструментов для выполнения общих статистических расчётов;
- Крайне высокая производительность, данный сервер легко держит нагрузку в промышленных масштабах;
- Надежность.

В качестве среды разработки приложения для наполнения БД использовалась среда PyCharm. Выбор данного программного обеспечения обусловлен следующими предоставленными возможностями:

- Интеграция с базой данных Elastic Search;
- Возможность формирования БД непосредственно в коде;
- Запись асинхронного кода простым и интуитивно понятным способом.

В качестве языка программирования использован язык Python 3, так как он обладает следующими преимуществами:

- объектно-ориентированность;
- универсальность использования;
- многофункциональность;
- относительная простота.

Для использования технологии Data Mining и проведения анализа данных, использовались дополнительные библиотеки, встраиваемые в среду Python:

- Numpy – библиотека позволяющая проводить математические расчёты.
- Matplotlib – библиотека расширяющая возможности построения диаграмм.
- Sklearn.manifold – библиотека позволяющая выполнять работу с данными в основе которой технологии Data Mining и Big Data.

2.2 Проектирование базы данных судопроизводства

На основе проведенного анализа структуры судебных решений была спроектирована схема базы данных.

База данных состоит из 25 сущностей со своими наборами атрибутов:

- Регионы (идентификатор региона, название региона);
- Города (идентификатор города, идентификатор региона, название города);
- Формы юридических лиц (идентификатор формы юридического лица, форма юридического лица);
- Юридические лица (идентификатор юридического лица, имя юридического лица, идентификатор формы юридического лица);
- Юридические лица дела (идентификатор юридического лица, идентификатор дела);

- Физические лица (идентификатор физического лица, имя физического лица);
- Физические лица дела (идентификатор физического лица, идентификатор дела);
- Истцы (идентификатор истца, тип истца, имя истца);
- Суды (идентификатор суда, название суда, идентификатор города);
- Виды дел (идентификатор вида дел, название вида дел);
- Подвиды дел (идентификатор подвида дел, название подвида дел);
- Дела (идентификатор дела, номер дела, идентификатор суда, идентификатор судьи, идентификатор адвоката, категория, дата суда, дата вступления в силу, идентификатор истца, идентификатор вида дела, идентификатор подвида дела, идентификатор инстанции);
- Судьи (идентификатор судьи, ФИО, пол судьи);
- Решения (идентификатор решения, идентификатор дела, идентификатор ответчика, идентификатор типа решения);
- Тип решения (идентификатор типа решения, тип решения);
- Инстанции (идентификатор инстанции, название инстанции);
- Наказания (идентификатор наказания, идентификатор решения, идентификатор типа наказания, размер наказания, идентификатор типа меры наказания);
- Статьи решения (идентификатор решения, идентификатор статьи);
- Адвокаты (идентификатор адвоката, ФИО);
- Типы наказаний (идентификатор типа наказания, название типа наказания);
- Типы мер наказаний (идентификатор типа меры наказания, название типа меры наказания);
- Кодексы (идентификатор кодекса, название кодекса);
- Статьи (идентификатор статьи, статья, идентификатор кодекса, номер статьи, идентификатор категории статьи);

- Категории статей (идентификатор категории статей, название категории статей).

У каждой сущности есть свой уникальный идентификатор, который позволяет однозначно определить элементы сущностей, а также помогает предотвратить дублирование данных. За исключением трех сущностей (Статьи решения, Юридические лица дел, Физические лица дел), которые имеют составной ключ, состоящий из двух соответствующих идентификаторов.

Кроме того, практически все сущности связаны с помощью бинарных связей, в данном случае используется связь один ко многим. Таким образом, Решение может принадлежать только одному Делу, но у Дела может быть множество Решений. Таким же образом связаны и остальные сущности: Юридические лица и Формы юридических лиц, Города и Регионы, Дела и Истцы, Суды и Города, Дела и Суды, Дела и Виды дел, Дела и Подвиды дел, Дела и Инстанции, Дела и Судьи, Дела и Адвокаты, Решения и Тип решения, Решения и Ответчики, Наказания и Решения, Наказания и Типы наказаний, Наказания и Типы мер наказаний, Статьи и Кодексы, Статьи и Категории статей. Исключение составляют сущности: Решения и Статьи, Дела и Юридические лица, Дела и Физические лица. В этих трех случаях между сущностями используется связь многие ко многим, сформированная за счет соответствующих промежуточных таблиц.

База данных создана в среде Microsoft SQL Server 2014 Management Studio в соответствии с данными проектирования. В приложении А представлена физическая схема БД.

2.3 Проектирование инструмента для извлечения атрибутов судебных решений

Извлечение данных представляет собой сложный процесс получения текстовых данных из одного или разных источников и приведения его к

унифицированному виду. Процессом извлечения конкретных данных, по заданным шаблонам называется парсингом. Обычно объектами парсинга являются веб-сайты, но также могут быть xml и pdf документы, и любые другие текстовые файлы и документы, хранящие информацию в текстовом виде.

В данной работе в качестве объектов для парсинга выступают xml документы. Xml файлы суда представляют собой набор нескольких параметров в xml тегах (имя судьи, дата суда, вид дела, подвид дела, инстанция, регион, город) и текст судебного дела.

Для получения параметров из текстов решений был разработан специальный инструмент. В котором применяются следующие методы и алгоритмы:

- Алгоритм шинглов — алгоритм, разработанный для поиска копий и дубликатов рассматриваемого текста в веб-документе. Шинглы — выделенные из текста подпоследовательности слов, идущих друг за другом. Выборка происходит внахлест, то есть каждый новый шингл сдвигается на одно слово [5].
- Расстояние Левенштейна между двумя строками — это минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую [6].
- Регулярные выражения — формальный язык поиска и осуществления манипуляций с подстроками в тексте, основанный на использовании метасимволов. По сути это строка-образец («шаблон», «маска»), состоящая из символов и метасимволов и задающая правило поиска [7].

Для каждого атрибута судебного дела была написана собственная функция, с соответствующими регулярными выражениями и алгоритмом поиска.

На рисунке 2 представлен алгоритм извлечения атрибутов из текстов судебных дел. Часть функций (извлечение номера дела, ФИО адвоката, истца, физических лиц, юридических лиц) не зависят от результатов остальных

функций. Однако некоторым функциям требуются данные из других функций. Так, для извлечения пола судьи требуется имя судьи, для решений требуется список ответчиков, для наказаний список ответчиков и решений, для статей список ответчиков и решений.

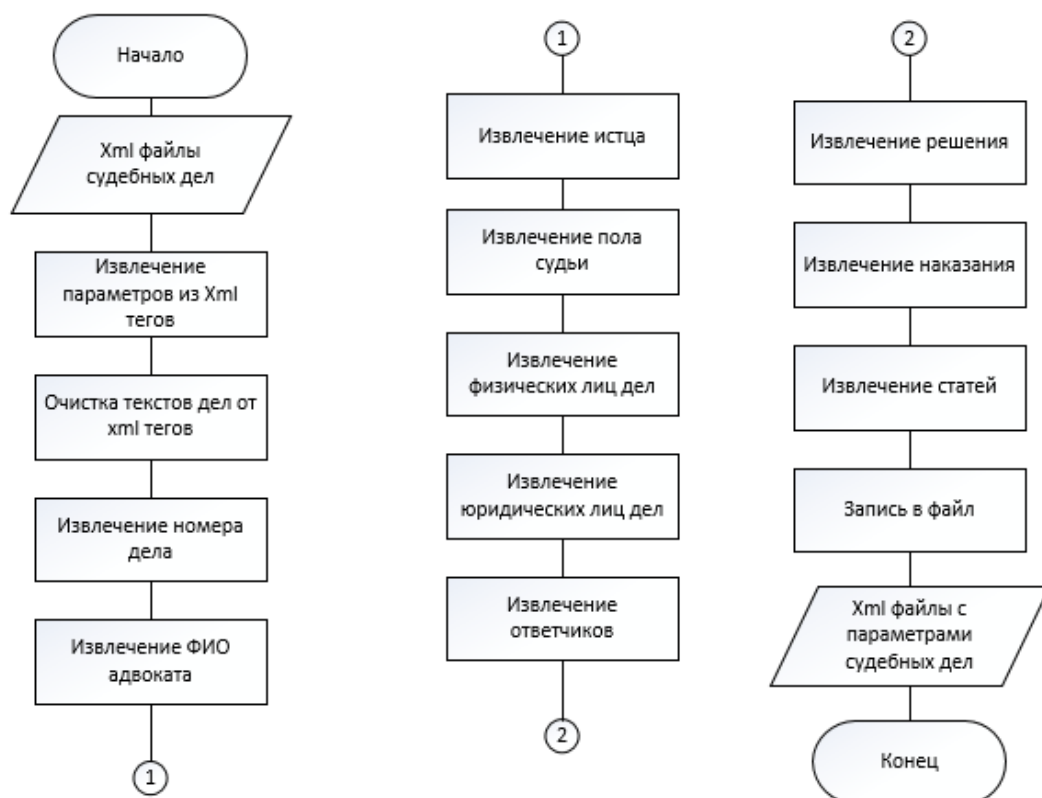


Рисунок 2 – Алгоритм извлечения атрибутов из текстов судебных решений

Извлечение параметров из xml тегов состоит в извлечении текста, заключенного в определенных xml тегах.

Выделение из текста дела резолютивной части, осуществляется с использованием регулярных выражений. Так, в тексте выполняется поиск заданных регулярных выражений (перед резолютивной частью ставятся следующие слова: постановление, решение и т.д.) при нахождении требуемого выражения выполняется обрезка текста от найденного выражения и до конца текста.

Выделение из резолютивной части дела фрагмента, относящейся к ответчику, осуществляется с использованием метода шинглов и расстояния Левенштейна. Поскольку имя ответчика может находиться в разных падежах в списке ответчиков и в тексте решения, текст разбивается на шинглы (с длиной

равной количеству слов в имени ответчика) и используя расстояние Левенштейна сравниваются шинглы решения и шингл имени ответчика.

Алгоритм для извлечения физических лиц находит в тексте все выражения, удовлетворяющие заданному набору регулярных выражений. После чего, используя расстояние Левенштейна, из списка удаляются все дубликаты. Для поиска дубликатов используется расстояние Левенштейна, поскольку в тексте дела одни и те же лица могут упоминаться в разных склонениях. Такой же алгоритм используется при поиске юридических лиц (рисунок 3).

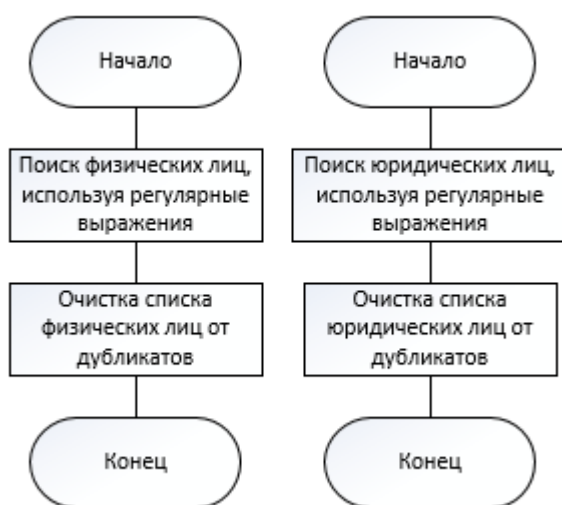


Рисунок 3 – Алгоритмы для извлечения физических и юридических лиц

Алгоритм для извлечения номера дела находит в тексте первое выражение, удовлетворяющее заданному набору регулярных выражений и отсекает ненужную более часть выражения, которая описывает слова, стоящие перед искомым номером дела. В качестве результата берется первое найденное выражение поскольку в деле может быть лишь один номер дела. Такой же алгоритм используется при поиске истца и адвоката. Алгоритм для извлечения пола судьи отличается тем, что поиск регулярных выражений происходит в ФИО судьи (рисунок 4).

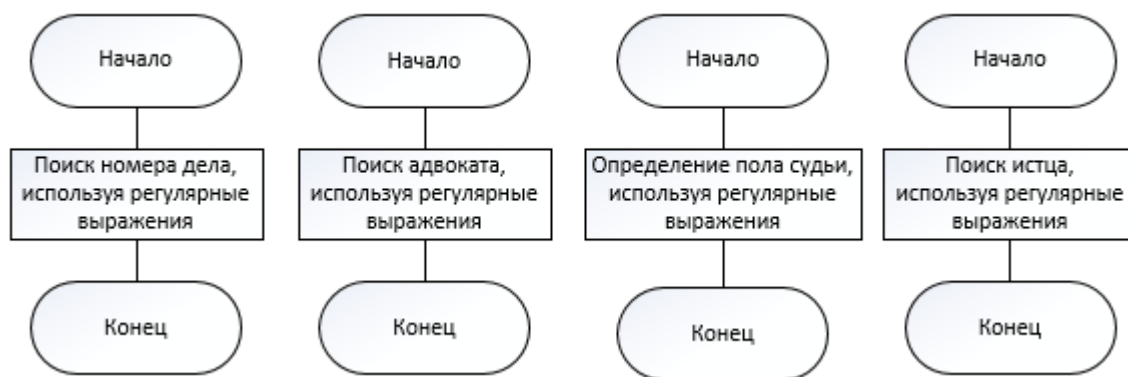


Рисунок 4 – Алгоритмы для извлечения номера дела, истца, адвоката и пола судьи

Алгоритм для извлечения ответчиков сперва выделяет часть текста с резолютивной частью. После этого идет по списку физических лиц и выделяет из решения часть текста, относящуюся к физическому лицу. Затем ищет в тексте заданные регулярные выражения, если выражение найдено, то в список ответчиков добавляется имя текущего физического лица. Если искомые выражения не найдены, то алгоритм переходит к следующему физическому лицу. Если не было найдено ответчиков среди физических лиц, то по такой же схеме, как для физических лиц, выполняется поиск ответчиков среди юридических лиц (рисунок 5).

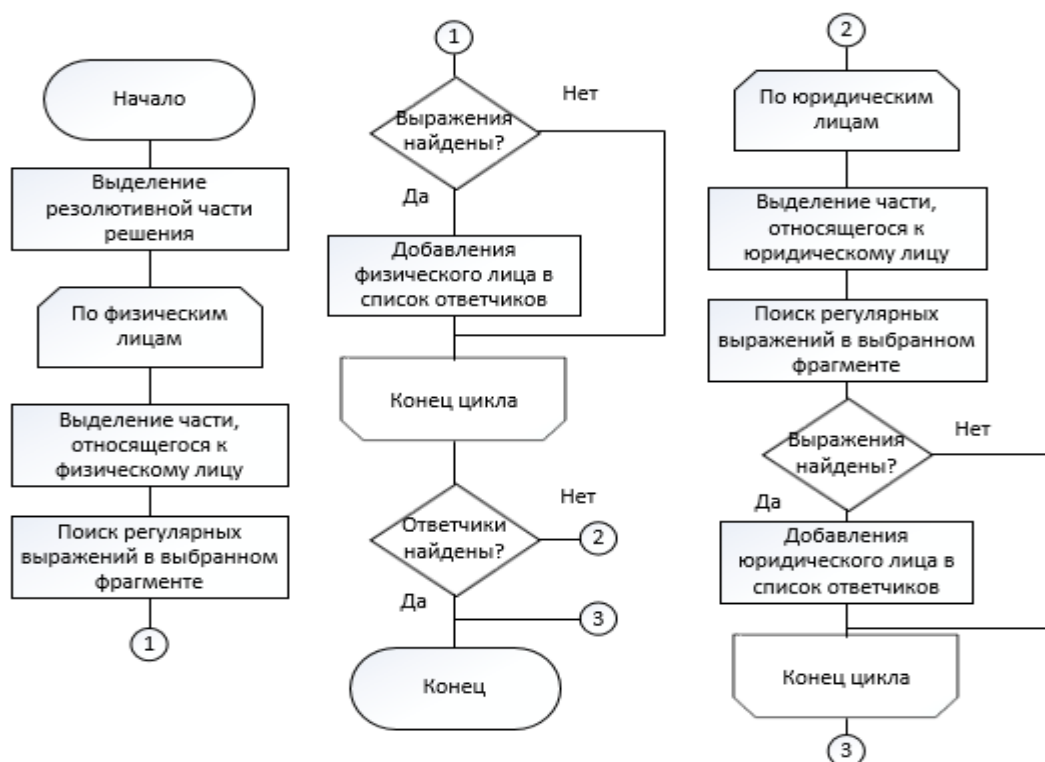


Рисунок 5 – Алгоритм для извлечения ответчиков

Алгоритм для извлечения решений сперва выделяет часть текста с резолютивной частью. Если есть ответчики, то алгоритм идет по списку ответчиков и выделяет из решения часть текста, относящуюся к ответчику. Затем ищет в тексте заданные регулярные выражения, если выражение найдено, то в список решений добавляется решение для текущего ответчика. Если искомые выражения не найдены, то алгоритм ведет поиск во всем решении по делу. Если же ответчиков нет, то поиск регулярных выражений ведется во всем тексте решения и результат записывается в список решений (рисунок 6).

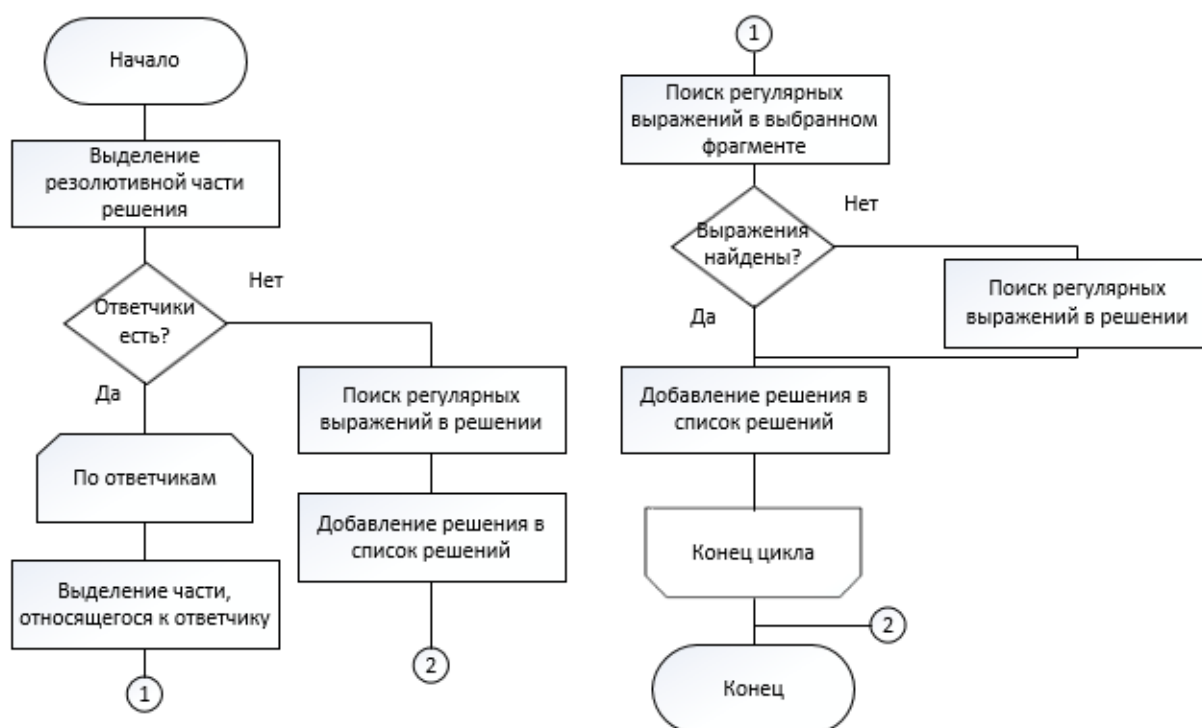


Рисунок 6 – Алгоритм для извлечения решений

Алгоритм для извлечения статей сперва выделяет часть текста с резолютивной частью. Если есть ответчики, то алгоритм идет по списку ответчиков и выделяет из решения часть текста, относящуюся к ответчику. Затем ищет в тексте регулярные выражения для номеров статей, если выражения найдены, то начинается поиск названия кодексов поблизости от номера статьи, с использованием регулярных выражений. Если название кодекса найдено, то в список статей добавляется номер статьи и название кодекса для текущего ответчика. Если кодекс не найден, то алгоритм переходит

к следующему ответчику. Если в выделенной части решения с ответчиком не были найдены номера статей, то поиск производится во всем решении. Если номера были найдены, то далее идет поиск названия кодекса как уже было описано ранее. Если номеров статей нет во всем решении, то алгоритм завершает выполнение. Если же ответчиков нет, то поиск регулярных выражений проводится в тексте решения. И если номеров статей нет, то алгоритм завершает выполнение (рисунок 7).

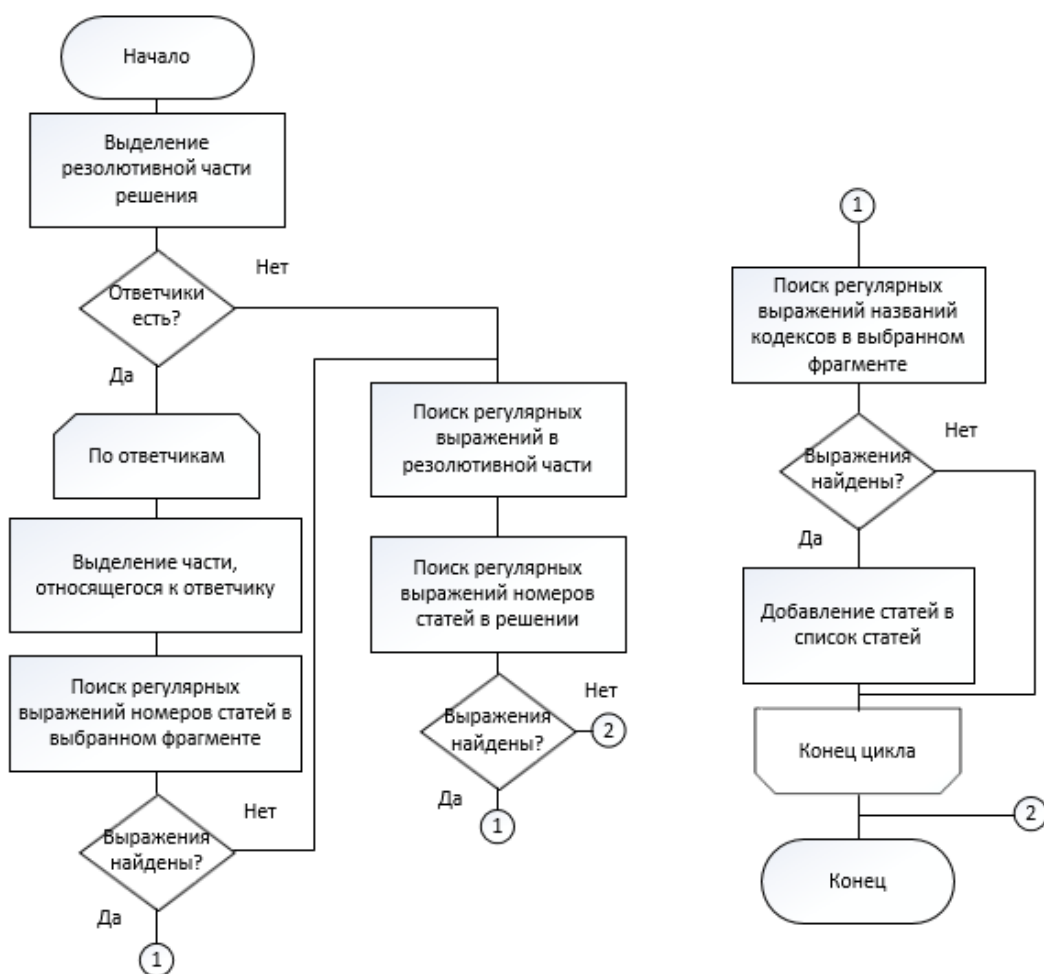


Рисунок 7 – Алгоритм для извлечения статей

Алгоритм для извлечения наказаний сперва выделяет часть текста с резолютивной частью. Если есть решения, то алгоритм идет по списку и выделяет из решения часть текста, относящуюся к текущему решению. Затем ищет в тексте заданные регулярные выражения наказаний, если выражение найдено, то в список наказаний добавляется наказание для текущего решения. Если искомые выражения не найдены, то алгоритм переходит к следующему и

поиск ведется во всем решении по делу. Если же решений нет, то алгоритм завершает работу (рисунок 8).

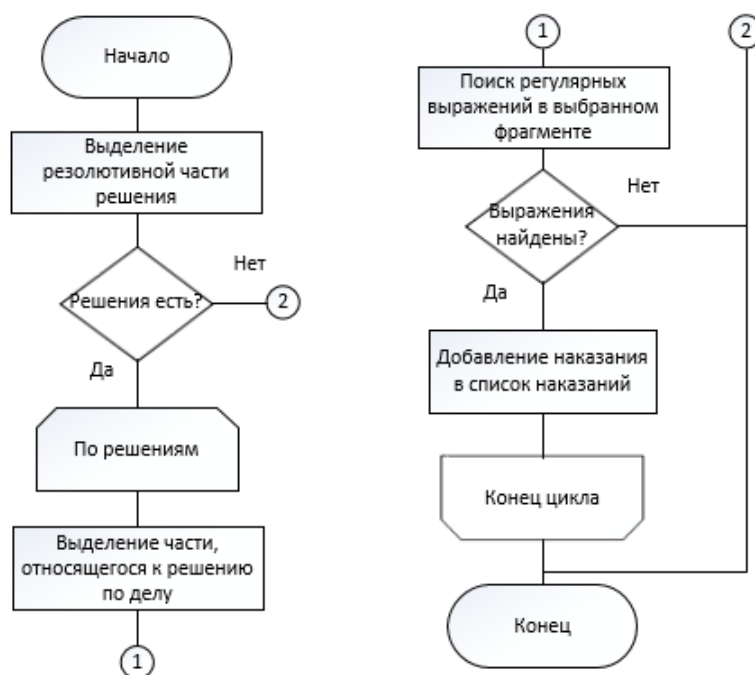


Рисунок 8 – Алгоритм для извлечения наказаний

Остальные атрибуты (имя судьи, вид и подвид дела, инстанция, дата суда, город и регион), находятся в xml тегах и очень просто извлекаются.

Извлеченные параметры записываются в xml файлы, каждый атрибут записывается на отдельную строку. Поскольку, атрибутов одного типа может быть несколько, то в файл записывается строка следующей структуры: «Название атрибута: Атрибут 1; ...; Атрибут N». Кроме того, у атрибута может быть несколько параметров (решение = имя ответчика + тип решения). То есть, «Атрибут 1» = «Параметр 1; ...; Параметр N».

2.4 Проектирование инструмента для наполнения базы данных

Наполнение БД довольно трудоемкое занятие, поскольку необходимо учитывать множество факторов. Так для каждой сущности необходимо выполнять проверку на дубликаты. Проверка выполняется с использованием расстояния Левенштейна, причем для атрибутов с именами нужно сравнивать не имя полностью, а отдельно фамилию и имя отчество. Поскольку в делах могут встречаться однофамильцы, поэтому в то время как фамилия может

немного отличаться, за счет разных падежей, то имя отчество должны проверяться на полное совпадение. Также имя и отчество может быть записано в виде инициалов или полное имя и отчество.

Также необходимо учитывать связи между различными сущностями. Так, например, нельзя сперва создать решение по судебному если еще этого дела еще нет, поскольку сущность «Решение» ссылается на идентификатор сущности «Дело». Кроме того, необходимо учитывать тип атрибута, так как БД требует точного их соблюдения.

У каждой сущности есть свой набор атрибутивной информации, включающий первичный и внешние ключи. У каждого экземпляра сущности есть свой уникальный ключ (он может состоять более чем из одного поля), который однозначно идентифицирует экземпляр. Внешние ключи, позволяют однозначно определить на какой экземпляр сущности 2 ссылается экземпляр сущности 1 [MS SQL Server 2012. Создание баз данных и разработка программ. Руководство для начинающих и профессионалов Бондарь Александр Григорьевич БХВ-Петербург, 2013 - Всего страниц: 608].

Упрощенный алгоритм записи в БД атрибутов судебного дела:

1. Первыми заполняются экземпляры сущностей, у которых нет внешних ключей (Адвокаты, Виды дел, Подвиды дел, Инстанции, Города, Судьи, Истцы).
2. Запись города, используя уникальный ключ региона.
3. Запись суда, используя уникальный ключ города.
4. Запись дела со всеми его атрибутами и внешними ссылками.
5. Записываются экземпляры сущность Юридические лица и одновременно заполняется промежуточная таблица юридических лиц, участвующих в деле.
6. Записываются экземпляры сущность Физические лица и одновременно заполняется промежуточная таблица физических лиц, участвующих в деле.

7. Затем, идет запись решений, вынесенных по делу. Причем, для каждого решения сперва записываются в соответствующие сущности ответчиков и типов вынесенного решения.

8. Далее для каждого решения записываются наказания, вынесенные по данному решению. Одновременно для каждого наказания записывается тип наказаний и тип меры наказания в соответствующие сущности.

9. Также для каждого решения записываются статьи, на которые оно ссылается. Перед записью статьи записывается название кодекса этой статьи в сущность, хранящую кодексы. После записи статьи в сущность добавляется запись в промежуточную таблицу статей решений.

Данный алгоритм выполняется для каждого файла с параметрами судебного дела.

2.5 Проектирование инструмента для анализа данных

В процессе использования алгоритмов кластеризации данных, была выявлена возможность применения результатов для более глубокого анализа данных. В качестве инструмента используются средства визуализации массивов данных, а именно координат точек на плоскости и метаданных к этим точкам. В качестве инструмента для визуализации может быть использован любой имеющий возможность интерпретировать матрицу данных в точечную диаграмму. В данной работе использовался онлайн (интернет) инструмент, называющийся `pyplot`, интегрируемый в язык Python. Также для работы с обработанными текстовыми решениями были использованы инструменты Tableau, Microsoft Excel, Open Office Calc. То есть данные сначала подготавливаются, далее проходят обработку через алгоритмы понижения размерности и кластеризации, и далее анализируется текстовая информация с учётом предыдущих манипуляций.

3 Разработка инструментов для обработки данных

3.1 Разработка инструмента для извлечения атрибутов судебного дела

Инструмент для извлечения атрибутов судебного дела реализован на языке Python с использованием его различных расширений, позволяющих работать над извлечением текстовой информации.

Основной проблемой при реализации инструмента было, то что у судебных дел нет единой структуры и имена ответчиков обезличены. Поэтому, для некоторых параметров приходилось составлять довольно большой список регулярных выражений. Причем, эти списки приходилось постоянно дополнять, поскольку решения по разным делам довольно сильно отличались. Таким образом, реализация инструмента проводилась итеративно.

В результате удалось получить следующие данные:

- номер дела;
- область;
- город;
- суд;
- вид дела;
- тип дела;
- инстанция;
- ФИО Судьи;
- пол судьи;
- дата вступления в силу;
- дата суда;
- вынесенное решение;
- название нормативного акта;
- статья;
- наказание;
- размер наказания;

- адвокат;
- ответчик;
- истец;
- юридические лица;
- физические лица.

Извлеченные атрибуты записываются в xml файлы (рисунок 9).

```

judge:Прохорова Людмила Петровна
region:Томская область
city:Томск
Advokat:Лодыгиной Е.В.
persony:Ясюкевич Е.И.;Истомина А.Г.;Лодыгиной Е.В.;Туров
С.А.;Председательствующий Л.П.;Прохорова Л.П.
Stat:Туров С.А.|ч. 1 ст. 111|Уголовный кодекс Российской
Федерации
court:Бакчарский районный суд (Томская область)
vid_dokumenta:приговор
podsud:Туров С.А.
urlica:ОГУ|Томский территориальный фонд ОМС;ОГУ|Томский
территориальный фонд обязательного медицинского страхования
number:№ 1-20/2011
etapd:Первая инстанция
vidpr:Уголовное
Nakaz:Туров С.А.|Взыскание|84951|рубли;Туров С.А.|Арест|2|годы
Reshenie:Туров С.А.|Виновен
Istec:Истомина А.Г.
Pol:ж
DateSuda:13.04.2011
DateVstVSily:2011-02-18

```

Рисунок 9 – Xml файл, хранящий извлеченные атрибуты

3.2 Разработка инструмента для наполнения базы данных

На основе результатов проведенного этапа проектирования был реализован инструмент для заполнения базы данных судопроизводства на языке C# в среде Visual Studio 2013. На входе инструмент получает набор Xml файлов (рисунок 9).

Основной проблемой разработки инструмента было правильно реализовать взаимосвязи между различными сущностями и предусмотреть все возможные варианты значений атрибутов. Так, имена могут записываться по-разному: фамилия имя отчество, фамилия инициалы имени и отчества, инициалы имени и отчества фамилия. Поэтому сперва имя разбивалось на

отдельные слова, затем определялось где фамилия, а где имя и отчество. После этого проводится сравнение фамилий расстоянием Левенштейна и отдельно сравнение имени и отчества на совпадение.

В результате работы инструмента была получена база данных наполненная атрибутами судебных дел. Все атрибуты были записаны в соответствующие поля сущностей.

3.3 Разработка инструмента для анализа данных

В результате использования алгоритмов кластеризации данных была выявлена возможность использования этих результатов для дальнейшего анализа, с помощью инструментов визуализации. Для этого результаты выполнения алгоритмов сохраняются в виде матрицы из трёх столбцов, в которых один столбец представляет собой значения ID судебных решений, а два других координаты на плоскости. Такие данные позволяют на выявленных кластерах более детально выяснять закономерности по каждому решению, либо по совокупности решений.

4 Анализ данных с помощью технологий DataMining

4.1 Основные типы задач, решаемые технологиями Data Mining

В данное время нет единого мнения какие именно задачи относятся к Data Mining, при этом многие авторитетные источники перечисляют следующие основные задачи:

- Классификация
- Кластеризация
- Ассоциация
- Последовательность
- Прогнозирование
- Определение отклонений и выбросов
- Оценивание
- Анализ связей
- Визуализация

Классификация (Classification)

Краткое описание. Наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных - классы; по этим признакам новый объект можно отнести к тому или иному классу.

Методы решения. Для решения задачи классификации могут использоваться методы: ближайшего соседа (Nearest Neighbor); k-ближайшего соседа (k-Nearest Neighbor); байесовские сети (Bayesian Networks); индукция деревьев решений; нейронные сети (neural networks).

Кластеризация (Clustering)

Кластеризация является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не predetermined. Результатом кластеризации является разбиение объектов на группы.

Пример метода решения задачи кластеризации: обучение "без учителя" особого вида нейронных сетей -самоорганизующихся карт Кохонена.

Ассоциация (Associations)

Краткое описание. В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных.

Отличие ассоциации от двух предыдущих задач Data Mining: поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.

Наиболее известный алгоритм решения задачи поиска ассоциативных правил - алгоритм Apriori.

Последовательность (Sequence), или последовательная ассоциация (sequential association)

Краткое описание. Последовательность позволяет найти временные закономерности между транзакциями. Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени (т.е. происходящими с некоторым определенным интервалом во времени). Другими словами, последовательность определяется высокой вероятностью цепочки связанных во времени событий. Фактически, ассоциация является частным случаем последовательности с временным шагом, равным нулю. Эту задачу Data Mining также называют задачей нахождения последовательных шаблонов (sequential pattern).

Правило последовательности: после события X через определенное время произойдет событие Y.

Пример. После покупки квартиры жильцы в 60% случаев в течение двух недель приобретают холодильник, а в течение двух месяцев в 50% случаев приобретается телевизор. Решение данной задачи широко применяется в маркетинге и менеджменте, например, при управлении циклом работы с клиентом (Customer Lifecycle Management).

Прогнозирование (Forecasting)

Краткое описание. В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей.

Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

Определение отклонений или выбросов (Deviation Detection), анализ отклонений или выбросов

Краткое описание. Цель решения данной задачи - обнаружение и анализ данных, наиболее отличающихся от общего множества данных, выявление так называемых нехарактерных шаблонов.

Оценивание (Estimation)

Задача оценивания сводится к предсказанию непрерывных значений признака.

Анализ связей (Link Analysis) - задача нахождения зависимостей в наборе данных.

Визуализация (Visualization, Graph Mining)

В результате визуализации создается графический образ анализируемых данных. Для решения задачи визуализации используются графические методы, показывающие наличие закономерностей в данных.

Пример методов визуализации - представление данных в 2D и 3D измерениях.

Подведение итогов (Summarization) - задача, цель которой - описание конкретных групп объектов из анализируемого набора данных.

4.2 Постановка задач

На этапе изучения предметной области, проектирования объектной модели судебных решений были выявлены следующие особенности текста в решениях судов:

- в каждом решении присутствуют текстовые маркеры разделяющие решения на части: вводная часть, описательная часть, мотивировочная часть, резолютивная часть;
- личные данные в некоторых решениях присутствуют в некоторых скрыты;
- данные судей всегда присутствуют;
- текст содержит достаточно подробную описательную часть

На основании перечисленных пунктов, а также на основании факта наличия большого количества решений можно сделать вывод о целесообразности использования методов Data Mining применительно к данным решений судов. Также на основании вышеперечисленного следует поставить основные направления задач:

- С помощью алгоритмов кластеризации провести поиск ранее неизвестных зависимостей и аномалий в судебных решениях.
- Создание механизма группировки судебных решений из ранее неструктурированных данных.
- Создание механизма по анализу сгруппированных судебных решений на основе методов кластеризации.
- Поиск отклонений в судебных решениях для дальнейшего использования.
- Поиск основных атрибутов влияющих на группировку в совокупности судебных решений.
- Сравнение алгоритмов обработки многомерных объектов на примере судебных решений.

4.3 Выбор алгоритмов для поставленных решения задач

Для определения возможности использования данных необходимо понять структуру данных и попытаться представить данные в числовом виде, пригодном для использования в различных алгоритмах, а также для возможности выполнения поставленных задач. Так как исследования проводятся фактически на строке с текстом, мы можем представить строку каждого решения как набор отдельных слов. В данном случае каждое слово выступает дополнительной мерой для объекта решения. К примеру, при описании какого-либо объекта используются различные атрибуты, и чем больше атрибутов, тем более отчётливо будет деление объектов на группы. В практическом применении очень легко представить на примере людей, люди - это объекты, на начальном этапе при отсутствии дополнительных атрибутов (дополнительных мер) весь набор объектов относится к одному классу “человек”, если ввести меру “пол”, то все объекты поделятся на два кластера соответственно “мужчины” и “женщины”, каждое дополнительное качество будет создавать новые группы, также различные комбинации атрибутов будут давать различную кластеризацию объектов.

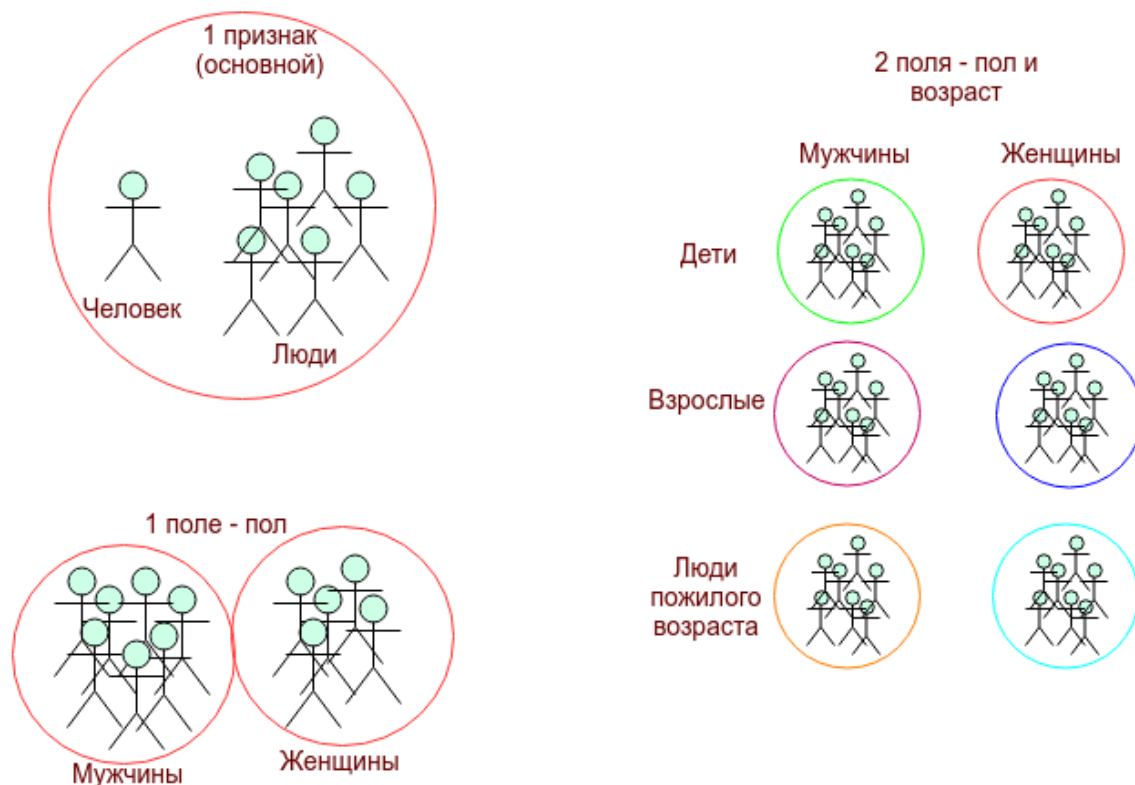


Рисунок 10 – Пример группирования одномерных и многомерных объектов

В нашем случае каждое слово влияет на положение объекта в многомерном пространстве, что в дальнейшем определяет его положение в том или ином кластере. То есть каждое слово - признак, что позволяет очень точно дифференцировать каждый объект в общей картине, и в то же самое время точно распределяют объекты по группам.

Далее возникает необходимость представления многомерных объектов в виде понятных к восприятию и применению данных, для этого необходимо каждый объект из многомерного пространства преобразовать в понятное и удобное для восприятия трёх или двумерное пространство, для чего появляется необходимость выбора соответствующего алгоритма, способного обработать многомерные данные. Многомерный объект несёт в себе большое количество полей, каждое из которых является мерой, или в математическом смысле каждое поле это есть координата в пространстве, таким образом наш объект является многомерным в многомерном пространстве. На приведённом рисунке

количества раз, это количество и есть индекс, пример можно видеть на рисунке 12 ниже.

суд	15
постановил	14
соответствие	7
изложить	6
отношение	6
надлежащий	6
верховный	5
состоять	5
ущерб	4
обоснованно	4
ООО	4
банковский	4

Рисунок 12 – Пример индексированного объекта

4.4 Интерпретация диаграмм

На приведённых выше диаграммах каждый знак “+” это объект решения. Каждый объект решения переданный на вход алгоритма влияет на своё положение и положение каждого другого решения, то есть на всю картину в целом. Присутствие в решении того или иного слова, воздействует на его принадлежность тому или иному кластеру, таким образом каждый элемент имеет значение во всей системе целиком и для каждого решения в отдельности.

Сходимость объектов.

В данной работе под сходимостью объектов понимается их взаимосвязь, схожесть относительно друг друга, именно в данной совокупности объектов, то есть два объекта могут быть сходимыми по ряду атрибутов (или полей), именно эти поля у объектов одинаковы, и различны по остальным атрибутам, чем больше атрибутов сходятся, тем сильнее сходны объекты друг с другом.

4.5 Подбор алгоритмов для решения выбранных задач.

Описание алгоритмов

При выборе алгоритмов для решения поставленных задач необходимо руководствоваться типами поставленных задач, имеющихся типов входных данных, возможность использования данных в тех или иных алгоритмах (возможность математического представления данных)

Критериями выбора алгоритма являются:

- обработка многомерных объектов
- быстрота выполнения
- эффективность использования ресурсов машины
- наглядность и точность результата

Для решения задач кластеризации с учётом вышеперечисленных условий подойдут алгоритмы принцип работы, которых заключён в снижении размерности объектов с сохранением сходимости этих объектов, и возможностью последующей группировкой, а именно алгоритмы Нелинейного Сокращения Размерности (Nonlinear dimensionality reduction). Это алгоритмы обучения. Алгоритм может узнать внутреннюю модель данных, которая может использоваться, чтобы отобразить пункты недоступные во время обучения с внедрением в процесс, часто называемый вневыборочным расширением.

К алгоритмам нелинейного сокращения размерности относятся следующие:

- Сопоставление Сэммона (Sammon mapping, иногда Проекция Сэммона).
- Самоорганизующиеся карты (Self-organizing map).
- Автокодировщики (Autoencoders).
- Диффузионные карты.
- Многочисленное выравнивание (Manifold alignment).
- Многомерное шкалирование (MDS) - классический МДС алгоритм.
- Алгоритм Isomap.
- Алгоритм LTSA (Local tangent space alignment - локальное выравнивание к касательному пространству).

- Алгоритм LLE (Locally-linear embedding) - Локальное линейное вложение.
- Локально-линейное вложение Гессе (Hessian LLE).
- Модифицированное локально-линейное вложение.
- Нелинейный метод главных компонент (Nonlinear Principal component analysis [PCA]).
- Топологически ограниченное изометрическое вложение.
- Алгоритм t-SNE - стохастическое вложение соседей с распределением Стьюдента.

Для работы с нашими данными были выбраны следующие алгоритмы:

Многомерное шкалирование(MDS) — метод анализа и визуализации данных с помощью расположения точек, соответствующих изучаемым (шкалируемым) объектам[8], в пространстве меньшей размерности чем пространство признаков объектов. Точки размещаются так, чтобы попарные расстояния между ними в новом пространстве как можно меньше отличались от эмпирически измеренных расстояний в пространстве признаков изучаемых объектов. Если элементы матрицы расстояний получены по интервальным шкалам, метод многомерного шкалирования называется метрическим. Когда шкалы являются порядковыми, метод многомерного шкалирования называется неметрическим. Мера различий расстояний в исходном и новом пространстве называется функцией стресса.

Области применения данного типа алгоритмов:

- Поиск скрытых переменных, объясняющих полученную из опыта структуру попарных расстояний между изучаемыми явлениями.
- Проверка гипотез о расположении изучаемых явлений в пространстве скрытых переменных.
- Сжатие полученного опытным путём массива данных путём использования небольшого числа скрытых переменных.
- Наглядное представление данных.

Алгоритм МДС стремится поместить каждый объект в N-мерном пространстве таким образом, что между объектами расстояния сохраняются.

Алгоритм LTSA - локальное выравнивание к касательному пространству (LTSA) представляет собой метод обучения многообразия, которое может эффективно свести нелинейное вложение из высокоразмерных данных в низкоразмерные координаты, а также может восстановить многомерные координаты из вложенных координат [9]. Данный алгоритм основан на интуиции, таким образом, что, когда мультиразмерные данные правильно развернуты, все касательные гиперплоскости к многообразию будут выстроены. Алгоритм начинает вычисления с K-ближайших соседей каждой точки. Он вычисляет касательное пространство в каждой точке путем вычисления d-первых главных компонент в каждой локальной окрестности, оптимизирует нахождение вложения, согласовывающего с касательными пространствами, при этом игнорирует информационные метки, перемещаемых выборок данных, и, таким образом, не может быть использован для классификации непосредственно.

Алгоритм LLE (Locally-linear embedding) - локальное линейное вложение. Локально линейное вложение (LLE) ищет нижнюю границу одномерной проекции данных, сохраняющее расстояния в пределах ближайших соседей. Его можно рассматривать как серию локальных PCA с глобальным сравнением для нахождения наилучшего нелинейного вложения [10].

Алгоритм t-SNE (t-distributed stochastic neighbor embedding) - стохастический алгоритм с распределением Стюдента и внедрением соседа [11].

Все вышеперечисленные алгоритмы осуществляют обработку многомерных данных с преобразованием их в двумерные и представлением на плоской (двумерной) диаграмме. Перечисленные выше четыре алгоритма оказались наиболее подходящими для работы с той конструкцией данных которая имеется в данной работе, эти четыре алгоритма оказались наиболее

продуманными с точки зрения получаемого результата, и наиболее эффективными для обработки данных. Стоит отметить что список алгоритмов, осуществляющих нелинейное сокращение размерности не ограничивается вышеуказанным в данной работе, однако именно перечисленные в данной работе алгоритмы наиболее применимы к существующим в текущей работе данным. Также стоит отметить что многие существующие алгоритмы являются или комбинацией алгоритмов, или различными модификациями существующих.

4.6 Основные этапы DataMining

1. Подготовка данных.
2. Индексация данных.
3. Нормализация данных для использования в алгоритмах обработки.
4. Выполнение алгоритмов и получение результатов.

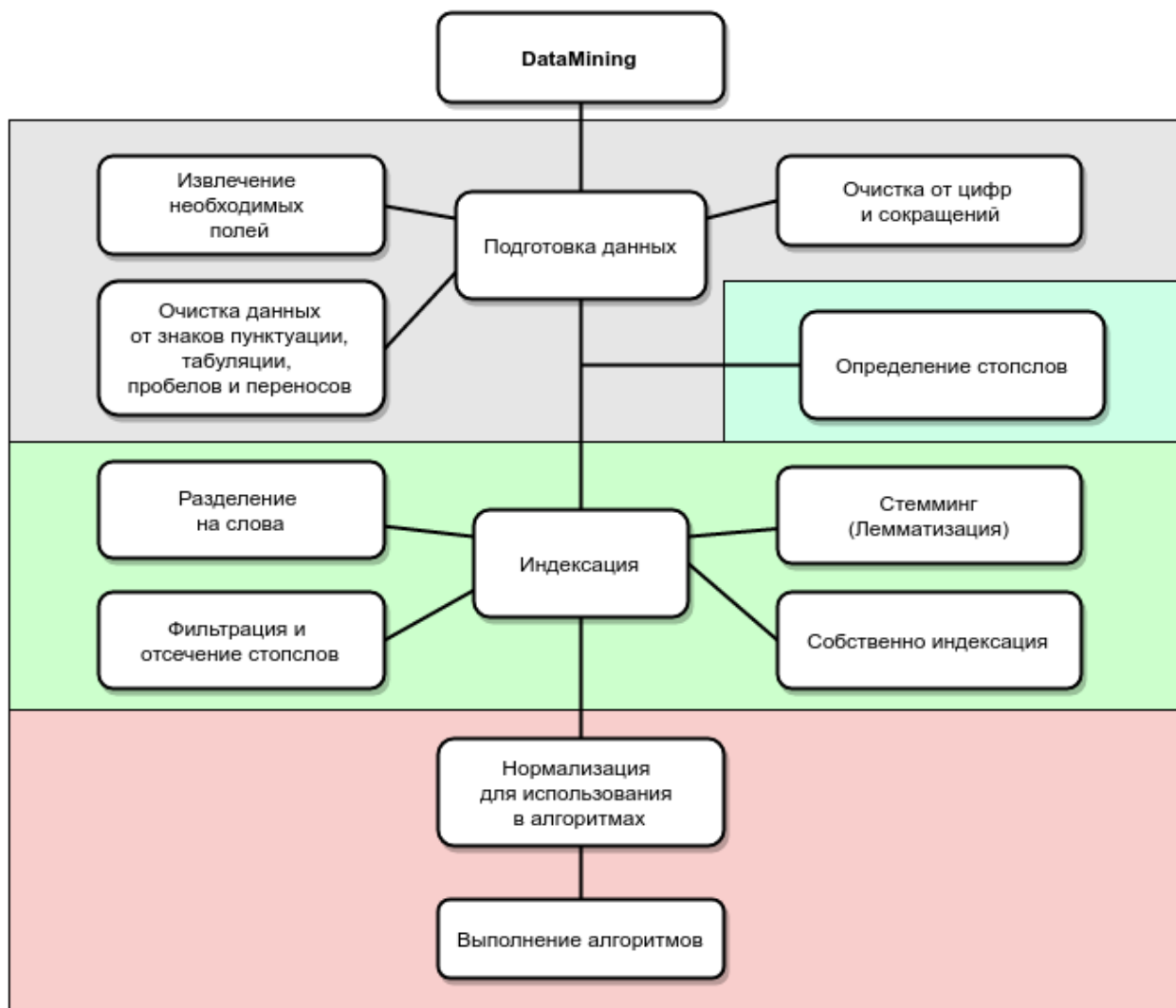


Рисунок 13 – Древоподобная схема последовательности этапов в использовании технологии Data Mining

1. Подготовка данных. На данном этапе подразумевается наличие данных полученных в результате парсинга сайтов и/или документов содержащих решения судов. Подготовка данных заключается в очистке текста от знаков препинания (.,!?:;), знаков табуляции, пробелов, символов переноса строк, устранение сокращений(кг, кв.м., м.куб., руб., ст., ч., п. и так далее) и ненужных цифр(даты, суммы, номера)

2. Индексацию данных можно условно разделить на несколько частей: разделение на слова, фильтрация и извлечение стопслов, стемминг или лемматизация, индексация.

Разделение на слова — это процедура дробления строки текста (в данной работе дробление строки текста нормативного акта суда) для получения массива данных в котором в качестве единичного элемента является слово из текста, так как все слова в тексте разделены символом пробела, то соответственно дробление строки осуществляется посредством символа пробела.

Фильтрация слов и извлечение стопслов — этап в котором из массива слов удаляются слова не представляющие значения для анализа текста в будущем. Стопслова — это слова, добавляющие оттенок в текст, дополняющие слова, которые не несут смысловой нагрузки по отдельности и являются вспомогательными, к ним же относятся и предлоги, частицы, союзы.

Стемминг или лемматизация слов — это приведение слов к нормальной форме. По сути эти два подхода отличаются друг от друга, но цель у них одна и та же. Стемминг представляет собой алгоритм отщепления от слова суффиксов, префиксов и окончаний, в результате чего получается фактически обрезанная форма слова, сохранившая только основную часть, данный подход проще в реализации, но бывает и ошибочным, так как дробление может быть некорректным [12].

Лемматизация слова — это алгоритм целью которого является приведение слова к нормальной форме в единственном числе именительном падеже. В реализации данный алгоритм сложнее, но результат более пригодный для использования в дальнейшей деятельности.

Индексация данных в текущей работе представляет собой счётчик слов в тексте, каждое слово, встретившееся один раз будет иметь индекс равный одному, соответственно сколько раз слово встречается в тексте такой индекс оно и будет иметь.

3. Для использования данных в различных алгоритмах необходимо данные привести в подходящую форму, просто текст или набор слов не всегда может подойти для выполнения операций, в первую очередь необходимо иметь модель обработки данных и понимание подходит ли набор данных для этой

модели или нет, если нет, то надо либо отказаться от применения данного алгоритма в связи с невозможностью использования, либо получить из существующего набора данных дополнительные данные.

4. Выполнение алгоритма это собственно запуск программы, которая реализует нужный алгоритм

4.7 Выбор оптимального алгоритма для будущего использования

4.7.1 Проверка каждого алгоритма непосредственно в использовании

Для использования алгоритмов нелинейного сокращения размерности была сделана выборка части судебных решений из общего числа в имеющейся базе данных решений данной работы, для этого все решения были перемешаны таким образом, что в каждом алгоритме будут фигурировать одинаковые входные данные и при этом в выборку попадут все типы решений исходя из теории вероятностей, таким образом была выбрана только часть решений, однако эта часть будет отражать картину аналогичную если бы были взяты все имеющиеся решения, была произведена выборка в количестве 4000 решений из 93000.

При использовании выбранных алгоритмов были получены графические результаты в значительной степени отличающиеся друг от друга, проведём анализ данных результатов.

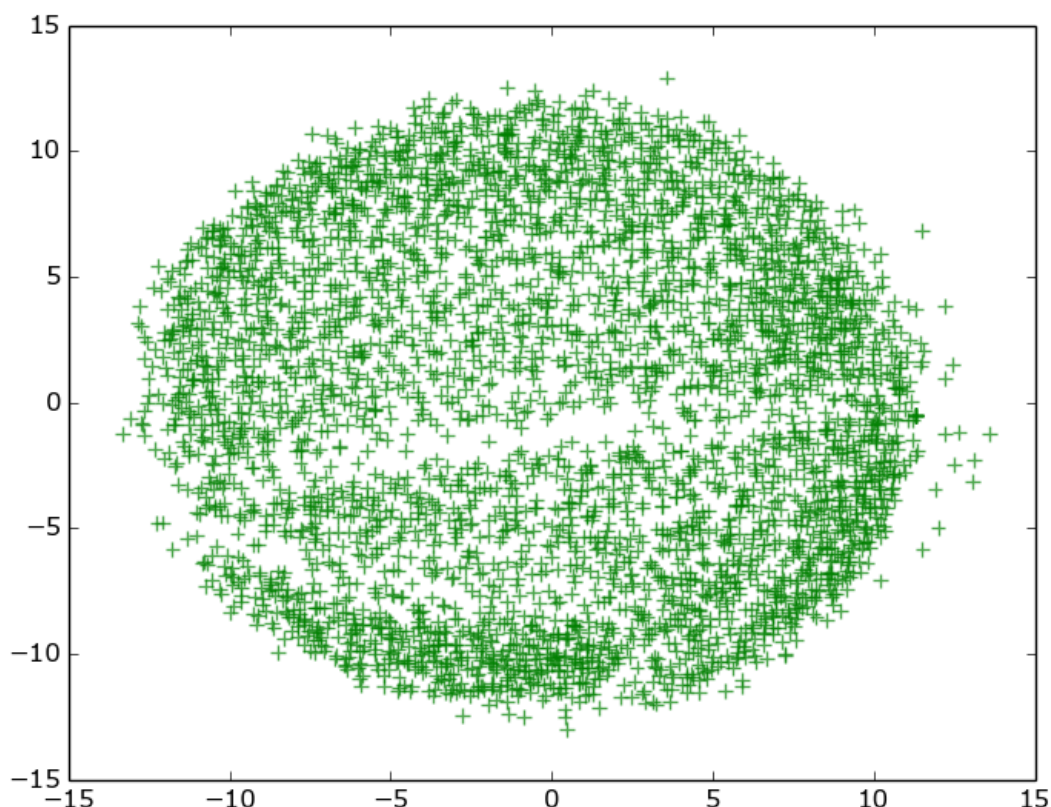


Рисунок 14 - результат выполнения алгоритма MDS

На рисунке 14 мы видим графическое представление совокупности судебных решений, полученное в результате обработки многомерных объектов судебных решений, преобразованные в двумерные в виде точек на плоскости в двумерном евклидовом пространстве с помощью алгоритма MDS, координаты данных точек представляют собой относительные единицы в пространстве, в графическом виде данная диаграмма даёт пока мало информации. Необходимо выделить кластеры более явно, для этого в данной работе обращаемся к алгоритму k-means (метод k-средних), это метод кластеризации точек разработанный в 50-х годах 20-го столетия математиком Гуго Штейнгаузом и почти одновременно Стюартом Ллойдом. Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров. В результате обработки совокупности

точек на плоскости методом k-средних, которые были предварительно обработаны алгоритмом MDS можно видеть на рисунке 14-а ниже.

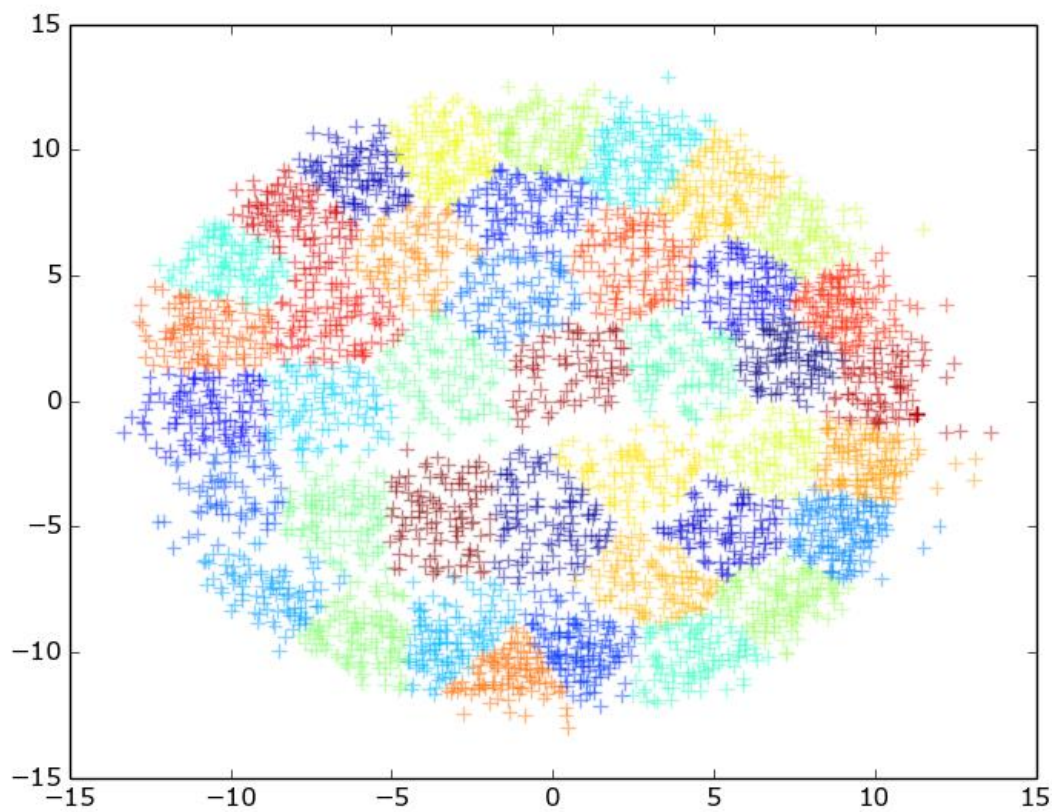


Рисунок 14-а - Кластеризация судебных решений представленных в виде точек на плоскости, каждая точка представлена в виде знака “+”

В данном случае кластеризация была представлено 40 кластерами, хотя в данной работе проводилась также кластеризация 100 кластерами.

Далее для обработки многомерных объектов решений судов, был использован алгоритм LTSA. На рисунке 15 мы видим графическое представление совокупности судебных решений, полученное в результате обработки многомерных объектов судебных решений этим алгоритмом, преобразованные в двумерные объекты, в виде точек на плоскости в двумерном евклидовом пространстве, также как и в предыдущем примере, представляют собой относительные единицы в пространстве.

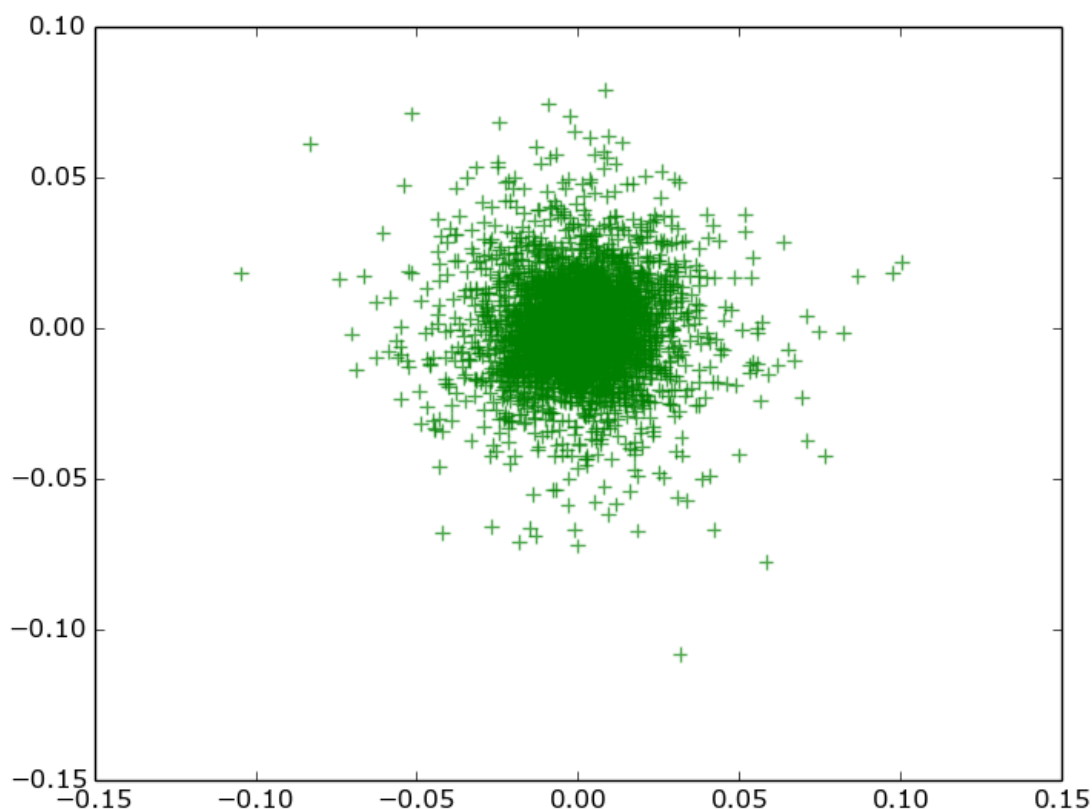


Рисунок 15 - результат выполнения алгоритма LTSA

Также, как и в предыдущем алгоритме, сложно увидеть явные группировки объектов на данной диаграмме. Поэтому также, как и для предыдущего варианта необходимо явно провести кластеризацию объектов, для чего воспользуемся алгоритмом k-means (метод k-средних), в данном случае также возьмём количество кластеров равное 40. Результат обработки совокупности точек на плоскости методом k-средних, которые были предварительно обработаны алгоритмом LTSA можно видеть на рисунке 15-а.

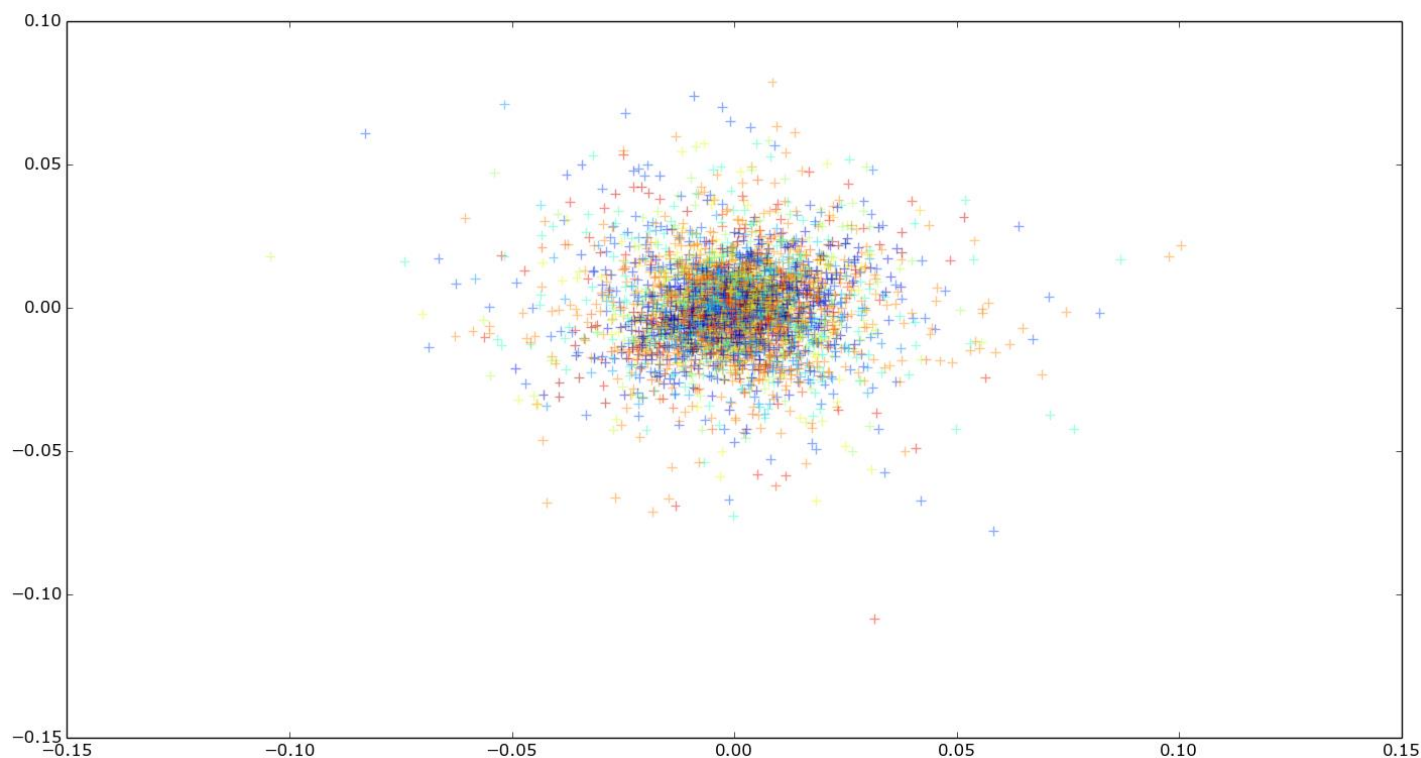


Рисунок 15-а – Кластеризация судебных решений представленных в виде точек на плоскости, каждая точка представлена в виде знака “+”

На рисунке 15-а весьма сложно определить кластеры по судебным решениям, так как визуально можно лишь выделять отдельные решения, но не группы, в следствии чего визуальные инструменты для объектов решений обработанных с помощью алгоритма LTSE не являются в данном случае показательными.

Для анализа объектов судебных решений перейдём к результатам алгоритма LLE, который также позволяет снижать размерность многомерных объектов, до двумерных объектов, на рисунке 16 можно увидеть графический результат выполнения данного алгоритма.

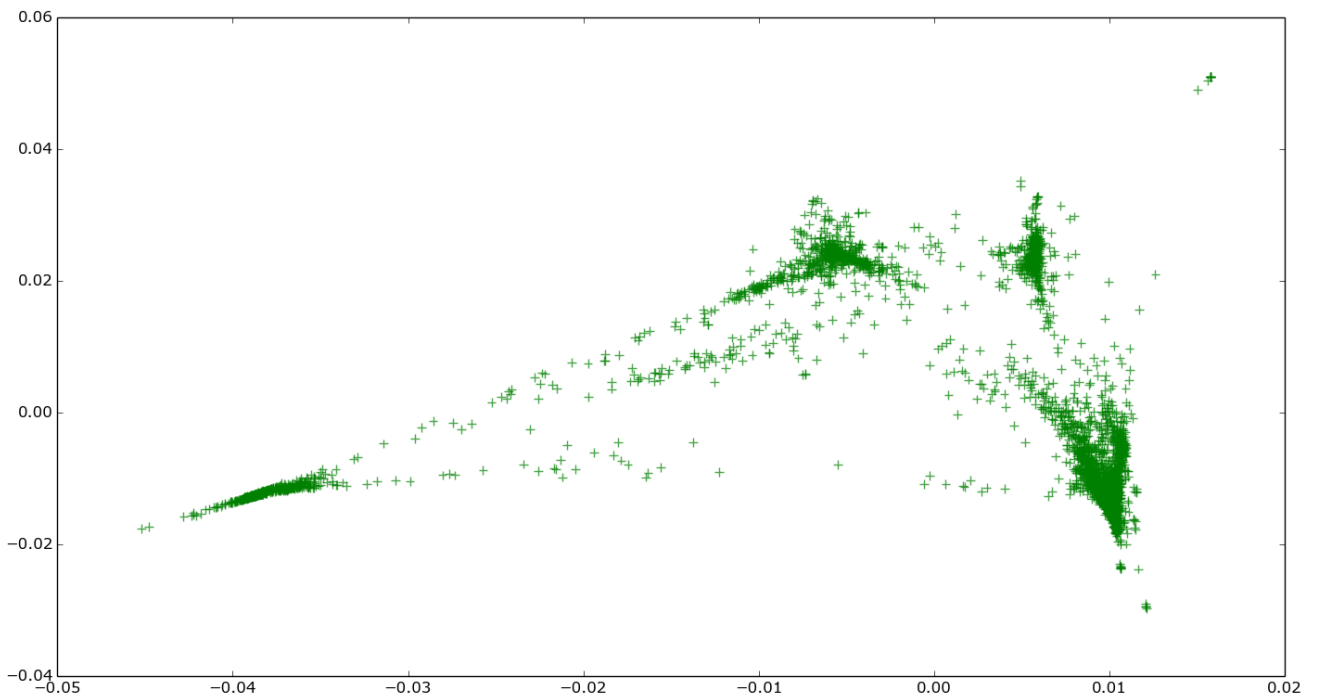


Рисунок 16 - результат выполнения алгоритма LLE

Визуально результат выполнения алгоритма LLE более показателен в плане группировки объектов решений, можно видеть явные совокупности объектов, однако на данной диаграмме также необходимо провести явную кластеризацию с областей разными цветами для наглядности результата. И здесь проведём это с помощью алгоритма k-means, также разделив на 40 кластеров. Результат выполнения можно видеть на рисунке 16-а.

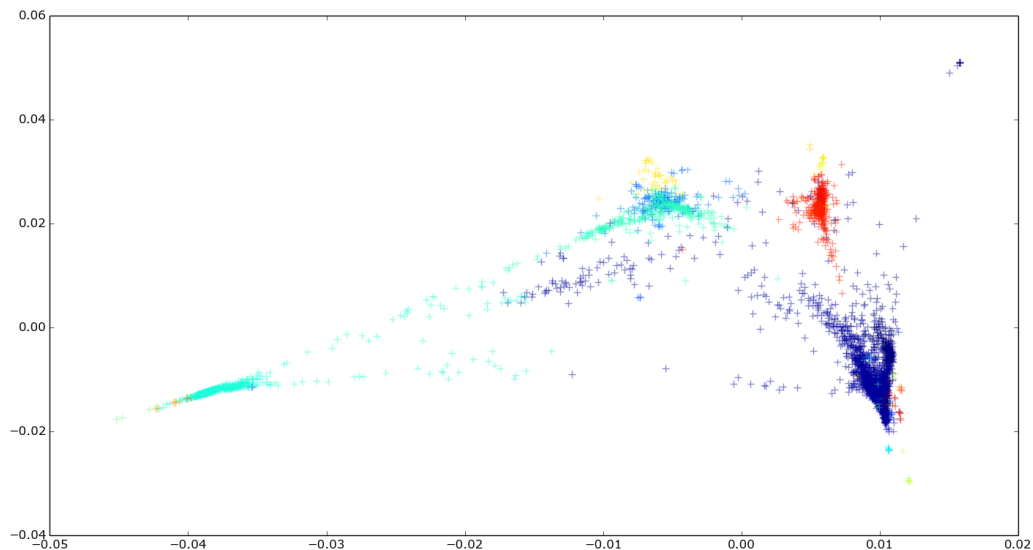


Рисунок 16-а – Кластеризация судебных решений представленных в виде точек на плоскости, каждая точка представлена в виде знака “+”

Цветная кластеризация на рисунке 15-а явно отличается от кластеризации на рисунке 15, хотя некоторые области всё же остались по-прежнему такими же.

Перейдём к следующему алгоритму и результатам его выполнения. Данным алгоритмом является t-SNE, он также понижает размерность объектов с сохранением их сходимости. Результат выполнения алгоритма t-SNE можно видеть на рисунке 17.

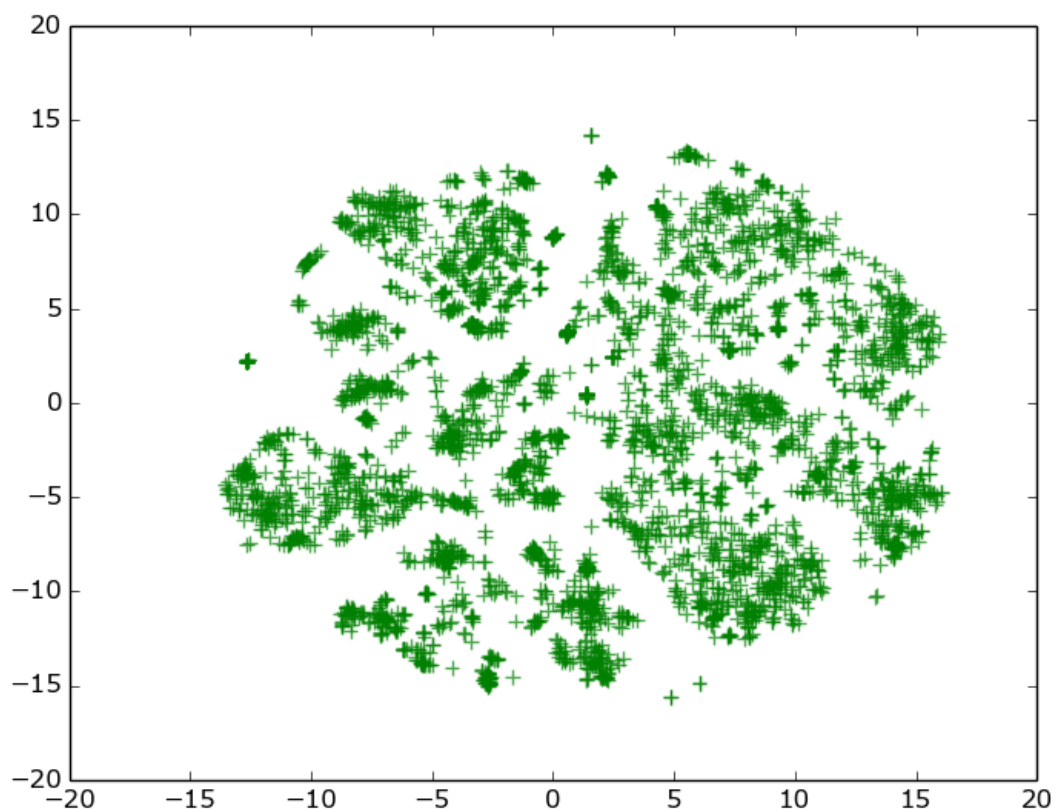


Рисунок 17 - результат выполнения алгоритма t-SNE

На данной диаграмме визуально очень хорошо выделены группы объектов судебных решений, и совокупности объектов очень хорошо дифференцированы, однако и в этом случае дополнительно обработать результаты выделив конкретные области и задав конкретное количество классов, для этого используем алгоритм k-means. В результате обработки получаем диаграмму, отображённую на рисунке 17-а

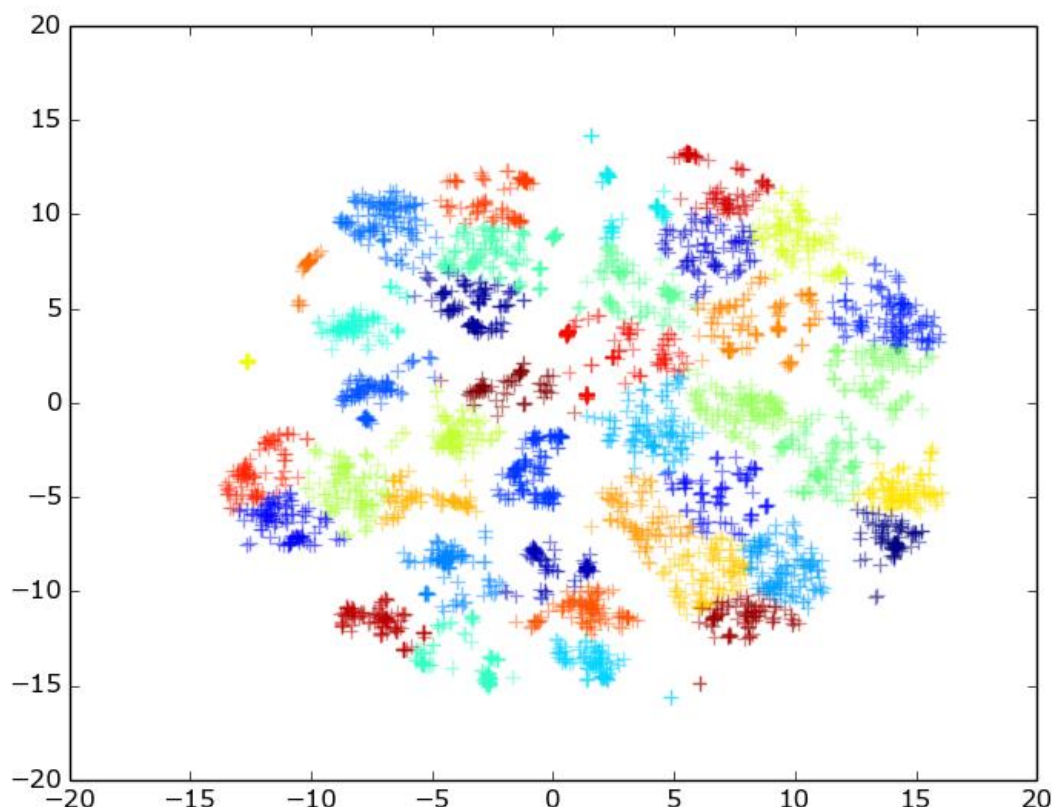


Рисунок 17-а – Кластеризация судебных решений представленных в виде точек на плоскости, каждая точка представлена в виде знака “+”

Результат на рисунке 17-а даёт более наглядную картинку по распределению решений и их сходимости

Также в процессе использования алгоритмов были выявлены данные о производительности этих алгоритмов:

- MDS - 4000 решений, время выполнения 25 минут.
- LTSA - 4000 решений, время выполнения 19 минут.
- LLE - 4000 решений, время выполнения 7 минут.
- t-SNE - 4000 решений, время выполнения 3 минуты.

Данные алгоритмы были использованы в одинаковых условиях на машине Lenovo G580, Core i3, RAM 8Gb, Ubuntu 14.04 operation system.

На приведённых выше диаграммах каждый знак “+” это объект решения. Каждый объект решения переданный на вход алгоритма влияет на своё положение и положение каждого другого решения, то есть на всю картину в

целом. Присутствие в решении того или иного слова, воздействует на его принадлежность тому или иному кластеру, таким образом каждый элемент имеет значение во всей системе целиком и для каждого решения в отдельности.

4.7.2 Анализ решений после обработки данных алгоритмами Нелинейного Сокращения Размерности и применения метода К-средних.

После полученных данных о сходимости объектов судебных решений необходимо провести их проверку, а именно наложив данные о сходимости(номера кластеров и возможно координаты точек полученные с помощью алгоритмов нелинейного сокращения размерности) и атрибуты судебных решений полученные ранее с помощью инструментов парсинга, в итоге станет ясно по каким атрибутам группируются судебные решения. Каждый объект судебного решения имеет уникальный идентификационный номер, что позволяет обращаться к каждому решению напрямую, а также сохранять уникальность решений. Таким образом

Для алгоритма LLE удалось обнаружить что происходит явная группировка решений по типу решения и по кодексу, это можно видеть на примере рисунка 18.

2408	34	126	Номер кластера	Административное	12.27 ч.2
2409	34	126		Административное	12.24 ч.1
2410	33	1317661200	419535312	Гражданское	Прочие исковые дела
2411	33	1317661200	419623832	Гражданское	Прочие исковые дела
2412	32	1299261600	103072711	Гражданское	Другие жилищные споры
2413	32	1336669200	104671066	Гражданское	Другие жилищные споры
2414	32	1339693200	105213345	Гражданское	Другие социальные споры
2415	32	1345482000	106925996	Гражданское	Другие жилищные споры
2416	32	1304442000	104073760	Гражданское	Другие жилищные споры
2417	32	1352307600	107105321	Гражданское	Другие жилищные споры
2418	32	1330362000	103071158	Гражданское	Другие жилищные споры
2419	32	1365440400	105074393	Гражданское	Другие жилищные споры
2420	32	1356109200	105074395	Гражданское	Другие жилищные споры
2421	32	1359392400	107265643	Гражданское	Другие жилищные споры
2422	32	1334595600	104671101	Гражданское	Другие жилищные споры
2423	32	1288112400	103073500	Гражданское	Другие жилищные споры
2424	31	1340643600	105449094	Административное	18.8 ч.1
2425	31	1315155600	101828470	Административное	18.8 ч.1
2426	31	1358269200	400047216	Административное	
2427	31	1312304400	101828481	Административное	18.8 ч.1
2428	31	1352739600	417636375	Административное	
2429	31	1345482000	106437989	Административное	18.8 ч.1
2430	31	1361206800	417601596	Административное	

Рисунок 18 – Группировка судебных решений по классам с соответствующими группировками по полям, после обработки алгоритмом LLE.

334	92	12689		158 ч.2 п.г; 158 ч.2 п.в	
335	92	12644	Номер кластера	158 ч.2 п.в	Обвинительный приговор
336	92	1255626000	102790962 Уголовное	158 ч.2 п.б	
337	92	1339002000	105109440 Уголовное		Обвинительный приговор
338	92	1327510800	100479237 Уголовное	158 ч.2 п.в	Обвинительный приговор
339	92	1338742800	106360204 Уголовное	161 ч.1	Обвинительный приговор
340	92	1285606800	101707017 Уголовное		228.2
341	92	1325782800	100378079 Уголовное	158 ч.3 п.а	Обвинительный приговор
342	91	1363885200	428848613 Административное	20.1 ч.2	Вынесено постановление о назначении административного наказания
343	91	1366909200	428819907 Административное	20.1 ч.1	Вынесено постановление о назначении административного наказания
344	91	1363366800	428848616 Административное	20.1 ч.2	Вынесено постановление о назначении административного наказания
345	91	1365354000	425281539 Административное	20.1 ч.2	Вынесено постановление о назначении административного наказания
346	90	1363107600	429578518 Административное	18.8 ч.1	Вынесено постановление о назначении административного наказания
347	90	1332349200	104326247 Административное	18.8 ч.1	Назначено административное наказание
348	90	1297706400	103663982 Административное		
349	90	1282496400	400920944 Административное	18.8 ч.1	Вынесено постановление о назначении административного наказания
350	90	1340643600	105524690 Административное	18.8 ч.1	Назначено административное наказание
351	90	1337187600	104878327 Административное	18.8 ч.1	Назначено административное наказание
352	90	1337014800	429596752 Административное	12.24 ч.2	Вынесено постановление о назначении административного наказания
353	89	1319734800	101828731 Гражданское	Иски о взыскании платы за жилую площадь и коммунальные платежи, тепло и электроэнергию	
354	89	1319734800	101828729 Гражданское	Иски о взыскании платы за жилую площадь и коммунальные платежи, тепло и электроэнергию	
355	88	1337014800	104709910 Гражданское		
356	88	1300989600	406814395 Гражданское	Прочие исковые дела	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
357	88	1303837200	102400726 Гражданское	Трудовые споры (независимо от Ф	Дело прекращено
358	87	1282582800	101849036 Административное	20.4 ч.1	Назначено административное наказание
359	87	1335286800	106065857 Административное	20.25 ч.1	Прекращено производство по делу по основаниям ст. 24.5 Кодекса
360	87	1359046800	419235565 Административное	12.24 ч.1	Вынесено постановление о назначении административного наказания
361	87	1320685200	101298423 Административное	11.8.1	Назначено административное наказание
362	87	1320080400	100398860 Административное	18.8 ч.1	Назначено административное наказание
363	87	1354035600	107144478 Административное	12.8 ч.1	Назначено административное наказание
364	87	1318266000	415161664 Административное	19.15 ч.1	Отменено с прекращением производства
365	87	1349024400	106786921 Административное	12.15 ч.1	Назначено административное наказание
366	87	1271610000	101706677 Административное	14.2; 14.1 ч.1	Назначено административное наказание
367	87	1335373200	104876749 Административное	14.16 ч.3	Назначено административное наказание
368	87	1364490000	427612468 Административное	9.1 ч.3	Вынесено постановление о назначении административного наказания

Рисунок 19 – Группировка судебных решений по классам с соответствующими группировками по полям, после обработки алгоритмом LLE.

На рисунке 18 можно увидеть, что классы 34, 33, 32 и 31 явно сходятся не только по номерам классов, а также по дополнительным полям Кодекс, Тип решения, на рисунке 19 также видны явные совпадения группировок данных и полученных классов по решениям судов, что свидетельствует о применимости алгоритма LLE для кластеризации текстовой информации.

Также подобные результаты получены и для всех остальных алгоритмов использованных для кластеризации объектов судебных решений. Результаты можно видеть на рисунках 20, 21, 22, 23.

14	99	1361813300	1069339890	Уголовное	158 ч.2 п.б; 30 ч.3. ; 158 ч.2 п.б	Вынесен ПРИГ ОБОР
15	99	13448	1069339890	Уголовное		Обв
16	99	1348765200	1069339890	Уголовное		Обв
17	99	1353517200	107203027	Уголовное	69 ч.5; 111 ч.4	Оставлено без изменения
18	99	1279040400	103074901	Уголовное	161 ч.2 п.г	
19	99	1346000400	107205877	Уголовное	116 ч.1; 119 ч.1	Обвинительный приговор изменен без изменения квалификации
20	99	1297965600	102059792	Уголовное	161 ч.1	
21	99	1330880400	103669138	Уголовное	161 ч.1	Оставлено без изменения
22	99	1287939600	101458987	Уголовное	161 ч.1; 162 ч.1; 119 ч.1	
23	99	1340211600	105422811	Уголовное	228 ч.1	Обвинительный приговор
24	99	1304355600	104032572	Уголовное	158 ч.2 п.в	Обвинительный приговор
25	99	1350579600	107071823	Уголовное	111 ч.4	Обвинительный приговор
26	99	1287594000	100402468	Уголовное	228 ч.2	
27	99	1288634400	100598738	Уголовное	158 ч.3 п.а	
28	99	1338483600	105358841	Уголовное	158 ч.2 п.а; 158 ч.2 п.б	Обвинительный приговор
29	99	1354726800	107238569	Уголовное	111 ч.1	
30	99	1331398800	105109442	Уголовное		Обвинительный приговор
31	99	1364317200	107302941	Уголовное		Обвинительный приговор
32	99	1342112400	105975411	Уголовное	228 ч.2	Обвинительный приговор
33	99	1278954000	105035559	Уголовное		
34	99	1290967200	100438329	Уголовное	158 ч.2 п.а; 158 ч.2 п.б	
35	99	1295373600	101832327	Уголовное	105 ч.1	Обвинительный приговор
36	99	1303664400	103075139	Уголовное	158 ч.2 п.в	
37	99	1337187600	104831180	Уголовное	70 ч.1; 74 ч.4; 112 ч.2 п.в	Оставлено без изменения
38	98	1280768400	1004015361	Гражданское	Иски о взыскании сумм по договору займа	
39	98	1293732000	105594532	Гражданское	Споры из нарушений пенсионного законодательства	Иск (заявление) удовлетворен (в том числе частично)
40	98	1365354000	428923900	Гражданское	Прочие искивые дела	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
41	98	1341564000	103071213	Гражданское		Иск (заявление) удовлетворен (в том числе частично)

Рисунок 20 – Группировка судебных решений по классам с соответствующими группировками по полям, после обработки алгоритмом MDS.

212	93	132880	1069339890	Уголовное	О возмещении ущерба от незаконных действий органов прокуратуры и суда	
213	93	134638	1069339890	Уголовное	18.15 ч.1	Оставлено без из
214	93	1338483600	102059792	Административно	19.5 ч.12	
215	93	1357750800	107241056	Административно	12.26 ч.1	Оставлено без изменения
216	93	1347382800	106633440	Административно	20.4 ч.1	Отменено с прекращением производства
217	93	1328202000	101298189	Административно	20.4 ч.3	
218	93	1338397200	104793475	Административно		Оставлено без изменения
219	93	1350234000	106919631	Административно	12.15 ч.4	Оставлено без изменения
220	93	1366995600	417796429	Административно	19.5 ч.1	Оставлено без изменения
221	93	1274115600	100402617	Уголовное	204 ч.3	
222	93	1345482000	106316684	Административно	12.8 ч.1	Оставлено без изменения
223	93	1330880400	102790791	Административно	12.8 ч.1	
224	93	1336669200	105115328	Административно	12.26 ч.1	Оставлено без изменения
225	93	1295892000	101298489	Административно	20.4 ч.1	Назначено административное наказание
226	93	1352134800	107077348	Административно	8.32 ч.1	
227	93	1307552400	103446152	Административно	12.8 ч.1	Оставлено без изменения
228	93	1303146000	101298436	Административно	20.4 ч.1	Назначено административное наказание
229	93	1298224800	101850069	Уголовное	256 ч.1	
230	92	1349024400	108886370	Гражданское	Другие жилищные споры	Иск (заявление) удовлетворен (в том числе частично)
231	92	1342976400	105942288	Гражданское	Прочие искивые дела	Иск (заявление) удовлетворен (в том числе частично)
232	92	1338483600	105968754	Гражданское	Другие споры, связанные с земл	Отказано
233	92	1359392400	107280757	Гражданское	О возмещении ущерба от незаконных действий (безде	Иск (заявление) удовлетворен (в том числе частично)
234	92	1342717200	106232986	Гражданское	Прочие искивые дела	Иск (заявление) удовлетворен (в том числе частично)
235	92	1313686800	103666553	Гражданское	Прочие искивые дела	Иск (заявление) удовлетворен (в том числе частично)
236	92	1286989200	101831142	Гражданское	Прочие искивые дела	Иск (заявление) удовлетворен (в том числе частично)
237	92	1336582800	104876793	Гражданское	Впервые предъявленные иски о возмещении вреда за	Иск (заявление) удовлетворен (в том числе частично)
238	92	1546279200	107308105	Гражданское	Жалобы на неправомерные действия (бездействие) р	Отказано
239	92	1356541200	424476392	Гражданское	Прочие искивые дела	Иск (заявление, жалоба) УДОВЛЕТВОРЕН ЧАСТИЧНО
240	92	1324314000	103665361	Гражданское	Иски о взыскании сумм по договору займа	
241	92	1298916000	101829908	Гражданское	Прочие искивые дела	Иск (заявление) удовлетворен (в том числе частично)
242	92	1343754000	106034650	Гражданское	Прочие искивые дела	Иск (заявление) удовлетворен (в том числе частично)
243	92	1354381200	104694989	Гражданское	Прочие искивые дела	

Рисунок 21 – Группировка судебных решений по классам с соответствующими группировками по полям, после обработки алгоритмом MDS.

883	Гражданское	Споры из нарушений пенсионного законодательства иски физ. лиц к Пенсионному фонду РФ	Иск (заявление) удовлетворен (в том числе частично)
175	Гражданское	Споры из нарушений пенсионного законодательства иски физ. лиц к Пенсионному фонду РФ	Иск (заявление) удовлетворен (в том числе частично)
182	Гражданское	Споры из нарушений пенсионного законодательства иски физ. лиц к Пенсионному фонду РФ	Иск (заявление) удовлетворен (в том числе частично)
251	Гражданское	Споры из нар. пенс. зак. - иски физ.лиц к Пенсионному фонду РФ	Иск (заявление, жалоба) УДОВЛЕТВОРЕН ЧАСТИЧНО
004	Гражданское	Споры из нарушений пенсионного законодательства иски физ. лиц к Пенсионному фонду РФ	Иск (заявление) удовлетворен (в том числе частично)
350	Гражданское	Споры из нарушений пенсионного законодательства иски физ. лиц к Пенсионному фонду РФ	Иск (заявление) удовлетворен (в том числе частично)
332	Гражданское	Споры из нарушений пенсионного законодательства иски физ. лиц к Пенсионному фонду РФ	Иск (заявление) удовлетворен (в том числе частично)
297	Гражданское	Споры из нарушений пенсионного законодательства иски физ. лиц к Пенсионному фонду РФ	Иск (заявление) удовлетворен (в том числе частично)
808	Гражданское	Прочие исковые дела	Отказано
607	Гражданское	Споры из нарушений пенсионного законодательства иски физ. лиц к Пенсионному фонду РФ	Иск (заявление) удовлетворен (в том числе частично)
651	Гражданское		
984	Гражданское	Споры из нар. пенс. зак. - иски физ.лиц к Пенсионному фонду РФ	ОТКАЗАНО в удовлетворении иска (заявлении, жалобы)
256	Гражданское	Споры из нарушений пенсионного законодательства иски физ. лиц к Пенсионному фонду РФ	Иск (заявление) удовлетворен (в том числе частично)
449	Гражданское	Споры из нарушений пенсионного законодательства иски физ. лиц к Пенсионному фонду РФ	Иск (заявление) удовлетворен (в том числе частично)
293	Гражданское	Споры из нар. пенс. зак. - иски физ.лиц к Пенсионному фонду РФ	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
052	Гражданское	Споры из нарушений пенсионного законодательства иски физ. лиц к Пенсионному фонду РФ	Иск (заявление) удовлетворен (в том числе частично)
998	Гражданское	Прочие исковые дела	Иск (заявление) удовлетворен (в том числе частично)
492	Гражданское	Прочие исковые дела	Иск (заявление) удовлетворен (в том числе частично)
793	Гражданское	Другие жилищные споры	Иск (заявление) удовлетворен (в том числе частично)
831	Гражданское	Споры о праве собственности на землю	Иск (заявление) удовлетворен (в том числе частично)
043	Гражданское	Другие жилищные споры	Иск (заявление) удовлетворен (в том числе частично)
922	Гражданское	Споры, связанные с наследованием имущества	Иск (заявление) удовлетворен (в том числе частично)
754	Гражданское	Другие споры, связанные с землепользованием	Отказано
623	Гражданское	Прочие исковые дела	Иск (заявление) удовлетворен (в том числе частично)
927	Гражданское	Связанные с приватизацией жилой площади	Иск (заявление) удовлетворен (в том числе частично)
377	Гражданское	Другие жилищные споры	Иск (заявление) удовлетворен (в том числе частично)
363	Гражданское	Другие жилищные споры	Иск (заявление) удовлетворен (в том числе частично)
091	Гражданское	Другие жилищные споры	Иск (заявление) удовлетворен (в том числе частично)
620	Гражданское	Споры, связанные с наследованием имущества	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
697	Гражданское	Споры в отношении имущества, не являющегося объектом хозяйственной деятельности	Иск (заявление) удовлетворен (в том числе частично)
101	Гражданское	Другие жилищные споры	Иск (заявление) удовлетворен (в том числе частично)
209	Гражданское	Прочие исковые дела	Иск (заявление) удовлетворен (в том числе частично)
796	Гражданское	Споры, связанные с наследованием имущества	Иск (заявление) удовлетворен (в том числе частично)
223	Гражданское	Другие споры, связанные с землепользованием	Отказано
983	Гражданское	Прочие исковые дела	Иск (заявление, жалоба) УДОВЛЕТВОРЕН ЧАСТИЧНО
843	Гражданское	Прочие исковые дела	Иск (заявление) удовлетворен (в том числе частично)

Рисунок 22 – Группировка судебных решений по классам с соответствующими группировками по полям, после обработки алгоритмом LTSA.

2585	34	1352394000	107196733	Гражданское	Другие споры, связанные с землепользованием	Др. касс.ц. опред. с удовлетворением жалоб и представлен
2586	34	1320168800	103668946	Гражданское	Другие споры, связанные с землепользованием	Отменено полностью с вынесением нового суд.постан., не пер
2587	34	1319475600	103665910	Гражданское	Прочие исковые дела	
2588	34	1315242000	103666446	Гражданское	О взыскании (независимо от принадлежности жилфонда): - иные без предоставления другого жилого помещения	
2589	34	1313686800	103665565	Гражданское	Прочие исковые дела	
2590	34	1319734800	103665979	Гражданское	Другие жилищные споры	
2591	34	1305219600	103667684	Гражданское	Другие жилищные споры	
2592	33	1317661200	419535312	Гражданское	Прочие исковые дела	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
2593	33	1317661200	419623832	Гражданское	Прочие исковые дела	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
2594	32	1296756000	103072558	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2595	32	1319994000	101299002	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2596	32	1326733200	101298752	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2597	32	1283187600	101831040	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2598	32	1282064400	103073516	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2599	32	1323884200	101298738	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2600	32	1325264400	103072544	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2601	32	1279126800	101299303	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2602	32	1282064400	103073514	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2603	32	1293732000	101299313	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2604	32	1307984400	103072212	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2605	32	1319648400	103071635	Гражданское	О признании движимой вещи бесхозяйной и признании права муницип	Иск (заявление) удовлетворен (в том числе частично)
2606	31	1334682000	106071190	Гражданское	Жалобы на неправомерные действия (бездействие) должностных лиц	Дело прекращено
2607	31	1282755600	102727326	Гражданское	Другие споры, связанные с землепользованием	Отказано
2608	31	1361379600	424439233	Гражданское	Жалобы на неправ. дейст. (безд.) - должностных лиц, государственн	ОТКАЗАНО в удовлетворении иска (заявлении, жалобы)
2609	31	1364922000	424363439	Гражданское	Жалобы на неправ. дейст. (безд.) - должностных лиц, государственн	ОТКАЗАНО в удовлетворении иска (заявлении, жалобы)
2610	31	1300730400	103285647	Гражданское	Прочие исковые дела	Иск (заявление) удовлетворен (в том числе частично)
2611	31	1315242000	106094750	Гражданское	Жалобы на неправомерные действия (бездействие) должностных лиц	Иск (заявление) удовлетворен (в том числе частично)
2612	31	1274806800	101831651	Гражданское	Прочие исковые дела	Иск (заявление) удовлетворен (в том числе частично)
2613	31	1354813200	424492656	Гражданское	Жалобы на неправ. дейст. (безд.) - должностных лиц, государственн	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
2614	31	1331398800	106075592	Гражданское	Прочие исковые дела	Иск (заявление) удовлетворен (в том числе частично)
2615	31	1310490000	101157741	Гражданское	Иски о взыскании сумм по договору займа	Иск (заявление) удовлетворен (в том числе частично)
2616	31	1354813200	424492654	Гражданское	Жалобы на неправ. дейст. (безд.) - должностных лиц, государственн	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
2617	31	1356541200	424476392	Гражданское	Прочие исковые дела	Иск (заявление, жалоба) УДОВЛЕТВОРЕН ЧАСТИЧНО
2618	31	1364835600	424328315	Гражданское	Жалобы на неправ. дейст. (безд.) - должностных лиц, государственн	ОТКАЗАНО в удовлетворении иска (заявлении, жалобы)
2619	31	1277559000	102401035	Гражданское	Прочие исковые дела	Иск (заявление) удовлетворен (в том числе частично)

Рисунок 23 – Группировка судебных решений по классам с соответствующими группировками по полям, после обработки алгоритмом LTSA.

Также стоит отметить, что наряду с явными группировками решений, присутствуют и некоторые аномальные данные, не вписывающиеся в общую картину решений, таковые, например, были найдены с помощью алгоритма t-SNE на рисунках далее.

547	82	1279472400	101158764	Уголовное	161 ч.2 п.г	
548	82	1327510800	100479237	Уголовное	158 ч.2 п.в	Обвинительный приговор
549	82	1329066000	102379874	Уголовное	158 ч.2 п.б	Обвинительный приговор
550	82	1268762400	105468912	Уголовное	158 ч.2 п.в	Обвинительный приговор
551	82	1275411600	105470633	Уголовное	161 ч.1	
552	82	1342717200	106167856	Уголовное	260 ч.2 п.г	Обвинительный приговор
553	82	1363194000	414106854	Уголовное	30 ч.3; ; 158 ч.2 п.а	Вынесен ПРИГОВОР
554	82	1310490000	104032538	Уголовное	161 ч.2 п.г	Обвинительный приговор
555	81	1329843600	103673274	Уголовное	70 ч.1; 79 ч.7; 105 ч.1	Приговор изменен без изменения квалификации
556	81	1337706000	105549318	Уголовное	69 ч.5; 162 ч.2	Приговор изменен без изменения квалификации
557	81	1307466000	103673720	Уголовное		
558	81	1306256400	103673731	Уголовное	111 ч.4	Приговор изменен с изменением квалификации со снижением меры наказания
559	81	1314723600	103673511	Уголовное	69 ч.5	Приговор изменен с изменением квалификации со снижением меры наказания
560	81	1307466000	103673726	Уголовное	69 ч.3; 30 ч.3; 228.1 ч.1	Приговор изменен с изменением квалификации со снижением меры наказания
561	81	1364490000	427612468	Административно	9.1 ч.3	Вынесено постановление о назначении административного наказания
562	81	1356454800	107273603	Уголовное	73 ч.1; 161 ч.2 п.а; 161 ч.2 п.г	Приговор изменен без изменения квалификации
563	81	1325005200	103673341	Уголовное	73 ч.1; 228 ч.2	Приговор изменен без изменения квалификации
564	81	1361898000	107307680	Уголовное	73 ч.1; 158 ч.3 п.а	Приговор изменен с изменением квалификации со снижением меры наказания
565	81	1393779600	107285349	Административно	14.4 ч.1	Отменено возвращением на новое рассмотрение
566	81	1323190800	103673466	Уголовное	30 ч.3; 158 ч.2 п.а	
567	81	1303837200	103673770	Уголовное	70 ч.1; 74 ч.5; 112 ч.1	Приговор изменен с изменением квалификации со снижением меры наказания
568	81	1338915600	105428616	Уголовное	161 ч.2 п.а; 161 ч.2 п.в; 161 ч.2 п.г	Приговор изменен с изменением квалификации со снижением меры наказания
569	81	1295978400	103673831	Уголовное		
570	81	1354035600	107256706	Уголовное	30 ч.1; 228.1 ч.3 п.г	Приговор изменен с изменением квалификации со снижением меры наказания
571	81	1295978400	103673790	Уголовное	69 ч.3; 70 ч.1; 30 ч.3; 228.1 ч.1; 234 ч.1	Приговор изменен с изменением квалификации со снижением меры наказания
572	81	1302627600	103673767	Уголовное	70 ч.1; 79 ч.7; 228 ч.1	Приговор изменен без изменения квалификации
573	81	1295892000	101298489	Административно	20.4 ч.1	Назначено административное наказание
574	81	1303146000	101298436	Административно	20.4 ч.1	Назначено административное наказание
575	81	1295978400	103673833	Уголовное		
576	80	1359046800	422713813	Гражданское	Трудовые споры - другие, возникающие из трудов	Иск (заявление, жалоба) УДОВЛЕТВОРЕН ЧАСТИЧНО
577	80	1318525200	100853804	Гражданское	Прочие иски	Оставлено без изменения
578	80	1358787600	107265259	Гражданское	О взыскании страхового возмещения (выплата	Оставлено без изменения
579	80	1324486800	103071257	Гражданское	Трудовые споры (независимо от форм собственности	Иск (заявление) удовлетворен (в том числе частично)
580	80	1297965600	103668543	Гражданское	Трудовые споры (независимо от форм собственности работодателя): - о восстановлении на работе	
581	80	1278435600	102727359	Гражданское	Трудовые споры (независимо от форм собственности	Иск (заявление) удовлетворен (в том числе частично)
582	80	1355331600	106507226	Гражданское	Трудовые споры (независимо от форм собственности	Иск (заявление) удовлетворен (в том числе частично)
583	80	1355850000	105553629	Гражданское	Трудовые споры (независимо от форм собственности работодателя): - другие, возникающие из трудовых отношений	
584	80	1079460000	100438197	Гражданское	Трудовые споры (независимо от форм собственности	Иск (заявление) удовлетворен (в том числе частично)
585	80	1273510800	102401031	Гражданское	Трудовые споры (независимо от форм собственности	Иск (заявление) удовлетворен (в том числе частично)
586	80	1331830800	104022798	Гражданское	Трудовые споры (независимо от форм собственности	Иск (заявление) удовлетворен (в том числе частично)

Рисунок 24 – Группировка судебных решений по классам с соответствующими группировками по полям, после обработки алгоритмом t-SNE.

На рисунке 24 в одном кластере с решениями по уголовным делам находятся решения и по административным делам, что следует проверить на наличие нарушений.

711	76	1304096400	100851716	Гражданское	Прочие из публично-правовых отношений	Иск (заявление) удовлетворен (в том числе частично)
712	76	1337878800	106037418	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
713	76	1357664400	415303930	Гражданское	Прочие иски	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
714	76	1308502800	102584996	Гражданское	Жалобы на правонарушения (бездей	Иск (заявление) удовлетворен (в том числе частично)
715	76	1350234000	107043794	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
716	76	1337187600	105883565	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
717	76	1366650000	425353325	Гражданское	Прочие иски	Иск (заявление) УДОВЛЕТВОРЕН
718	76	1351702800	105918791	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
719	76	1353862800	428863819	Гражданское	Трудовые споры - об оплате труда	Производство по делу ПРЕКРАЩЕНО (ст.220 ГПК)
720	76	1356886800	105358397	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
721	76	1305478800	101157888	Гражданское	Споры, вытекающие из права собственности	Иск (заявление) удовлетворен (в том числе частично)
722	76	1319562000	100399566	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
723	76	1339952400	101298722	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
724	76	1326387600	102400615	Гражданское		Иск (заявление) удовлетворен (в том числе частично)
725	75	1341766800	105914076	Гражданское	Впервые предъявленные иски о возмещении в	Иск (заявление) удовлетворен (в том числе частично)
726	75	1349715600	106886368	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
727	75	1328806800	100399089	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
728	75	1352394000	107093406	Гражданское	Прочие иски	Отказано
729	75	1295892000	105035463	Гражданское		Иск (заявление) удовлетворен (в том числе частично)
730	75	1327510800	101828939	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
731	75	1326301200	102313745	Гражданское	Впервые предъявленные иски о возмещении в	Иск (заявление) удовлетворен (в том числе частично)
732	75	1343754000	106034650	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
733	75	1352394000	107093406	Гражданское	Прочие иски	Отказано
734	75	1306947600	106702260	Гражданское	Прочие иски	Отказано
735	75	1306342800	100400248	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
736	75	1318438800	100854265	Гражданское	Иски о возмещении ущерба от ДТП	Иск (заявление) удовлетворен (в том числе частично)
737	75	1365440400	107310418	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
738	75	1557165600	103071831	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
739	75	1312995600	100399906	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
740	75	1170698400	400931056	Гражданское	Прочие иски	Иск (заявление, жалоба) УДОВЛЕТВОРЕН ЧАСТИЧНО
741	75	1353517200	406766160	Гражданское	Прочие иски	Иск (заявление, жалоба) УДОВЛЕТВОРЕН ЧАСТИЧНО
742	75	1314637200	100598220	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
743	75	1334854800	104713775	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
744	75	1309107600	100854475	Гражданское	Прочие иски	Отказано
745	75	1316538000	102250390	Гражданское	О защите чести, достоинства, деловой репутации	Иск (заявление) удовлетворен (в том числе частично)
746	75	1328029200	103285125	Гражданское		
747	75	1361725200	417621535	Гражданское	Впервые пред. иски о возм. вреда за увечье и	Иск (заявление, жалоба) УДОВЛЕТВОРЕН ЧАСТИЧНО

Рисунок 25 – Группировка судебных решений по классам с соответствующими группировками по полям, после обработки алгоритмом t-SNE.

На рисунке 25 видно явное отсутствие типа дела, однако судя по изображению скорее всего в данных решениях пропущено название «Прочие иски»

1192	65	1293040800	103446887	Гражданское	О защите прав потребителей - из договоров с	Иск (заявление) удовлетворен (в том числе частично)
1193	65	1301331600	100400366	Гражданское	Иски о взыскании сумм по договору займа	Иск (заявление) удовлетворен (в том числе частично)
1194	65	1300212000	101830125	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
1195	65	1309194000	103072064	Гражданское	Иски о взыскании сумм по договору займа	Иск (заявление) удовлетворен (в том числе частично)
1196	65	1332176400	104525233	Гражданское	Трудовые споры (независимо от форм собственности работодателя)	Иск (заявление) удовлетворен (в том числе частично)
1197	65	1308798800	102726717	Гражданское	Трудовые споры (независимо от форм собственности работодателя)	Отказано
1198	64	1363712400	107308213	Уголовное	30 ч.3; 228.1 ч.2 п.а	Другое постанов. с удовлет. надзорной жалобы или представл
1199	64	1326387600	103665122	Гражданское	Прочие иски	
1200	64	1318352400	103673582	Уголовное	30 ч.3; 159 ч.4; 159 ч.1; 30 ч.3; 159 ч.4; 186 ч.1	
1201	64	1357664400	107275515	Уголовное	159 ч.1; 161 ч.1	Другое постанов. с удовлет. надзорной жалобы или представл
1202	64	1353517200	107212785	Уголовное	30 ч.3; 228.1 ч.3 п.г	Оставлено без изменения
1203	64	1355504400	107255132	Гражданское	Трудовые споры (независимо от форм собственности работодателя): - другие, возникающие из трудовых отношений	
1204	64	1349802000	107122575	Уголовное	69 ч.5; 71 ч.1 п.в; 116 ч.1	Другое постанов. с удовлет. надзорной жалобы или представл
1205	64	1301331600	103668157	Гражданское	Споры, связанные с социальными гарантиями	Отменено с направлением дела на новое рассмотрение
1206	64	1279731600	102790908	Гражданское	Споры о праве собственности на землю	Иск (заявление) удовлетворен (в том числе частично)
1207	64	1315328400	103673625	Уголовное	158 ч.4 п.а	
1208	64	1356886800	103665057	Гражданское	Иски о возмещении ущерба от ДТП	
1209	64	1327251600	101511340	Гражданское	Прочие иски	Отказано
1210	64	1327251600	103669617	Уголовное	162 ч.2	Оставлено без изменения
1211	64	1341421200	105556184	Уголовное	115 ч.1; 116 ч.1	Оставлено без изменения
1212	64	1361120400	107315497	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
1213	64	1361120400	107283965	Уголовное	158 ч.3 п.а	
1214	64	1355331600	107241835	Уголовное	161 ч.2 п.г	
1215	64	1302022800	103673783	Уголовное	162 ч.2	Другое постанов. с удовлет. надзорной жалобы или представл
1216	64	1327251600	405022493	Гражданское	Прочие иски	ОТКАЗАНО в удовлетворении иска (заявления, жалобы)
1217	64	1330018400	103664843	Гражданское	Трудовые споры (независимо от форм собственности работодателя): - о возмещении ущерба, причиненного при исполнении трудовых обязанностей	Судебное постановление 1 инстанции по существу дела отменено
1218	64	1316538000	103673601	Уголовное	161 ч.2 п.г	
1219	64	1348419600	106847881	Уголовное	30 ч.3; 228.1 ч.2 п.б	
1220	64	1363798800	107301645	Уголовное	166 ч.4	
1221	64	1295892000	103668797	Гражданское		Оставлено без изменения
1222	64	1310490000	103673637	Уголовное	159 ч.2	
1223	64	1334509200	104558375	Уголовное	161 ч.2 п.г	Оставлено без изменения
1224	64	1325005200	103673389	Уголовное	161 ч.2 п.г; 162 ч.2	Кассационное определение изменено
1225	63	1334682000	106071190	Гражданское	Жалобы на правонарушения (бездей	Дело прекращено
1226	63	1305824400	406310312	Гражданское	Другие жилищные споры	Производство по делу ПРЕКРАЩЕНО (ст.220 ГПК)
1227	63	1336644000	104706014	Гражданское	Прочие иски	Дело прекращено

Рисунок 26 – Группировка судебных решений по классам с соответствующими группировками по полям, после обработки алгоритмом t-SNE

На рисунке 26 присутствует явно смешанная группа из решений по уголовным и гражданским типам дел, что тоже необходимо проверить.

На основе сравнительного анализа результатов выполнения алгоритма, был выбран алгоритм t-SNE как наиболее оптимальный алгоритм для дальнейшего использования.

Таким образом для решения задач кластеризации и анализа данных судебных решений, был выбран алгоритм t-SNE, так как он показал наилучшие результаты по кластеризации как визуальных способах, так и в виде сводных таблиц. В дальнейшем именно этот алгоритм будет рассматриваться в данной работе.

4.8 Алгоритм t-SNE

Алгоритм - стохастическое вложение соседей с распределением Стьюдента (t-distributed stochastic neighbor embedding) (t-SNE) - это алгоритм машинного обучения с сокращением размерности разработан Лоренсом ван дер Маатеном и Джеффри Хинтоном.

Это нелинейный метод сокращения размерности, что особенно хорошо подходит для встраивания многомерных данных в пространстве двух или трех измерений, которые затем могут быть отображены в диаграмме. В частности, этот алгоритм моделирует каждый многомерный объект в двух - или трехмерные точки таким образом, что подобные объекты моделируются близлежащими точками и разнородные объекты моделируются отдаленными точками.

Алгоритм t-SNE состоит из двух основных этапов. Во-первых, t-SNE создает распределение вероятностей над парами многомерных объектов таким образом, что подобные объекты имеют высокую вероятность быть выбранным, в то время как несхожие точки имеют бесконечно малую вероятность быть выбранным. Во-вторых, t-SNE определяет аналогичное распределение вероятностей по точкам в низкоразмерной карте, а также сводит к минимуму

расстояния Кульбака–Лейблера между распределениями относительно местоположения точек на карте. Следует обратить внимание, что в то время как оригинальный алгоритм использует Евклидово расстояние между объектами, как базы его показателей сходства, и это должно быть соответствующим образом изменено.

У алгоритма t-SNE широкий диапазон использования, включая научные исследования по компьютерной безопасности, анализ музыки, исследования рака, и биоинформатика.

4.9 Обработка данных

Входные данные для алгоритма.

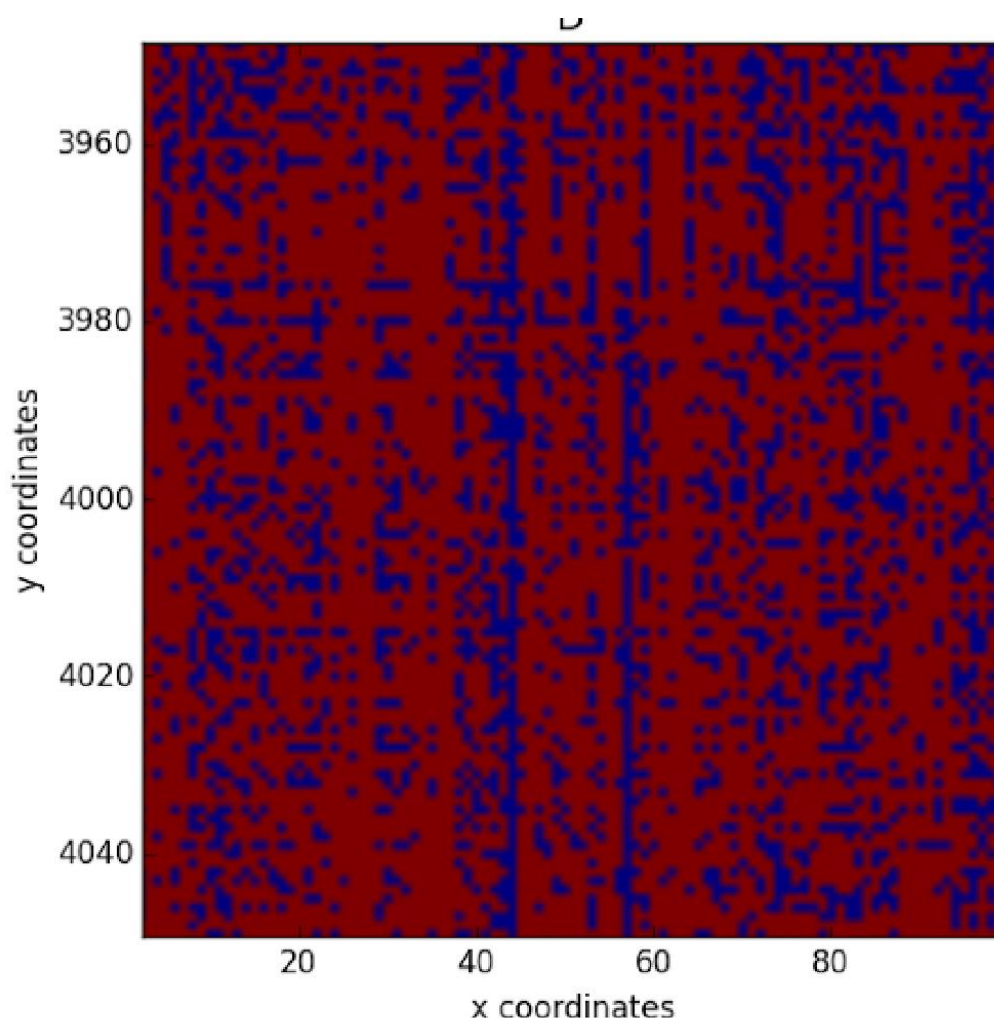


Рисунок 27 – Представление входных данных в графическом виде Python

В качестве входных данных для использования в алгоритме t-SNE используются слова решения, заранее подготовленные, отфильтрованные по

стопсловам, приведённые к нормальной форме. В рисунке 27 приведён пример частичного отображения входных подготовленных данных каждая синяя точка — это наличие слова в решении, по оси Y расположены судебные решения по оси X слова, уже из входных данных видно присутствие слов, которые есть во многих решениях, из исследования стало ясно что этими словами являются специфичные для судебной практики слова к примеру «постановил», «судебный», «взыскать» и так далее, но эти слова оставлены для чистоты исследования.

Обработка данных с помощью алгоритма t-SNE

Как ранее рассматривалось в алгоритме t-SNE используются многомерные объекты для исследования, в качестве многомерного объекта представлены судебные решения и каждой мерой является слово в судебном решении приведённое к нормальной форме. Чтобы представить каждое слово по важности проводится индексация текста, и по сути обозначает количество раз которое встречается слово в тексте, таким образом мы полностью представили массив слов из каждого решения в виде числовых эквивалентов, для того чтобы можно было применять математические модели. Алгоритм t-SNE снижает многомерность нелинейно, и отражает отношение одного судебного решения к другому, то есть сравнение судебных решений опираясь на набор слов, которые они содержат. Ещё одна особенность данного алгоритма в том, что он отображает результат на двумерной плоскости в графическом виде и в случае наличия некоторых зависимостей и особенно при наличии кластеров распределяет точки на плоскости так, что кластеры не накладываются друг на друга. Результат можно видеть на рисунке 28.

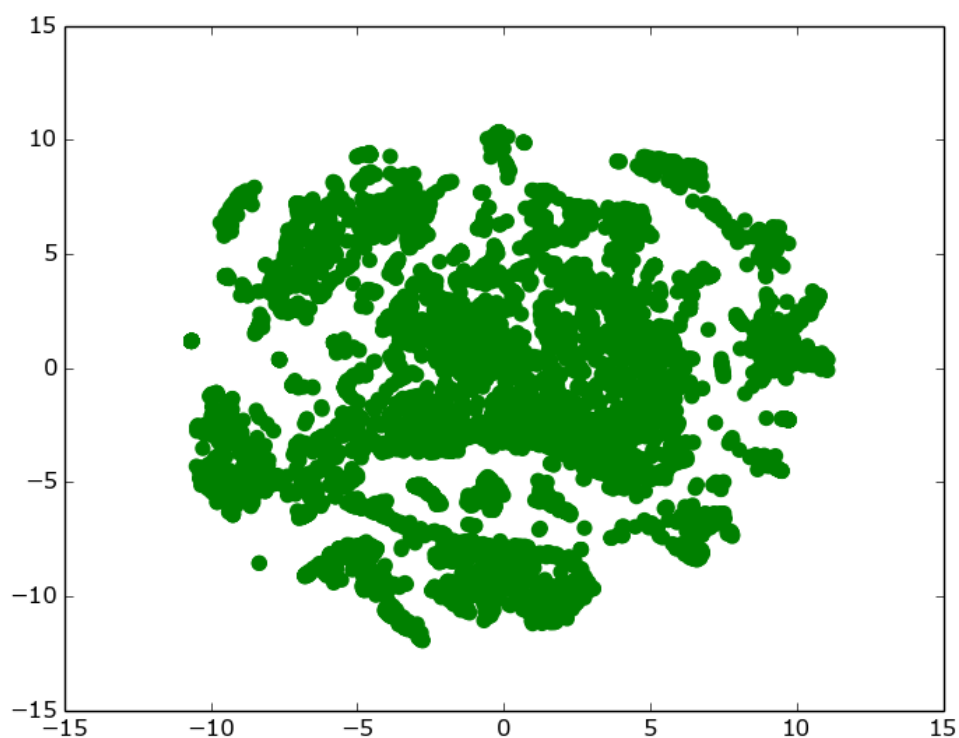


Рисунок 28 – Представление результата выполнения алгоритма t-SNE в виде точечной диаграммы

Как видно из рисунка 28 точки накладываются друг на друга, и кластеризация хоть и заметна, но выглядит нечётко, для этого меняется способ отображения, каждый элемент выводится в виде знака “+”, а также добавляется свойство прозрачности для каждого элемента, таким образом кластеризация чётко видна в местах где она есть, и те места на диаграмме в которых присутствие элементов не значительно ими соответственно не заполнено.

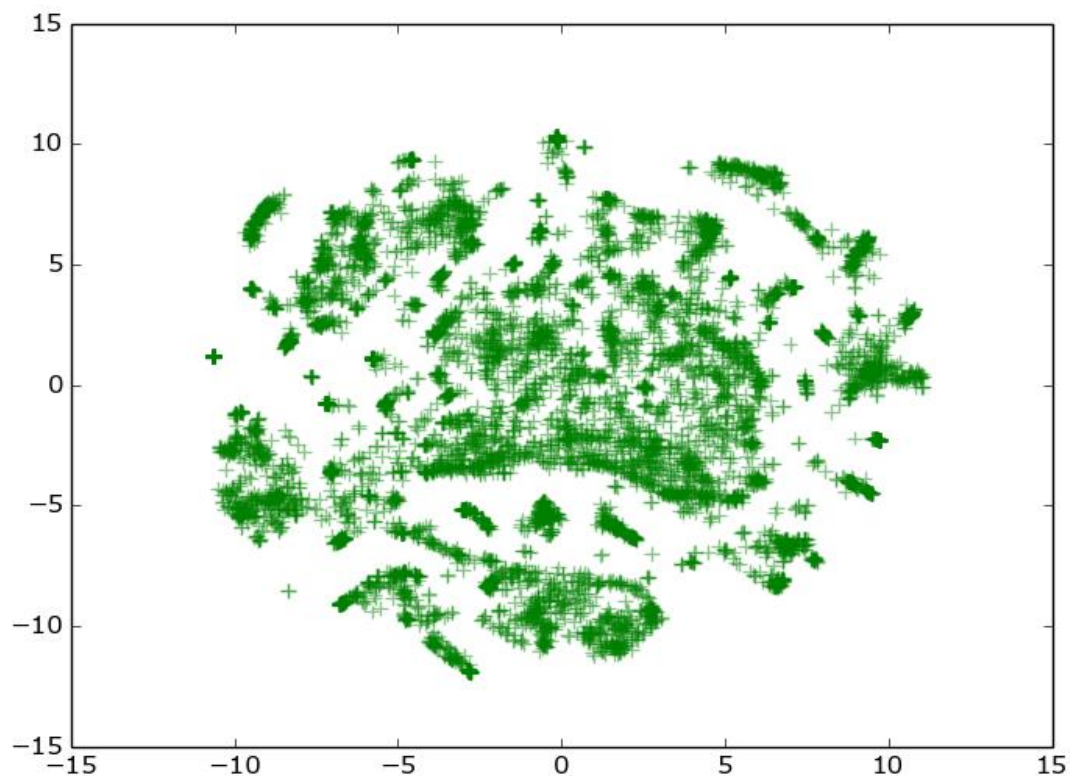


Рисунок 29 – Представление результата выполнения алгоритма t-SNE в виде точечной диаграммы с использованием знака “+” и альфа-прозрачности

На рисунке 29 отчётливо видны кластеры и отдельные объекты между кластерами. Схожесть текстового состава приближает объекты друг к другу на диаграмме, соответственно различия отдаляют, также происходит группировка объектов по группам слов что и создаёт кластеры на диаграмме.

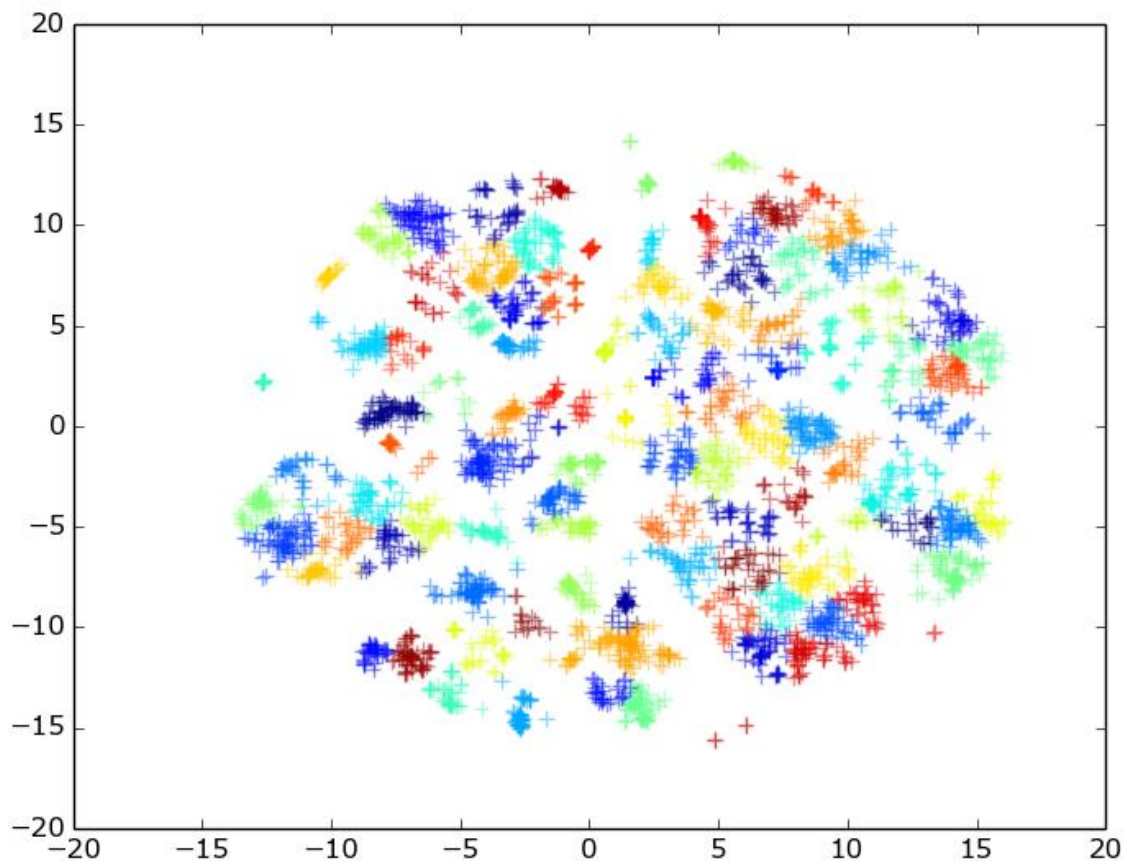


Рисунок 30 – Представление результата выполнения алгоритма t-SNE в виде точечной диаграммы с обозначенными кластерами

На рисунке 30 кластеры дополнительно раскрашены в цвета, и явно выделяются.

4.10 Результат обработки и анализа данных

В результате обработки и анализа судебных решений были выявлены определённые факты и закономерности, а также возможности для упрощения работы с большим количеством текстовых документов.

49	98	1290103200	103288100	По материалам					
50	98	1366822800	107314708	Уголовное		Оставлено без изменения			
51	98	1362416400	107317401	Уголовное		Оставлено без изменения			
52	97	1287939600	103214221	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров об отмене условного осуждения с освобождением от наказания				
53	97	1305133200	102462756	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров об отмене условного осуждения с освобождением от наказания				
54	97	1356368400	107217771	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о возложении обязанностей				
55	97	1325350800	102727694	По материалам					
56	97	1316451600	102727735	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о возложении обязанностей				
57	97	1340298000	107224901	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о замене одного вида наказания на другой				
58	97	1360602000	107291817	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров иные материалы в порядке исполнения приговора				
59	97	1329152400	102945839	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о возложении обязанностей				
60	97	1303750800	102790923	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров об отмене условного осуждения с освобождением от наказания				
61	97	1360602000	107291821	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о продлении срока условного осуждения (ч. 2 ст. 74 УК РФ)				
62	97	1309971600	102462448	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о возложении обязанностей				
63	97	1374426000	106561477	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров				
64	97	1291226400	103074323	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о замене исправительных работ лишением свободы				
65	97	1348074000	106599742	По материалам					
66	97	1353085200	107153381	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о возложении обязанностей				
67	97	1297015200	103074236	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров иные материалы в порядке исполнения приговора				
68	97	1308762000	102462478	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о возложении обязанностей				
69	97	1320253200	102283311	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о замене обязательных работ лишением свободы				
70	97	1297706400	102727932	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о возложении обязанностей				
71	97	1291744800	103073941	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о возложении обязанностей				
72	97	1355936400	107208997	По материалам	Производство по материалам, разрешаемым в порядке исполнения приговоров о возложении обязанностей				

Рисунок 31 – Отсутствующие атрибуты в судебных решениях.

Как видно из рисунка 31 в процессе обработки решений судов, появляются документы с неопределёнными полями, данный инструмент позволяет их восполнить и создать базу данных документов судов более полной и валидной.

96	1310922000	101157826	Гражданское	Споры, связанные с наследованием имущества	Иск (заявление) удовлетворен (в том числе частично)
96	1361206800	400102393	Гражданское	Другие жилищные споры	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
96	1388422800	400102381	Гражданское	Споры, связанные с наследованием имущества	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
96	1339002000	105207472	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
96	1299693600	101511444	Гражданское	Споры, связанные с наследованием имущества	Иск (заявление) удовлетворен (в том числе частично)
96	1325264400	103072690	Гражданское	Споры, вытекающие из права собственности: государственное	Отказано
96	1312909200	405758320	Гражданское	Прочие иски	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
96	1388422800	417621525	Гражданское	Споры, связанные с наследованием имущества	Иск (заявление, жалоба) УДОВЛЕТВОРЕН
96	1330362000	104428454	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
96	1330880400	104428455	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
96	1314032400	101157745	Гражданское	Прочие иски	Иск (заявление) удовлетворен (в том числе частично)
96	1336755600	105883561	Гражданское	Споры, связанные с наследованием имущества	Иск (заявление) удовлетворен (в том числе частично)

Рисунок 32 – Результат судебного решения выбивается из списка остальных решений, хотя находится в той же группе.

Как видно из рисунка 32 есть совокупность судебных решений сходных по атрибутам и находящиеся в одной группе, однако среди них присутствует решение, результат которого отличен от остальных решений, что может свидетельствовать о несоответствии результата и сути судебного дела.

При изучении отдельных объектов из кластеров были выявлены прямые зависимости между полями объектов-решений и расположением объектов-решений на точечных диаграммах и в таблицах для анализа решений, зависимость проявляется в первую очередь от следующих полей:

- имя судьи
- название суда

- тип суда
- норма права
- тип документа
- вид спора
- исход дела

Также следует отметить что данная работа проводилась только для судебных решений города Томска и Томской области, но была проведена попытка обработки данных из разных регионов РФ, в результате была выявлена дополнительная кластеризация по регионам. Заметную роль в кластеризации играют названия фирм и типы юридических лиц.

В ходе применения алгоритма t-SNE были использованы различные способы отображения результатов.

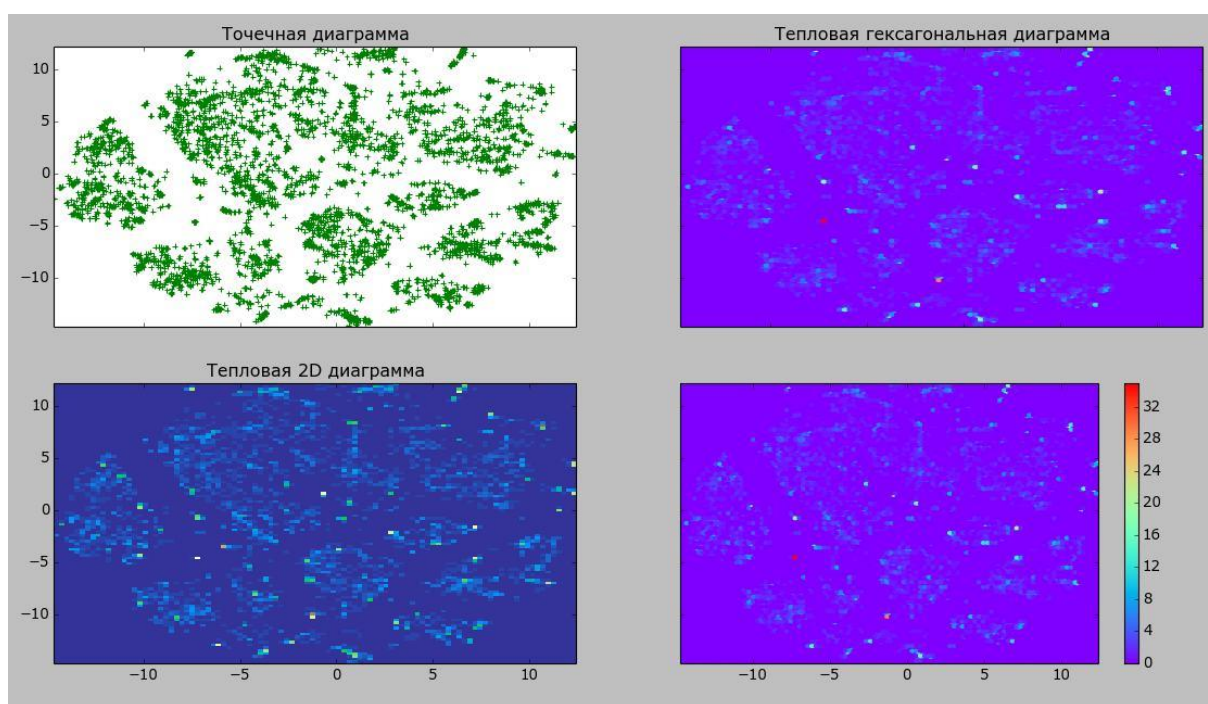


Рисунок 33 – Сравнение различных диаграмм отображения результата выполнения алгоритма t-SNE

Как видно в сравнении различных диаграмм, именно точечная диаграмма даёт наиболее наглядную картину во взаимосвязях объектов и их кластеризацию.

Полученный результат:

- выявленные основные факторы, влияющие на сходимость решений друг с другом;
- было проведено сравнение алгоритмов обрабатывающих многомерные объекты методами нелинейного снижения размерности и кластеризации, определён наилучший алгоритм для обработки таких многомерных объектов;
- получен инструмент, позволяющий ускорить процесс анализа судебных решений;
- были выявлены некоторые отклонения в судебных решениях, которые необходимо изучать в дальнейшем;
- получена совокупность инструментов для программной классификации судебных решений в автоматическом режиме, позволяющий структурировать ранее неструктурированные данные, также данный способ может быть применён и к другим текстовым документам, не только к судебным решениям.

4.11 Вывод

В ходе выполнения работы был освоен весь цикл технологий, используемый в Data Mining. В частности, подготовка данных, очистка данных, индексирование данных, формирование необходимой структуры данных для успешного применения алгоритмов Data Mining. Были выбраны алгоритмы для обработки многомерных объектов и проведён сравнительный анализ выбранных алгоритмов. Были обнаружены явные группы решений, связанные между собой различными атрибутами, причём данные группы формируются полностью автоматически с помощью технологий и алгоритмов Data Mining. В процессе поиска неявных зависимостей были обнаружены некоторые факты, которые могут послужить для дальнейшей проверки решений, такие как

причисление к группам решений с уголовными статьями решений с административными статьями.

Стоит учесть, что количество публикуемых судебных решений возрастает и в будущем станет возможным провести обработку ещё большего количества решений, в том числе по территории всей Российской Федерации. А также при определённом изменении исходного кода можно добиться значительного улучшения в подготовке и изъятии данных, что повысит их точность, как следствие с одной стороны увеличится количество решений, а с другой уменьшит число исключённых фильтрами судебных решений.

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ИМ4А	Хлопонин Александр Александрович

Институт	кибернетики	Кафедра	ВТ
Уровень образования	магистр	Направление/специальность	ИСиТ

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:

1. Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих	Работа с информацией, представленной в российских и иностранных научных публикациях, аналитических материалах, статистических бюллетенях и изданиях, нормативно-правовых документов.
2. Нормы и нормативы расходования ресурсов	
3. Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования	

Перечень вопросов, подлежащих исследованию, проектированию и разработке:

1. Оценка коммерческого и инновационного потенциала НТИ	Проведение предпроектного анализа. Определение целевого рынка и проведение его сегментирования. Выполнение SWOT-анализа проекта.
2. Разработка устава научно-технического проекта	Определение целей и ожиданий, требований проекта. Определение заинтересованных сторон и их ожиданий.
3. Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок	Составление календарного плана проекта. Определение бюджета НТИ.
4. Определение ресурсной, финансовой, экономической эффективности	Приведение оценки экономической эффективности разработки информационно-аналитической системы.

Перечень графического материала (с точным указанием обязательных чертежей):

1. Доли участия исполнителей проекта в запланированных работах
2. Распределение трудоемкости запланированных работ по исполнителям
3. Линейный график запланированных работ
4. Изменение степени готовности проекта в процессе выполнения запланированных работ
5. Структура цены проекта
6. Оценка научно-технического уровня проекта

Дата выдачи задания для раздела по линейному графику	
---	--

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент каф. менеджмента ИСГТ	И.С. Антонова	к.э.н., доцент		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ИМ4А	Хлопонин Александр Александрович		

Заключение

В ходе выполнения данной работы были выполнены следующие задачи: анализ предметной области и текстов судебных решений, проектирование и реализация базы данных для хранения извлеченных параметров, проектирование и разработка инструментов для извлечения атрибутов судебных дел и наполнения базы данных, Data Mining анализ данных.

Изучение предметной области позволило выявить актуальность исследования, определить перечень атрибутов, которые возможно и необходимо извлечь из судебных решений, а также спроектировать структуру базы данных и инструментов для извлечения атрибутов судебных дел и наполнения базы данных. На основе спроектированного решения была выполнена программная реализация инструмента для извлечения атрибутов судебных дел в среде разработки Pycharm на языке программирования Python. А также выполнена программная реализация инструмента для анализа решений судов, хранимых в подготовленном виде.

Data Mining анализ данных проводился в среде Pycharm, а также с помощью библиотек numpy, sklearn, matplotlib. Проведенный анализ позволил получить статистическую информацию и выявить некоторые закономерности между различными атрибутами решений. Так, например, основными атрибутами решений, влияющих на их сходимость являются имя судьи, название суда, тип решения, тип наказания и т.д.

В дальнейшем данное исследование можно развивать, увеличив объем анализируемых судебных решений и анализировать не только судебные дела Томской области, а дела всей Российской Федерации. Кроме того, можно увеличить объем параметров, извлекаемых из судебных решений.

Список публикаций

1. Щукова К. Б. , Хлопонин А. А. , Паршина Д. М. Извлечение и анализ данных о судопроизводстве в г. Томске с помощью технологий OLAP и DATA MINING // Технологии Microsoft в теории и практике программирования: сборник трудов XII Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых, Томск, 25-26 Марта 2015. - Томск: ТПУ, 2015 - С. 105-106.

2. Хлопонин А. А. , Паршина Д. М. Обработка решений судов Томской области и города Томска с помощью технологий OLAP и Data Mining // Технологии Microsoft в теории и практике программирования: сборник трудов XIII Всероссийская научно-практическая конференция студентов, аспирантов и молодых ученых, Томск, 22-23 Марта 2016. - Томск: ТПУ, 2016.

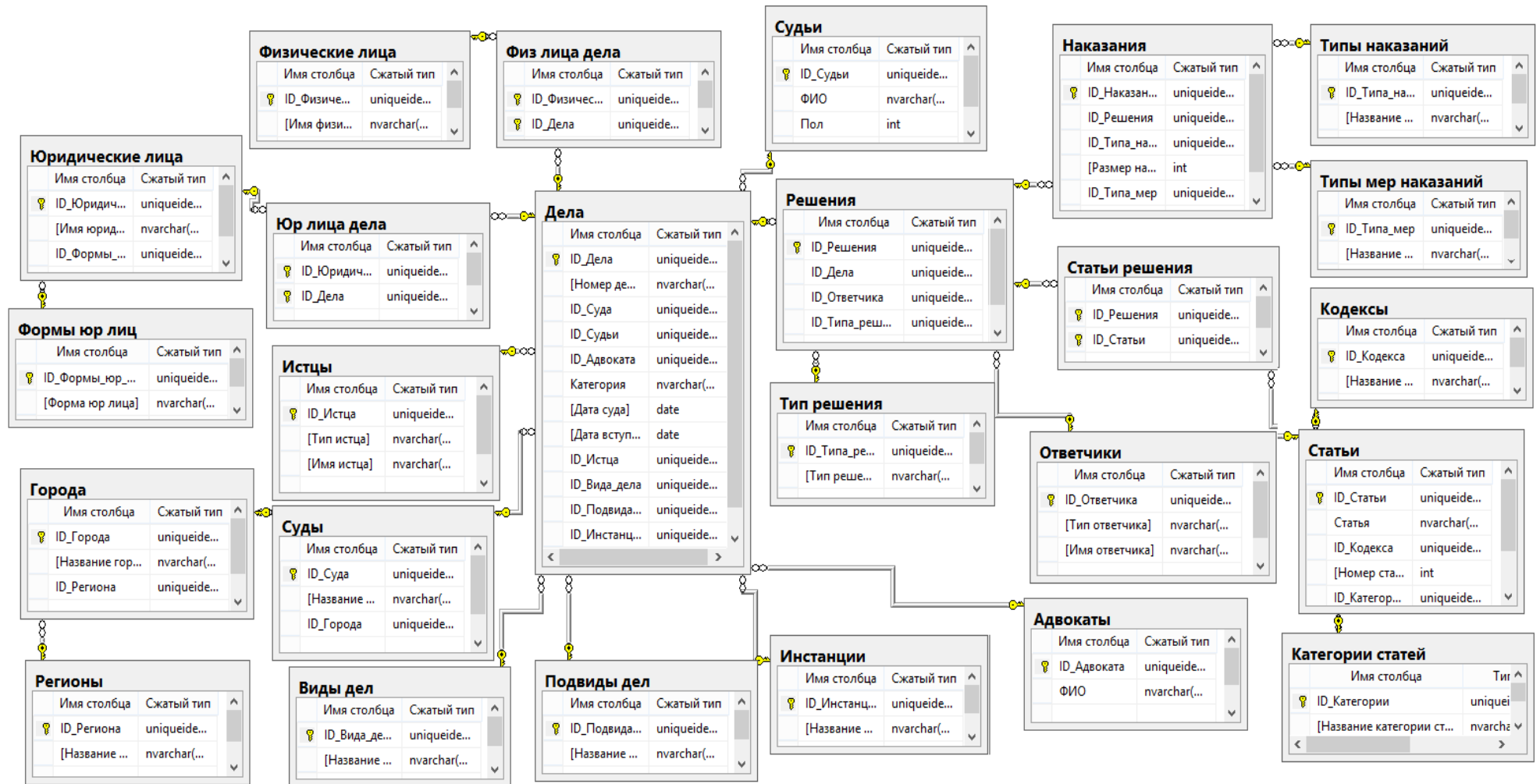
Список использованных источников

1. Федеральный закон N 262-ФЗ от 22.12.2008 (ред. от 05.04.2016) «Об обеспечении доступа к информации о деятельности судов в Российской Федерации».
2. «Гражданский процессуальный кодекс Российской Федерации» от 14.11.2002 N 138-ФЗ (ред. от 02.03.2016).
3. CaseLook [Электронный ресурс]. URL: <http://caselook.ru/> (дата обращения: 28.05.2016).
4. РосПравосудие [Электронный ресурс]. URL: <http://www.rospravosudie.com/> (дата обращения: 28.05.2016).
5. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов / Ю. Г. Зеленков, И. В. Сегалович // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL 2007: Сборник работ участников конкурса. — Переславль-Залесский, Россия. — 2007.
6. Двоичные коды с исправлением выпадений, вставок и замещений символов / В. И. Левенштейн // Доклады Академий Наук СССР – 1965.
7. Регулярные выражения / Фридл, Дж. — СПб., 2001. — 352 с.
8. «An Introduction to MDS» Wickelmaier, Florian [Электронный ресурс]. URL: <https://homepages.uni-tuebingen.de/florian.wickelmaier/pubs/Wickelmaier2003SQRU.pdf> (дата обращения: 10.05.2016).
9. Adaptive Neighborhood Graph for LTSA Learning Algorithm without Free-Parameter. Xianlin Zou, Qingsheng Zhu [Электронный ресурс]. URL: <http://www.ijcaonline.org/volume19/number4/pxc3873070.pdf> (дата обращения: 11.05.2016)
10. Nonlinear Dimensionality Reduction I: Local Linear Embedding. Sam T. Roweis [Электронный ресурс]. URL: <http://www.stat.cmu.edu/~cshalizi/350/lectures/14/lecture-14.pdf> (дата обращения: 11.05.2016)

11. t-SNE. Laurens van der Maaten. [Электронный ресурс]. URL: <https://lvdmaaten.github.io/tsne/> (дата обращения 12.05.2016)
12. Влияние морфологического анализа на качество информационного поиска. М.В. Губин, А.Б. Морозов [Электронный документ]. URL: http://rcdl.ru/doc/2006/paper_67_v2.pdf (Дата обращения 09.04.2016).
13. СанПиН 2.2.2/2.4.1340 – 03. Санитарно – эпидемиологические правила и нормативы «Гигиенические требования к персональным электронно – вычислительным машинам и организации работы». – М.: Госкомсанэпиднадзор, 2003.
14. СНиП 23-05-95 от 20.05.1995. Строительные нормы и правила Российской Федерации «Естественное и искусственное освещение».
15. СанПиН 2.2.2/2.4.1340-03 Гигиенические требования к персональным электронно-вычислительным машинам и организации работы: с изменениями от 25 апреля 2007 г. – М.: Информационно-издательский центр Минздрава России, 2003.
16. СанПиН 2.2.1/2.1.1.1200-03 Санитарно-защитные зоны и санитарная классификация предприятий, сооружений и иных объектов. – М.: Госкомсанэпиднадзор России, 2003.
17. Научно-информационный портал Винити. Экология [Электронный ресурс]: – Режим доступа: <http://science.viniti.ru>
18. Технический регламент о требованиях пожарной безопасности: Федеральный закон от 22 июля 2008 года N 123-ФЗ.
19. ГОСТ Р 22.3.03 – 94. Безопасность в чрезвычайных ситуациях. Защита населения.

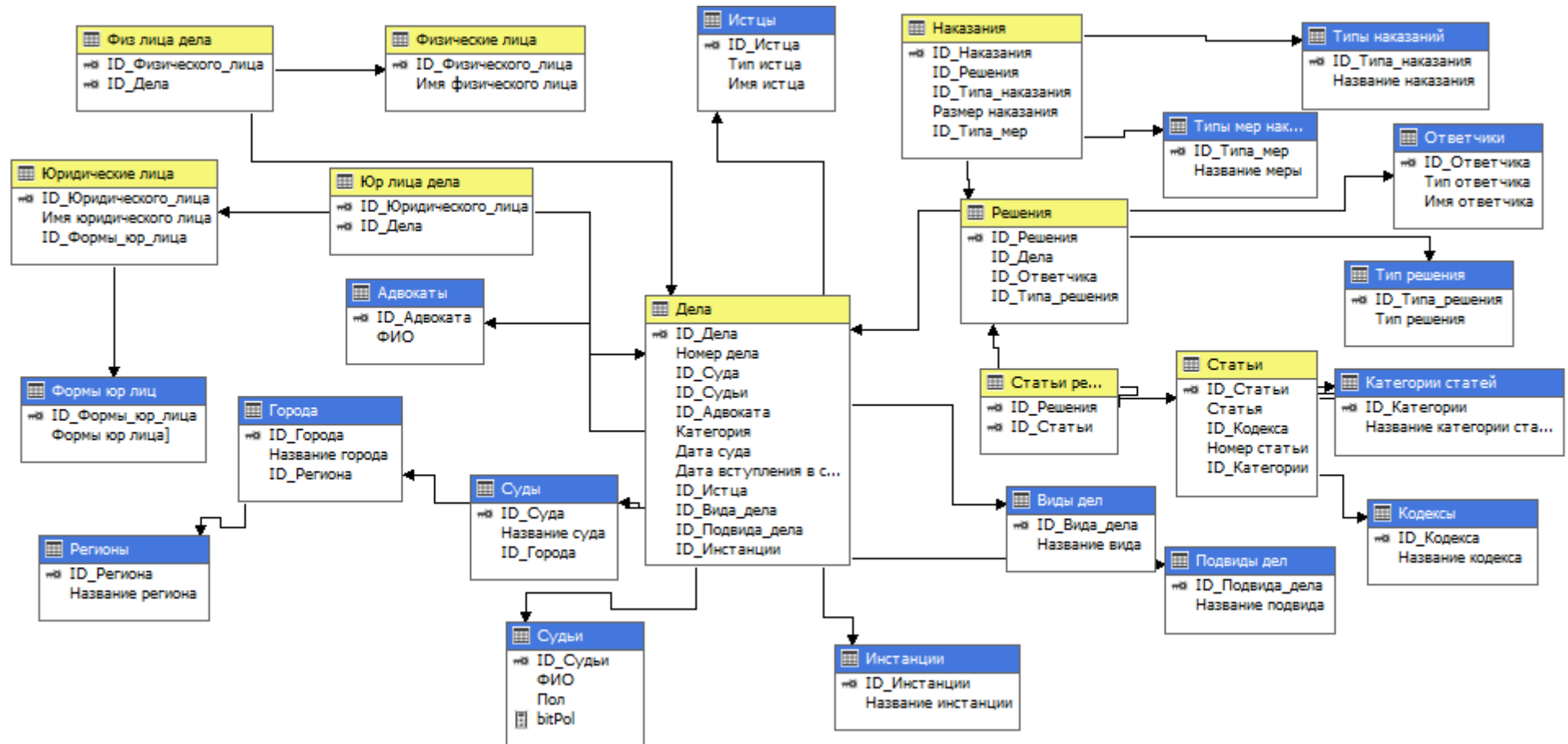
Приложение А

Физическая схема базы данных



Приложение Б

Схема куба данных судопроизводства



Приложение В
Разделы на иностранном языке

Разделы 2, 3, 4
Design of data processing tools.
Development of data processing tools.
Data analysis judicial decisions using Data Mining technologies

Студент:

Группа	ФИО	Подпись	Дата
8ИМ4А	Хлопонин Александр Александрович		

Консультант кафедры ВТ:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент каф. ВТ	Мирошниченко Е.А.	к.т.н.		

Консультант – лингвист кафедры иностранных языков ИК:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель	Шепетовский Д.В.			

Термины и определения	Terms and definitions
Добыча данных(дословный перевод)	Data Mining
Аналитическая обработка данных в реальном времени (обычно термин оставляют в виде аббревиатуры на английском языке OLAP)	OLAP(online analytical processing)
Индексация данных	Indexing the data
Парсинг данных	Parsing the data
Стемминг	Stemming
Лемматизация	Lemmatization
База данных	Database
Не только SQL (обычно используется аббревиатура на английском языке - NoSQL)	NoSQL(not only SQL)
Стохастическое вложение соседей с распределением Стюдента (обычно используется аббревиатура на английском языке - t-SNE)	t-SNE(t-distributed stochastic neighbor embedding)
Судебные решения	Judicial decisions

2 Design of data processing tools

2.1 Choice of development tools

The Python language was used for extracting of judicial decisions parameters in the Pycharm development environment.

Main advantages of chosen environment and language are:

- readable syntax;
- simple in learning;

- high-level language;
- object oriented language;
- wide set of plugins.

For the database creating was used the Microsoft SQL Server 2014

Management Studio environment due to:

- usable graphical interface;
- rich features;
- high performance;
- reliability.

The Visual Studio 2013 development environment was used for filling database of application. The choice of this software is conditioned with next opportunities:

- working with database objects in SQL Server object explorer;
- tables definition in new constructor of tables;
- development and testing of applications of databases in SQL Server

Express Local DB

- Writing of asynchronous code by simple and intuitive way;
- Obtaining information about the calling object which will help with tracing and debugging;

• The C# programming language was used because it has following advantages:

- object oriented;
- the commitment to safety code;
- versatility;
- relative simplicity.

3 Development of data processing tools

3.1 Development of the attributes extraction tool from decisions

The tool for extracting attributes from judicial decisions was implemented in the Python programming language using its different extensions allowing to extract text information.

The main problem of implementation was absence of unified structure and defendants was unnamed. Therefore, for some parameters we had to make huge list of regular expressions. At that these lists had to complement constantly because decisions in different cases had large differences. Thus, implementation of this tool was iterative.

As a result was obtained the following data: number of case, region, city, kind of case, type of case, instance, name of judge, gender of judge, date of becoming effective, name of standard regulation, article, punishment, the amount of punishment, lawyer, defendant, complainant, juridic persons, physical persons.

3.2 Development of the database filling tool

Judicial proceedings database filling tool was implemented based on the results of the design on C# programming language in Visual Studio 2013 environment. The tool gets the collection of XML files on input.

The main problem in development of this tool was correctly implement relations between different entities and to envisage all possible variants of attributes values. So names can be written in different ways: surname, name, middle name; surname, initials of name and middle name; initials of name and middle name then surname. Therefore at first names were splitted into separated words then it was determined which of them is the surname and which is the name or patronymic. Then the surnames are compared by Levenstein distance and separately comparing of name and middle name for coincidence.

The filled database was get and filled of judicial decisions attributes as a result of work. All attributes was written into matched fields of entities.

4 Analysis of a data by Data Mining technology

The main phases of Data Mining:

1. Data preparing
2. Indexing of data
3. Normalize of data for using in handling algorithms
4. Applying of algorithms and getting results

1) Data preparing.

At that phase we must have the data as a result of parsing from sites or documents of courts. That is at this phase assumes presence of text information.

The purpose of this phase is to parse your collected text data, and to store it in a structured format for later use. The goals of the phase are as follows:

1. Visually inspect text files.
2. Parse the files.
3. Serialize complex objects with JSON
4. **Visual inspection of the text documents** is an important part of the process of analyzing. During the visual inspection it is identifies parts of documents, in particular there are necessary text parts in judicial decisions, there are parts of text which not necessary takes place as different modifications or completely absent in some judicial decisions. For example, each decision has an information about justice, date of pronouncement of judgment etc. Also, each decision is separated into the text of justification of decision and text of sentence. Besides, each decision has sequence of specific tokens and parts of text that gives possibility for applying different algorithms in handling of text of decisions.

5. **Parsing of text files** is an analysis of text using multivariate analyzer with next separation single information entities and complicated object-model of judicial decisions in order to improve its capabilities. For example there is an entire

text in beginning of information handling, and after parsing and enrichment of data the object has extra fields and it gives extra opportunities of learning of judicial decisions, opportunity of grouping, sorting and filtering in dataset, decision can have any quantity of fields. That is parsing allows to allocate maximally possible dataset of entities for managing of judicial decisions.

6. **Saving of complexed JSON objects** is a moving prepared data into store for next working with them. The database is a NoSQL by principle it is not relational. Such method of data storing provides simplicity of data extracting, exclude the need for complex queries for getting whole object of judicial decision, lets to get simple statistical values on the fly.

2) Indexing of data

Indexing of data can be separated on a few parts:

splitting into words, filtering and extraction of stopwords, stemming or lemmatization, indexing itself.

Splitting into words is a procedure of text fragmentation (in this work it is a string splitting of judicial decision) to obtain the data array which has single element as word, because all words are splitted by space and accordingly fragmentation of string is performed in accordance with by space symbol.

Filtering of words and extraction of stop words is phase with removing of not important words which not affect the analysis of text in future.

Stopwords are words that added subtext to meaning, the extra words which not contain meaning separately and they are auxiliary, also these include prepositions, particles, conjunctions.

Stemming and lemmatization of words is a bringing of words to a normal form. These two methods are different but have similar goal. Stemming is an algorithm, that detach prefixes, suffixes and endings and gives in result cut form of word that has only base part of word. This method is simpler in implementation but can be wrong because splitting can be incorrect.

Lemmatization of word is an algorithm which goal has bringing of word to normal form singular number and the nominative case. This algorithm is harder in implementation but result is better for next using in future.

Indexing of data is the counter of words in the text, each word encountered one time has index 1 accordingly how many times word found in text, that index it has.

3) For using of data in different algorithms it is necessary to bring the data to right form simple text or words dataset not always can match for operations, primarily it is necessary to have the data processing model and understanding whether dataset match with this model or not. If not this algorithm should be rejected in conjunction to loss of use or it needs to get extra information from current dataset

4) The execution of algorithm is running of software which has implemented algorithm for computing of dataset and to obtain prepared information in needed format