Оглавление

[Введение 5](#_Toc511700123)

[1. Техническое задание 6](#_Toc511700124)

[1.1 Введение 6](#_Toc511700125)

[1.2 Требования 6](#_Toc511700126)

[1.2.1 Требования к функциональным характеристикам 6](#_Toc511700127)

[1.2.2 Требования к графическому интерфейсу пользователя 6](#_Toc511700128)

[1.2.3 Требования к надежности 6](#_Toc511700129)

[1.2.4 Требования к составу и параметрам технических средств 7](#_Toc511700130)

[1.2.5 Требования к информационной и программной совместимости 7](#_Toc511700131)

[1.4 Основания для выполнения работы 7](#_Toc511700132)

[1.5 Заказчик и исполнитель работ 7](#_Toc511700133)

[2. Анализ технического задания 8](#_Toc511700134)

[3. Анализ программных средств 12](#_Toc511700135)

[3.1 Выбор операционной системы 12](#_Toc511700136)

[3.2 Выбор языка программирования 13](#_Toc511700137)

[3.3 Выбор среды разработки 14](#_Toc511700138)

[3.4 Выбор почтового клиента 15](#_Toc511700139)

[4. Информационная модель программы 18](#_Toc511700140)

[4.1 Система отображения информации 18](#_Toc511700141)

[4.2 Интерфейс 18](#_Toc511700142)

[4.3 Представление информации отображаемой на экране. 19](#_Toc511700143)

[4.3.1 Зоны программы 19](#_Toc511700144)

[4.3.2 Информация, отображаемая в системе сортировки 20](#_Toc511700145)

[4.3.3 Информация, отображаемая в зоне №1. 21](#_Toc511700147)

[4.3.4 Информация, отображаемая в зоне №2 22](#_Toc511700148)

[4.3.5 Информация отображаемая №3. 22](#_Toc511700149)

[5. Байесовский классификатор 23](#_Toc511700150)

[5.1 Теорема Байеса 24](#_Toc511700151)

[5.2 Общая структура байесовского классификатора 25](#_Toc511700152)

[5.3 Наивный байесовский классификатор 26](#_Toc511700153)

[5.4 Оптимальный байесовский классификатор 27](#_Toc511700154)

[6. Сетевая папка 29](#_Toc511700155)

[6.1 Настройка сети 29](#_Toc511700156)

[6.2 Общий доступ к папке 31](#_Toc511700160)

[7. Примеры 35](#_Toc511700163)

[7.1 Сортировка писем 35](#_Toc511700164)

[8. Исследовательская часть 39](#_Toc511700195)

[9. Заключение 43](#_Toc511700196)

[10. Список литературы 44](#_Toc511700198)

# Введение

В настоящее время различная информация рассматривается как ресурс, который, можно добывать, использовать, перерабатывать, распространять, как и традиционные ресурсы. Но информация еще не будет иметь силу и практическое применение, пока она грамотно не переработана, не сохранена. В современном обществе информация является ресурсом производства, важным элементом жизни общества.

Деятельность организаций неразрывно связана с большим потоком документации. С помощью документа информация передается в пространстве и времени. Документация, как носитель информации, выступает в качестве непременного элемента организации деятельности учреждения. Документы находятся в учреждении в огромном количестве, и каждый из них имеет свое назначение. Документооборот учреждения включает в себя входящие и исходящие документы. Они очень важны, т.к. поддерживают связь учреждения с другими организациями и органами власти. Приказы, постановления, являются в первую очередь входящими документами. От правильного их движения зависит своевременность и быстродействие исполнения. Организация оборота документов важна потому, что современное общество не стоит на месте, совершенствуется, развивается. От оперативности обработки и движения документов в конечном итоге зависит быстрота и правильность принятия решений.

**Вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввывывывывуцуцавпвапвапааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыавыааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыаываааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааа**

**ааааааааааааааааааааа**

1. **Техническое задание**
   1. **Введение**

Наименование: система автоматизированного анализа писем.

Область применения: Обработка входящих писем, приходящих на электронную почту.

* 1. **Требования** 
     1. **Требования к функциональным характеристикам**

Должна обеспечивать:

- Прием, обработку входящей документации;

- Выполнять сортировку входящих писем в раздел по темам.

- Программа должна сохранять файл, если он был изменен.

- Программа должна иметь режим обучения.

* + 1. **Требования к графическому интерфейсу пользователя**

Средствами графического интерфейса программа должна обеспечивать следующий состав отображаемой информации:

- кнопку указывающий путь к папке с входящими письмами

- количество обработанных документов;

- режим отображения количества писем в данной категории;

- режим обучения системы;

- кнопку сортировки писем»;

* + 1. **Требования к надежности**

Программное обеспечение СОИ должно удовлетворять требованиям по надежности:

- ГОСТ 28806-90 «Качество программных средств. Термины и определения»

Программа должна корректно функционировать и не иметь деструктивных последствий для операционной системы и прочих программ и данных, хранящихся на компьютере.

* + 1. **Требования к составу и параметрам технических средств**

• PC-компьютер с доступом в Интернет;

• Процессор (CPU) - не ниже Pentium III-600;

• Оперативная память (RAM) не ниже 256 Мб;

• Web-браузер с поддержкой javascript (IE6+, Fx2+, Opera 8+, GoogleChrome)

• Клавиатура

• Монитор

* + 1. **Требования к информационной и программной совместимости**

- специальное программное обеспечение должно функционировать под управлением операционной системы «MS Windows»;

- специальное программное обеспечение должно быть разработано на языке C++

- требуется обеспечить корректное функционирование других программ, которые работают с клавиатурой;

- исполняемый файл должен быть в формате \*.exe-модуля.

##### Назначение и область применения

Система сортировки писем предназначена для взаимодействия органов местного самоуправления и повышения уровня информатизации.

Область применения – взаимодействие органов местного самоуправления, с помощью технологии Интернет.

1. **Анализ технического задания**

В ходе выполнения магистерской работы необходимо разработать систему автоматизированного анализа писем, которая в соответствии с требованиями ТЗ должна обеспечивать сортировку входящих писем электронной почты:

- кнопку указывающий путь к папке с входящими письмами

- режим отображения количества писем в данной категории;

- режим обучения системы;

- кнопку сортировки писем»;

- сортировку входящей документации.

Отображаемая информация (рис. 3.1 Схема состава отображаемой информации):

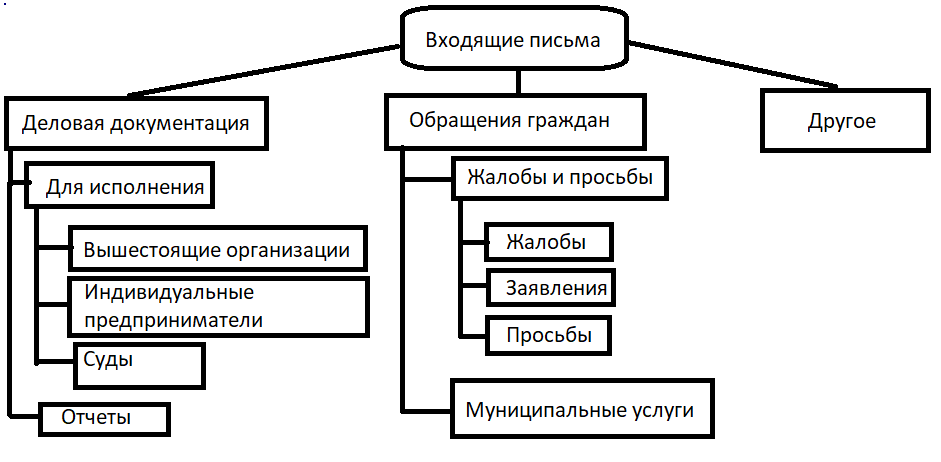


Рис. 3.1 - Схема состава отображаемой информации

Входящие письма делятся на три категории:

* Деловая документация
* Обращения граждан
* Другое

Деловая документация - это официально-деловые документы, регламентирующие деятельность организаций, учреждений, предприятий, фирм.

Деловая документация включает в себя:

* Для исполнения
* Отчеты

Документы для исполнения приходят от вышестоящих организации, индивидуальных предпринимателей или судов. Данные документы перенаправляются органам местного самоуправления для работы.

Отчетность — это система взаимосвязанных показателей, характеризующих условия и результаты деятельности организации или составляющих его элементов.

Составление отчетности — завершающий этап учетного процесса, который состоит из обобщающих итоговых показателей, получающих в конце отчетного периода путем соответствующей обработки данных текущего учета. Отчетность может содержать как количественные, так и качественные показатели. Отчетность представляет собой важнейший источник информации для анализа и принятия решений (см. рис. 2.1 - Пример 1).

1. **Анализ программных средств**

Прежде чем приступить к разработке системы отображения информации необходимо провести анализ программных средств. В программные средства входят выбор операционной системы, языка программирования, среды разработки.

* 1. **Выбор операционной системы**

В дипломном проекте была выбрана операционная система Windows (рис. 3.1 - Windows).



Рис. 3.1 – Windows

Данная операционная система одна из самых распространенных и популярных на данный момент систем, она обладает рядом существенных преимуществ.

Это единый (для всех программ и приложений Windows), простой и стандартизированный интерфейс для взаимодействия с пользователем. Начинающий пользователь персонального компьютера, не будет нуждаться в усилиях для освоения навыков работы с данным программным средством.

ОС Windows имеет программный интерфейс (Windows API) для разработчика программных средств – это большой набор различных функций ядра ОС, предоставляемых для использования в своих проектах и содержащих большой объем кода, являющийся общим для всех программ-приложений Windows.

Операционная система Windows поддерживает сетевое взаимодействие удалённых компьютеров, что позволяет создавать программные приложения, работающие через сеть посредством стандартных протоколов взаимодействия.

Таким образом, учитывая вышеперечисленные положительные свойства данной операционной системы, выбор данной платформы для создания программных средств, считается весьма обоснованным.

* 1. **Выбор языка программирования**

Для реализации алгоритмов, применённых в программе, был выбран язык программирования С++.

В рамках данной дипломной работы основными задачами будут являться: проектирование и разработка системы сортировки входящих писем.

Языки программирования служат двум целям: они дают программисту аппарат для реализации алгоритмов, и формируют концепции, которыми пользуется программист при выполнении поставленных задач. Первой цели идеально отвечает язык, который легко позволяет управлять всеми основными машинными аспектами. С таким умыслом первоначально задумывался C. Второй цели идеально отвечает язык, который настолько «близок к решаемой задаче», чтобы концепции её решения можно было выражать прямо и коротко.

С++ - является уникальным языком программирования. Он позволяет сделать программирование наиболее приятным и простым. За исключением некоторых деталей С++ является объектно-ориентированным расширением языка программирования С.

Основной целью разработчиков C было создание языка, который мог заменить ассемблер при решении задач системного программирования. К достоинствам языка С относятся лаконичность записи алгоритмов, а также стройность написанных программ, их переносимость между компьютерами с различной архитектурой и различными операционными средами. Компиляторы языка разработаны практически для всех существующих в настоящее время платформ. Языка программирования С «доверяет» программисту даже в таком вопросе, как преобразование типов данных. Чаще всего, написанные на этом языке программы, сравнимы по быстродействию с программами, написанными на языке ассемблера. Но программы на С более наглядны и просты в сопровождении. Всё это делает язык С удобным и эффективным средством для решения разнообразных задач.

С++ предоставляет гибкие и эффективные средства определения новых типов. Используя определения новых типов, программист может разделять разрабатываемую программу на легко поддающиеся контролю части. Такой метод построения программ часто называют абстракцией данных. Информация о типах содержится в некоторых объектах типов, определенных пользователем. Такие объекты просты и надежны в использовании в тех ситуациях, когда их тип нельзя установить на стадии компиляции. Программирование с применением таких объектов часто называют объектно-ориентированным. При правильном использовании этот метод дает более

короткие, проще понимаемые и легче контролируемые программы.

Различие между C и С++ состоит в первую очередь в степени внимания, уделяемого типам и структурам. Когда известны типы объектов, компилятор может правильно обрабатывать выражения, тогда как в противном случае программисту пришлось бы максимально подробно описывать каждое действие. Знание типов объектов также позволяет компилятору обнаруживать ошибки, которые в противном случае остались бы до тестирования. Немаловажно то, что использование системы типов для того, чтобы получить проверку параметров функций, защитить данные от случайного искажения, задать новые операции и т.д., само по себе не увеличивает расходов по времени выполнения и памяти. Система типов особенно полезна в нетривиальных задачах.

# Выбор среды разработки

Для реализации поставленной задачи мною была выбрана среда разработки Microsoft Visual Studio , так как она наиболее подходит для создания данного проекта, а именно автоматизированной информационной системы, имеет понятный интерфейс и большой набор функций, инструментов (рис. 3.2 - Microsoft Visual Studio).



Рис. 3.2 - Microsoft Visual Studio

Microsoft Visual Studio — продукт компании Майкрософт, включающих интегрированную среду разработки программного обеспечения и другие инструментальные средства. Данный продукты позволяют разрабатывать консольные приложения, приложения с графическим интерфейсом, с поддержкой технологии Windows Forms, веб-сайты, веб-приложения, веб-службы для всех платформ, поддерживаемых Microsoft Windows, Windows Mobile, Windows CE, .NET Framework, Xbox, Windows Phone .NET Compact Framework и Microsoft Silverlight.

В Visual Studio есть редактор исходного кода с возможностью простейшего редактора кода и поддержкой технологии IntelliSense. Встроенный отладчик может работать как отладчик уровня исходного кода, так и как отладчик машинного уровня. Остальные инструменты включают в себя редактор форм для создания графического интерфейса приложения, веб-редактор, дизайнер классов и дизайнер схемы базы данных.

Visual Studio позволяет создавать и подключать различные сторонние дополнения (плагины) для расширения функциональности, включая добавление поддержки систем контроля версий исходного кода ( например, Subversion и Visual Source Safe), добавление новых наборов инструментов (например, для редактирования и визуального проектирования кода на предметно-ориентированных языках программирования или инструментов для прочих аспектов процесса разработки программного обеспечения.

Каждое обновление программы состоит из новейших инструментов и различных технологий, позволяющих разрабатывать приложения с учетом особенностей и моментов современных платформ. Разработчики могут создавать новые и модернизировать уже существующие приложения, предназначенные для ранних версий ОС Windows. Стоит отметить, что в процессе использования поддерживаемых системой вариантов исходные файлы, проекты и решения в программе Visual Studio будут работоспособными, но исходный код может нуждаться в изменениях.

* 1. **Выбор почтового клиента**

Для автоматической обработки писем и скачивания вложений был выбран почтовый клиент Mozilla Thunderbird (Рис. 3.3 - Mozilla Thunderbird) и расширение AttachmentExtractor.



Рис. 3.3 - Mozilla Thunderbird

Mozilla Thunderbird - свободно распространяемый почтовый клиент, обладающий многими удобными и полезными возможностями. Преимущества Thunderbird – высокая безопасность при работе с электронной почтой, высокая скорость программы, простой и понятный интерфейс. Почтовый клиент имеет фильтры спама и быстрые механизмы поиска сообщений.

Mozilla Thunderbird поддерживает почтовые протоколы POP и IMAP, а также формат HTML почты, функцию импорта настроек электронной почты и сообщений из других почтовых клиентов (Outlook Express, Microsoft Outlook, The Bat!). Есть проверка правописания во время введения текста.

Thunderbird поддерживает шифровку сообщений, цифровые подписи, проверку сертификатов. Вложения не открываются без команды, обеспечивая этим защиту от многих вирусов.

Вся электронная почта, проходящая через Thunderbird, проверяется на спам. Программа имеет свои методы фильтрации нежелательной почты. Thunderbird использует фильтры спама, установленные на сервере поставщика услуг электронной почты.

Механизм автоматических обновлений почтового клиента Mozilla Thunderbird обеспечивает своевременную и быструю актуализацию программы, как в плане повышения безопасности, так и при модификации клиента.

Для работы с большим количеством входящих писем с вложениями было выбрано расширение Attachment Extractor для Mozilla Thunderbird (рис. 3.4 - Attachment Extractor). С помощью Attachment Extractor можно автоматизировать процесс сохранения входящей документации с вложениями в папку на операционную систему.

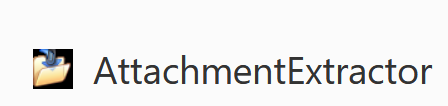


Рис. 3.4 - Attachment Extractor

Основное назначение расширения заключается в сохранении вложений из группы выделенных сообщений. Кроме того, допускается извлечение вложенных файлов из всей почтовой папки.

Attachment Extractor позволяет заносить уникальные имена для вложений. Расширение допускает пометку сообщений как прочитанных, а также их удаление после завершения процедуры извлечения вложенных файлов.

В настройках Attachment Extractor можно указать путь, куда должны по-умолчанию складываться файлы вложений. По-умолчанию документы переносятся в папку загрузок Mozilla Thunderbird.

С помощью данного расширения можно вынести на панель инструментов кнопку вызова извлечения вложенных файлов. Вызов Attachment Extractor осуществляться из контекстного меню. Во время работы расширения на экране отображается диалоговое окно, содержащее информацию о процессе работы. После завершения всех процедур выводится специальное уведомление.

**вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввывывывывуцуцавпвапвапааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыавыааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыаывааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыкуукаааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааа**

**вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввывывывывуцуцавпвапвапааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыавыааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыаывааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааа**

# Информационная модель программы

## Система отображения информации

Система отображения информации - совокупность технических и программных средств, обеспечивающих представление необходимых данных человеку. Система отображения информации является материальной основой для формирования информационной модели и состоит из отдельных, зачастую взаимосвязанных между собой элементов зрительной и акустической индикации.

СОИ классифицируются по ряду признаков:

1) С индивидуальным способом представления информации, с представлением информации в обобщенной форме, с регулируемым потоком информации, с иерархическим способом предъявления информации;

2) Индивидуального или коллективного пользования;

3) по конструктивному исполнению.

Разработка СОИ состоит из нескольких этапов:

- психологический анализ органов местного самоуправления и определение информации, необходимых им для выполнения заданной работы;

- выбор конкретных функции, наиболее полно соответствующих характеру решаемых задач и возможностям по переработки информации;

- композиционное решение и определение конкретной структуры

- разработка, оценка полученных решений и проведение последовательной коррекции структуры для получения результатов.

## Интерфейс

Интерфейс – это визуальная среда управления системой сортировки писем, позволяющая одновременно взаимодействовать  программам для обмена информацией между ними. Интерфейс должен быть простым, надежным, безопасным и эстетичным, так как грамотно разработанный интерфейс с учетом всех требований способствует быстрому принятию решений органами местного самоуправления.

Система должна обеспечивать следующий состав отображаемой информации, указанной в ТЗ:

- кнопку указывающий путь к папке с входящими письмами

- кнопку каталога для сортировки

- количества обработанных документов;

- режим отображения количества писем в данной категории;

- режим обучения системы;

- кнопку сортировки писем

* 1. **Представление информации отображаемой на экране.**
     1. **Зоны программы**

Экран разделен на 4 зоны (рис. 4.1 - распределение рабочего поля экрана):



Рис.4.1 - Распределение рабочего поля экрана

**Зона №1** служит для отображения системных кнопок приложения и указания пути к папке с входящими письмами.

**Зона №2** служит информационным полем по статистике обработанных входящих писем. В данной зоне указывается структура входящих писем.

**Зона №3** служит для управления сортировки писем по папкам и включения режима обучения.

* + 1. **Информация, отображаемая в системе сортировки**

Визуальная среда управления системой сортировки писем (рис. 4.2 - вид экрана)

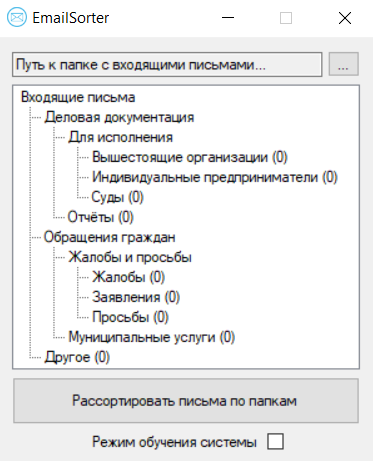


Рис. 4.2 Вид экрана

Отображение зоны №1 (рис. 4.3 - Зона 1).

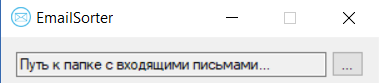


Рис. 4.3 Зона 1

Отображение зоны №2 (рис. 4.4 - Зона 2).

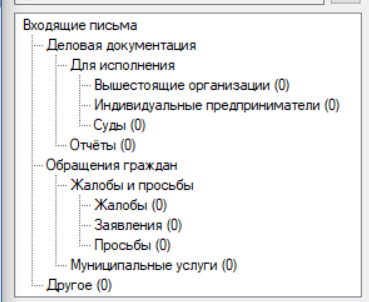


Рис. 4.4 Зона - 2

Отображение зоны №3 (рис. 4.5 - Зона 3).

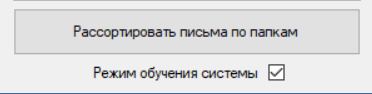


Рис. 4.5 - Зона 3

* + 1. **Информация, отображаемая в зоне №1.**
       1. **Папка с входящими письмами**

Каталог с входящей документацией (рис. 4.6 - Папка с входящими письмами).

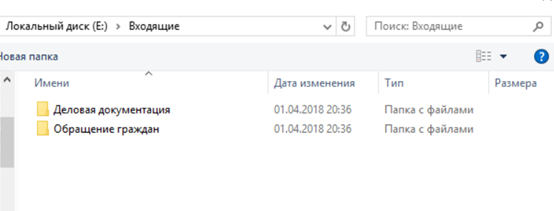


Рис. 4.6 - Папка с входящими письмами

В данной функции можно выбрать папку с входящими письмами или прописать путь к данной папке.

**Кнопки панели управления**

Наименование и функции кнопок:

«…» - Указывается путь к папке с входящими письмами;

« \_ » - Свернуть программу;

« Х » - Выход из программы;

**вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввывывывывуцуцавпвапвапааааааааааааааааа**

* + 1. **Информация, отображаемая в зоне №2**

В зоне №2 отображается структура входящих писем в виде дерева (рис. 4.7 Схема входящих писем).

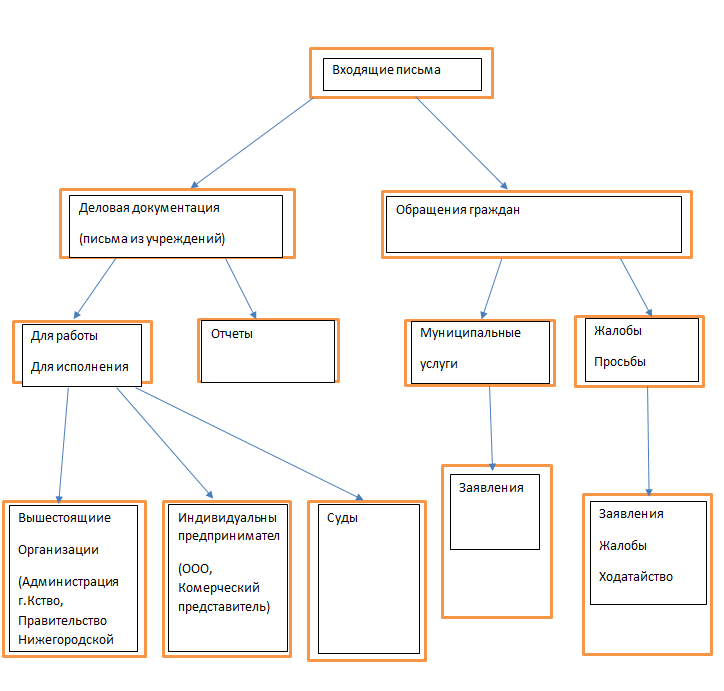


Рис. 4.7 - Схема входящих писем

* + 1. **Информация отображаемая №3.**

В зоне №3 отображается панель управления с кнопками выбора и режима обучения.

**Кнопки панели управления**

Наименование и функции кнопок:

«Рассортировать письма по папкам» - Идет сортировка писем по папкам;

«Режим обучения» - Включение режима обучения для программы;

**ввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввв**

1. **Байесовский классификатор**

Данный пункт содержит описание байесовского классификатора, который содержится в данной системе. В настоящее время статистические методы широко применяются для классификации текстов по признакам жанрового, авторского, гендерного и других стилей. Байесовская теория принятия решений составляет основу статистического подхода к задаче классификации объектов. Этот подход основан на предположении, что задача выбора решения сформулирована в терминах теории вероятностей и известны все представляющие интерес вероятностные величины. В основе байесовской классификации лежит правило Байеса.

Вероятность - это численная мера степени объективной возможности наступления случайного события. Она может быть равна значению от 0 до 1. Она также приближенно выражена посредством пропорции числа случаев на протяжении всего эксперимента. Если вероятность события равна нулю, то оно не произойдет. Тем не менее, если она равна 1, событие обязательно произойдет.

Предмет теории вероятностей. Для описания закономерной связи между некоторыми условиями S и событием А, наступление или не наступление которого при данных условиях может быть точно установлено

а) при каждом осуществлении условий S наступает событие А. Такой вид, например, имеют все законы классической механики, которые утверждают, что при заданных начальных условиях и силах, действующих на тело или систему тел, движение будет происходить однозначно определённым образом.

б) При условиях S событие А имеет определённую вероятность P (A / S. Так, например, законы радиоактивного излучения утверждают, что для каждого радиоактивного вещества существует определённая вероятность того, что из данного количества вещества за данный промежуток времени распадётся какое-либо число N атомов.

Статистические закономерности, то есть закономерности, описываемые схемой типа (б), были впервые обнаружены на примере азартных игр, подобных игре в кости. Очень давно известны также статистические закономерности рождения, смерти. Конец 19 в. и 1-я половина 20 в. отмечены открытием большого числа статистических закономерностей в физике, химии, биологии и т.п.

Возможность применения методов теории вероятностей к изучению статистических закономерностей, относящихся к весьма далёким друг от друга областям науки, основана на том, что вероятности событий всегда удовлетворяют некоторым простым соотношениям. Изучение свойств вероятностей событий на основе этих простых соотношений и составляет предмет теории вероятностей.

* 1. **Теорема Байеса**

Теорема Байеса (формула Байеса) — одна из основных теорем элементарной теории вероятностей, позволяющая определить вероятность любого события при условии, что было произведено другое статистически взаимозависимое с ним событие. По формуле Байеса можно, наиболее точно определить вероятность, приняв в расчет ранее известную информацию, так и данные новых наблюдений. Формула Байеса может быть выведена из основных аксиом теории вероятностей. Особенность данной теоремы заключается в том, что для её применения требуется очень большое количество вычислений, расчетов, поэтому байесовские оценки стали в основном использовать только после революции в компьютерных и сетевых технологиях.

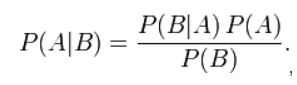
При возникновении теоремы Байеса вероятности, используемые в теореме, подвергались целому ряду вероятностных интерпретаций. В одной из интерпретаций сообщалось, что применение особого подхода к статистическому анализу напрямую связано с выводом формулы. При использовании байесовской интерпретации вероятности, теорема показывает, как личный уровень доверия может изменяться вследствие количества наступивших событий. В этом заключаются выводы Байеса, которые стали основополагающими для байесовской статистики. Теорема используется не только в байесовском анализе, но так же и активно применяется для большого ряда других расчетов.

Эксперименты психологов показали, что люди неверно оценивают вероятность события, на основе своего полученного опыта (апостериорная вероятность), игнорируя саму вероятность предположения (априорная вероятность). Исходя из этого правильный результат по формуле Байеса может значительно отличаться от интуитивно.

Теорема Байеса названа в честь Томаса Байеса (1702 — 1761) — английского математика и священника, который предложил использование данной теоремы для корректировки убеждений, основываясь на обновлённых данных. Его работа «An Essay towards solving a Problem in the Doctrine of Chances» была опубликована в 1763 году, через 2 года после смерти автора. До того, как посмертная работа Байеса была принята и прочитана в Королевском обществе, она была отредактирована и обновлена Ричардом Прайсом. Однако все эти идеи не предавались публичной огласке до тех пор, пока не были вновь открыты и развиты Лапласом, впервые опубликовавшим современную формулировку теоремы в своей книге 1812 года «Аналитическая теория вероятностей».

Формула Байеса:

ввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввв



P(A) — априорная вероятность гипотезы A;

P(A | B) — вероятность гипотезы A при наступлении события B (апостериорная вероятность);

P(B | A) — вероятность наступления события B при истинности гипотезы A;

P(B) — вероятность наступления события B.

ввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввв

Формула Байеса позволяет «переставить причину и следствие»: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной.

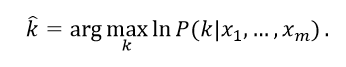
ввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввв

События, отражающие действие «причин», в данном случае обычно называют гипотезами, так как они — предполагаемые события, повлекшие данное. Безусловную вероятность справедливости гипотезы называют априорной (насколько вероятна причина вообще), а условную - с учетом факта произошедшего события — апостериорной.

вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввв

* 1. **Общая структура байесовского классификатора**

В основе классификатора лежит следующее правило. Классификатор вычисляет апостериорную вероятность 𝑃(𝑘|𝑥) каждого класса 𝑘, которому



может принадлежать испытуемый объект, и относит этот объект к апостериорно наиболее вероятному классу 𝑘 ̂:

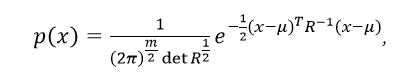


где 𝑃(𝑘) – априорная вероятность того, что объект относится к -ому классу, 𝑝(𝑘) и 𝑝(𝑥1,…,𝑥𝑚|𝑘) - безусловная и условная многомерные плотности распределения вектора признаков, компоненты которого обычно статистически зависимы.

Таким образом, байесовский классификатор предполагает, что многомерная совместная плотность распределения признаков известна для всех классов.

Аналитическое представление многомерной плотности вероятности известно только для нормального распределения. Вместе с тем многомерная нормальная плотность распределения дает подходящую модель для одного важного случая, а именно когда значения векторов признаков 𝑥 для данного класса 𝑘 представляются непрерывнозначными, слегка искаженными версиями единственного типичного вектора, или вектора -прототипа, 𝜇𝑘. Именно этого ожидают, когда классификатор выбирается так, чтобы выделять те признаки, которые, будучи различными для образов, принадлежащих различным классам, были бы, возможно, более схожи для образов из одного и того же класса.

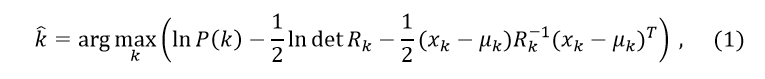
Многомерная нормальная плотность распределения в общем виде представляется выражением



где µ – 𝑚-компонентный вектор среднего значения, 𝑅 – ковариационная матрица размера 𝑚 × 𝑚, 𝑇 – знак транспонирования.

Отметим, что если все недиагональные элементы равны нулю, то 𝑝(𝑥) сводится к произведению одномерных нормальных плотностей компонент вектора 𝑥.

Поэтому для многомерного нормального распределения удаётся выразить в аналитически замкнутой форме (с точностью до несущественных слагаемых) алгоритм байесовской классификации:



где 𝜇𝑘 – 𝑚-вектор-строка математических ожиданий значений признаков объектов класса 𝑘, 𝑅𝑘 – 𝑚 × 𝑚-матрица ковариаций векторов признаков класса 𝑘.

Диагональные элементы матрицы образуют -вектор 𝐷𝑘 дисперсий признаков объектов класса 𝑘.

* 1. **Наивный байесовский классификатор**

В наивном байесовском классификаторе делается предположение о независимости признаков объекта. Если пренебречь статистическими связями между компонентами вектора признаков, тогда матрица 𝑅𝑘 будет диагональной с вектором 𝐷𝑘 на главной диагонали и классификатор (1) станет наивным байесовским классификатором.

Также предполагается, что маргинальная плотность распределения 𝑝(𝑥𝑗|𝑘) любого признака является нормальной для любого класса.

Но на практике так бывает далеко не всегда, то есть наблюдаемые данные не подчиняются нормальному закону распределения (в общем случае закон вообще неизвестен) и имеет место статистическая зависимость, поэтому область применения классификатора сужается.

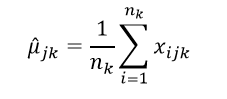
* 1. **Оптимальный байесовский классификатор**

Так как оптимальный байесовский классификатор является модификацией наивного байесовского классификатора, то в качестве решающего правила также берётся формула (1).

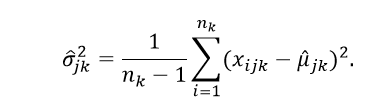
Основная идея состоит в том, чтобы, максимально используя обучающую выборку и гауссову копула-функцию, обойти два «наивных» предположения. Модификация позволяет, во-первых, учесть статистические связи между наблюдаемыми признаками. И во-вторых, адаптировать классификатор к неизвестному действительному распределению путем приведения сглаженных маргинальных функций распределения признаков к нормальному виду. Другими словами, с помощью нелинейных гауссовых копула-функций негауссовы данные преобразуются в гауссовы, которые можно подавать на вход классификатору.

Рассмотрим алгоритм байесовской классификации с обучением, который состоит из двух этапов:

1. Этап обучения. Если векторы математических ожиданий и дисперсий и матрицы 𝑅𝑘 неизвестны, то по обучающей выборке 𝑥𝑖𝑘 = (𝑥𝑖1𝑘,…,𝑥𝑖𝑗𝑘,…,𝑥𝑖𝑚𝑘),𝑖 = 1,𝑛𝑘 ̅̅̅̅̅̅,𝑘 = 1,𝐾 ̅̅̅̅̅, для каждого класса 𝑘 строятся эмпирические оценки математического ожидания



и дисперсии:



Затем находятся эмпирические оценки ковариационной матрицы для каждого класса 𝑘:

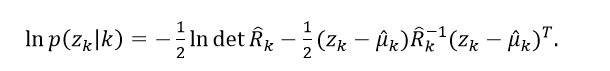


где 𝜇̂𝑘 = (𝜇̂1𝑘,…,𝜇̂𝑚𝑘) – вектор-строка эмпирических оценок математических ожиданий значений признаков, 𝑇 - знак транспонирования.

По найденным оценкам предоставляется возможным обучить байесовский классификатор (1) распознавать классы объектов, заменив в нем неизвестные 𝜇𝑘 и 𝑅𝑘 на их оценки.

2. Этап классификации.

Наблюдаемый вектор-строка признаков (𝑥1,…,𝑥𝑚) испытуемого объекта преобразуется по формуле в нормально распределённые векторы 𝑧𝑘 = (𝑧1𝑘,…,𝑧𝑗𝑘,…,𝑧𝑚𝑘), 𝑘 = 1,𝐾 ̅̅̅̅̅ (для каждой гипотезы 𝑘 о классе), по которым вычисляется логарифм функции правдоподобия каждого класса:



В качестве класса, к которому принадлежит испытуемый объект, выбирается максимально правдоподобный класс k ̂ в соответствии с решающим правилом для эквивалентных признаков z.

Вышеописанный алгоритм позволяет построить обученный и адаптированный к эмпирическому распределению наблюдений байесовский классификатор, который является оптимальным.

**вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввывывывывуцуцавпвапвапааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыавыааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыаывааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыкуукаааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааа**

1. **Сетевая папка**

В настоящее время локальные вычислительные (ЛВС) получили очень широкое распространение. Локальные сети позволяют использовать почтовый ящик для передачи сообщений на другие компьютеры, что позволяет в наиболее короткий срок передавать документы с одного компьютера на другой. Локальные сети, при наличии специального программного обеспечения, служат для организации совместного использования файлов. Сеть позволяет (зарегистрированным на файл-сервере) пользователям получать доступ к той информации, к которой их допускает оператор сети.

* 1. **Настройка сети**

Необходимо открыть "Проводник Windows" и открыть пункт "Сеть" (Рис - 6.1 Сеть).

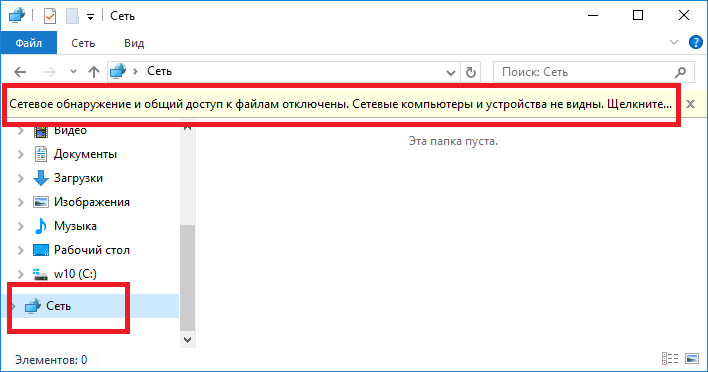


Рис. 6.1 - Сеть

Нужно выбрать пункт "Включить сетевое обнаружение и общий доступ к файлам" (см. рис. 6.2 - Сетевое обнаружение):

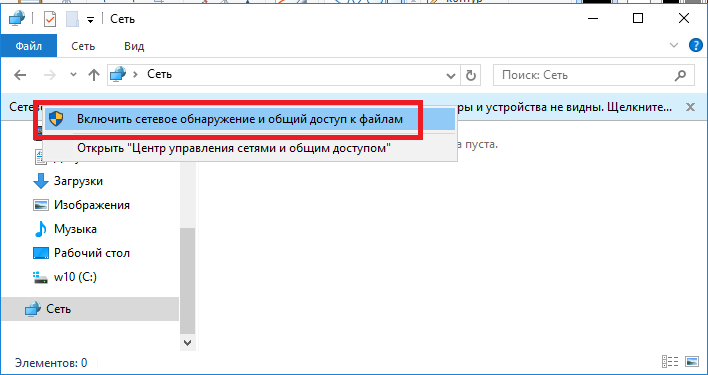


Рис. 6.2 - Сетевое обнаружение

После этого "Проводник Windows" выдаст запрос на выбор типа сети, там нужно выбрать первый вариант (см. рис. 6.3 - Сетевое обнаружение и общий доступ к файлам):

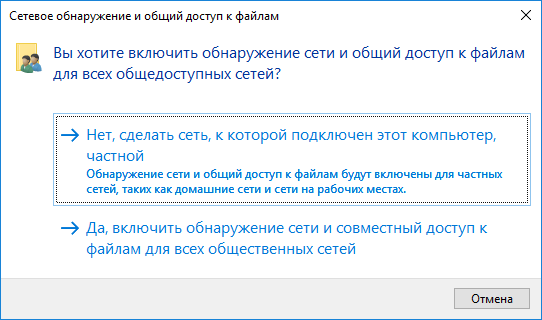


Рис. 6.3 - Сетевое обнаружение и общий доступ к файлам

**вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввывывывывуцуцавпвапвапааааааааааааааааа**

* 1. **Общий доступ к папке**

Для создания сетевого ресурса на компьютере, необходимо создать папку «Отсортированные письма» и «Входящие письма» . Далее необходимо открыть свойства папки (рис. 6.4 - Свойства).

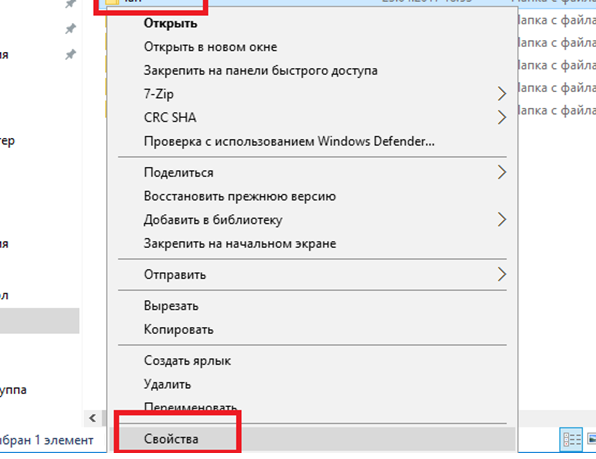


Рис. 6.4 - Свойства

Необходимо открыть общий доступ (рис. 6.5 - Общий доступ).

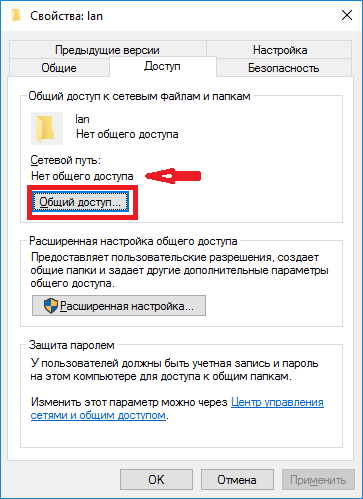


Рис. 6.5 - Общий доступ

В следующем окне нужно открыть список локальных пользователей (учетные записи на этом компьютере) (рис. 6.6 - Общий доступ):

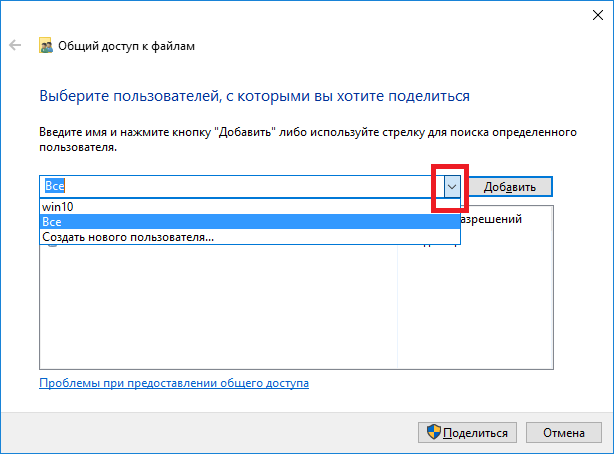


Рис. 6.6 - Общий доступ

После этого нажать кнопку "Добавить".

Далее нужно, для группы "Все", указать права доступа на чтение и запись (Рис. 6.7 - Общий доступ):

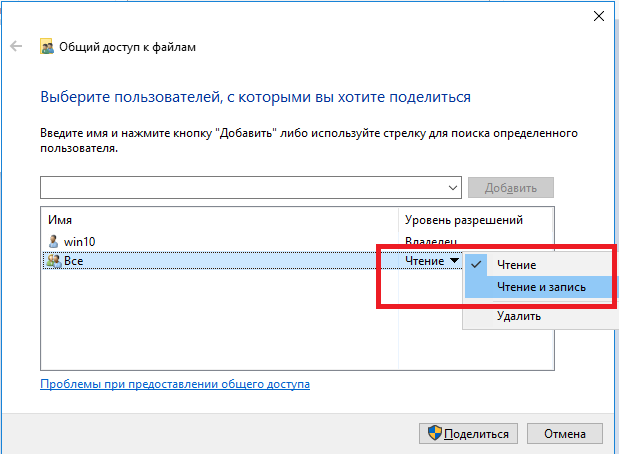


Рис. 6.7 - Общий доступ

После необходимо открыто окно "Свойства папки". В нем открыть вкладку "Безопасность", там должен быть полный доступ для группы "Все" (Windows автоматически изменяет права файловой системы NTFS) (рис. 6.8 - Свойства):

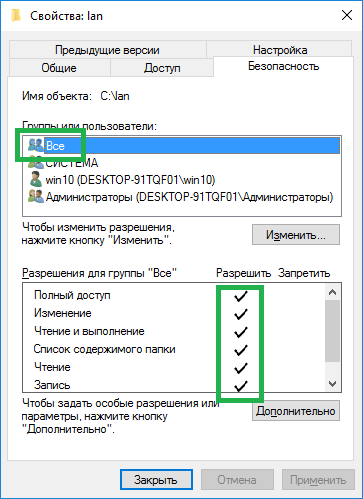


Рис. 6.8 - Свойства

Папки «Отсортированные письма» и «Входящие письма» готовы для сетевого использования.

Далее необходимо открыть "Центр управления сетями и общим доступом" (см. рис. 6.9 - Центр управления сетями и общим доступом):

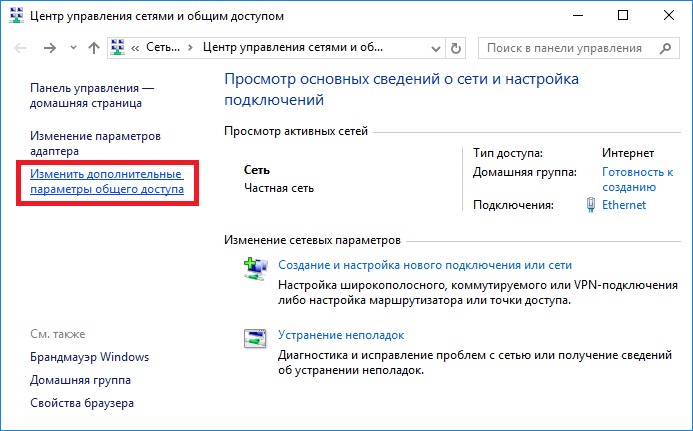


Рис. 6.9 - Центр управления сетями и общим доступом

Нужно отключить параметр "общий доступ с парольной защитой" "Сохранить изменения" (рис. 6.10 - Центр управления сетями):

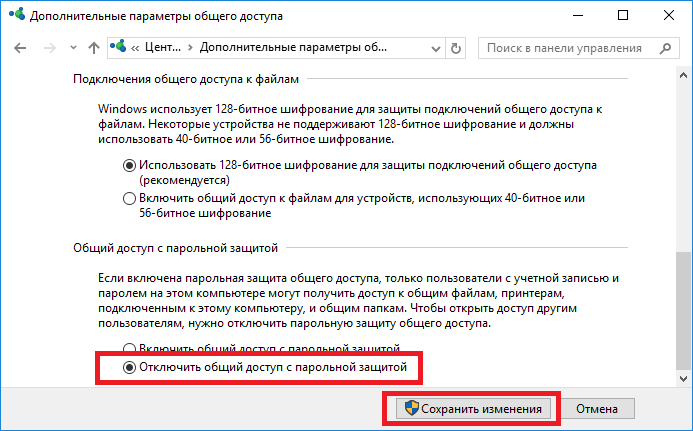


Рис. 6.10 - Центр управления сетями

В ходе выполнения данных действий была создана сетевая папка отсортированных писем с общим доступам всех пользователей сети.

**вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввывывывывуцуцавпвапвапааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыавыааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыаывааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыкуукаааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааа**

**вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввв**

1. **Примеры**
   1. **Сортировка писем**

Запускаем на главном компьютере почтовый клиент mozilla thunderbird и производим начальную настройку (рис. 8.1 - mozilla thunderbird).

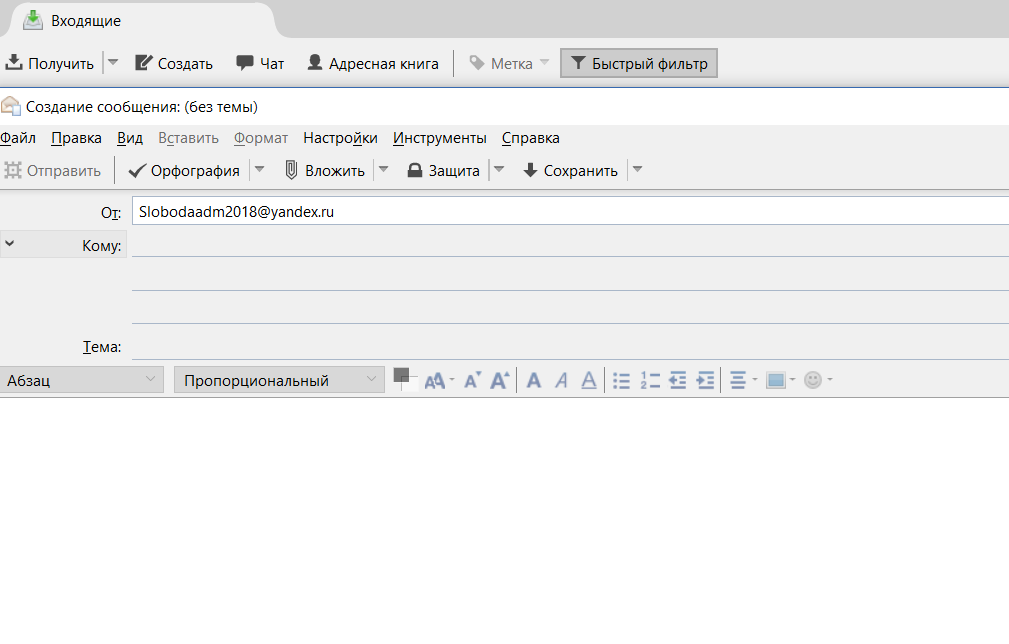


Рис. 8.1 - mozilla thunderbird

Необходимо извлечь все вложения в каталог на жестком диске компьютера с помощью дополнения AttachmentExtractor (рис. 8.2 - AttachmentExtractor).

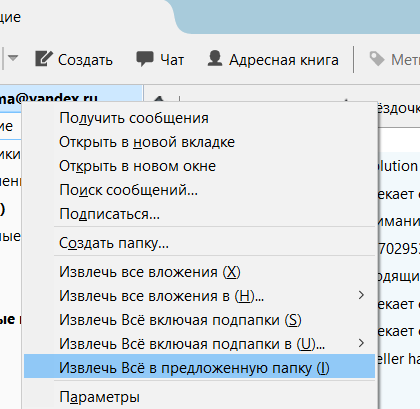


Рис. 8.2 - AttachmentExtractor

В корневом диске находится папка с входящими письмами (рис. 8.1 - Папка с входящими письмами).

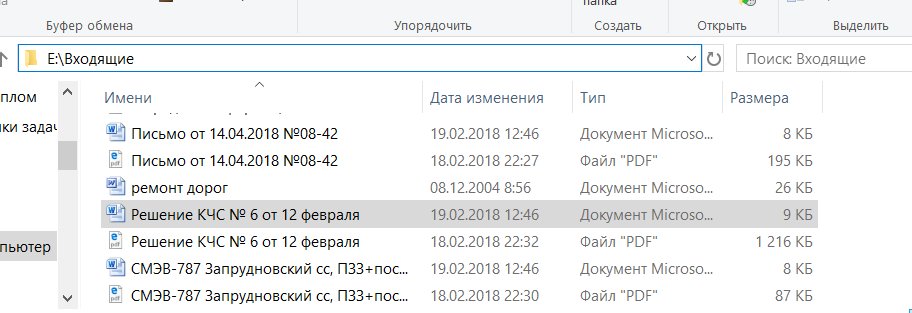


Рис. 8.1 - Папка с входящими письмами

Запускаем программу для сортировки писем на рабочем столе (рис. 8.2 - EmailSorter).



Рис. 8.2 - EmailSorter

Далее указываем путь, где хранятся входящие письма, нажав на кнопу  (рис. 8.3 - Путь к входящим письмам; рис. 8.4 - входящие письма)

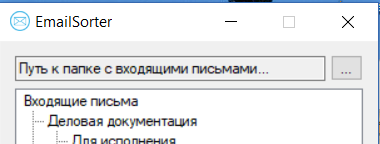


Рис. 8.3 - Путь к входящим письмам

**вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввывывывывуцуцавпвапвапааааааааааааааааа**

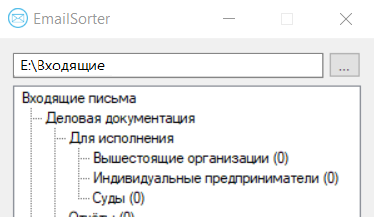


Рис. 8.4 - входящие письма

Далее необходимо нажать на кнопку «режима обучения системы», и «рассортировать письма по папкам» (рис. 8.5 - Сортировка писем по папкам).

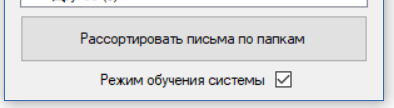


Рис. 8.5 - Сортировка писем по папкам

После сортировки отобразится информация по выполненной работе. Сколько писем находится в конкретной категории (Рис. 8.6 - Статистика писем).

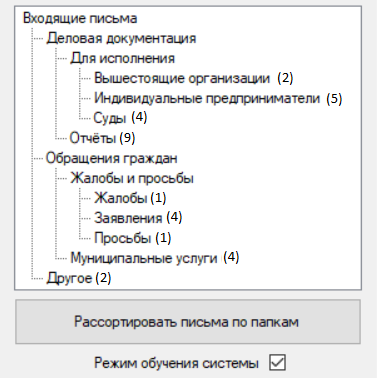


Рис. 8.6 - Статистика писем

После обработки письма будут рассортированы по папкам. В примеры указаны уже отсортированные письма в каждой из папок (рис. 8.7 - Пример 1, рис. 8.8 - Пример 2, рис. 8.9 - Пример 3).

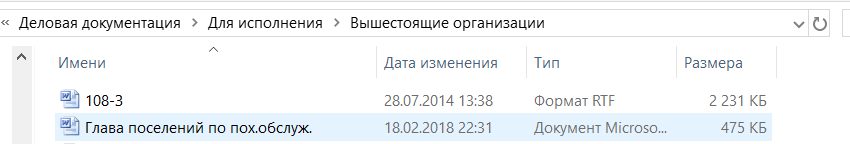


Рис. 8.7 - Пример 1

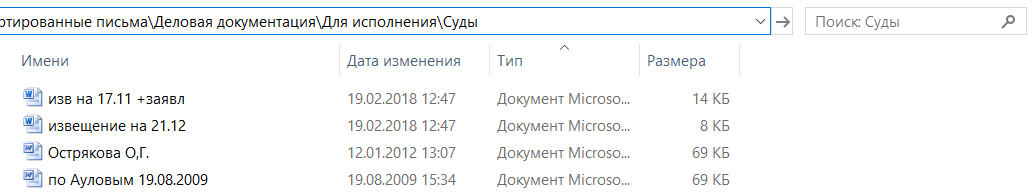


Рис. 8.8 - Пример 2

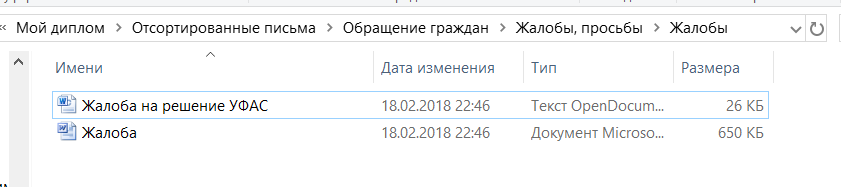


Рис. 8.9 - Пример 3

**вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввваааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыаываа**

1. **Заключение**

Результатом данной работы является создание системы автоматического анализа писем. Данная система удовлетворяет поставленным требованиям, и поддерживает следующие возможности:

- Прием, обработку и сортировки входящей документации в раздел по темам;

- Программа должна иметь режим обучения.

Программа проста в использовании, не требует от пользователя каких-либо специальных навыков работы с электронно-вычислительными машинами.

**ввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввывывывывуцуцавпвапвапааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыавыааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыаывааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааавыкуукаааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааааа**

**вввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввввывывывывуцуцавпвапвапааааааааааааааааа**