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

Институт Физико-технический
Направление подготовки: Прикладная математика и информатика
Кафедра высшей математики и математической физики

БАКАЛАВРСКАЯ РАБОТА

Тема работы
Оценка релевантности текста для географической диверсификации компании

УДК 005.591.61

Студент

Группа	ФИО	Подпись	Дата
0В31	Булыгин Лев Эдуардович		02.06.17

Руководитель

Должность	ФИО	Ученаястепень, звание	Подпись	Дата
Доцент	Семенов М.Е.	к. ф.-м. н.		8.6.17

КОНСУЛЬТАНТЫ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученаястепень, звание	Подпись	Дата
Доцент	Верховская М.В.	к. э. н.		16.05.17.

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор	Федорчук Ю.М	д. т. н		18.05.17г.

ДОПУСТИТЬ К ЗАЩИТЕ:

Зав. кафедрой	ФИО	Ученая степень, звание	Подпись	Дата
ВММФ	Трифонов А.Ю.	д. ф.-м. н.		09.06.17

Томск – 2017 г.

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ООП

Код результата	Результат обучения (выпускник должен быть готов)
Профессиональные компетенции	
ПК-1	К самостоятельной работе
ПК-2	Использовать современные прикладные программные средства и осваивать современные технологии программирования
ПК-3	Использовать стандартные пакеты прикладных программ для решения практических задач на ЭВМ, отлаживать, тестировать прикладное программное обеспечение
ПК-4	Настраивать, тестировать и осуществлять проверку вычислительной техники и программных средств
ПК-5	Демонстрировать знание современных языков программирования, операционных систем, офисных приложений, Интернета, способов и механизмов управления данными; принципов организации, состава и схемы работы операционных систем
ПК-6	Решать проблемы, брать на себя ответственность
ПК-7	Проводить организационно-управленческие расчеты, осуществлять организацию и техническое оснащение рабочих мест
ПК-8	Организовывать работу малых групп исполнителей
ПК-9	Определять экономическую целесообразность принимаемых технических и организационных решений
ПК-10	Владеть основными методами защиты производственного персонала и населения от возможных последствий аварий, катастроф, стихийных бедствий
ПК-11	Знать основные положения законы и методы естественных наук; выявлять естественнонаучную сущность проблем, возникающих в ходе профессиональной деятельности, использовать для их решения соответствующий естественнонаучный аппарат
ПК-12	Применять математический аппарат для решения поставленных задач, способен применять соответствующую процессу математическую модель и проверять ее адекватность

ПК-13	Применять знания и навыки управления информацией
ПК-14	Самостоятельно изучать новые разделы фундаментальных наук
Универсальные компетенции	
ОК-1	Владеть культурой мышления, иметь способности к обобщению, анализу, восприятию информации, постановке цели и выбору путей ее достижения
ОК-2	Логически верно, аргументировано и ясно строить устную и письменную речь
ОК-3	Уважительно и бережно относится к историческому наследию и культурным традициям, толерантно воспринимать социальные и культурные различия; понимать движущие силы и закономерности исторического процесса, роль насилия и ненасилия в истории, место человека в историческом процессе, политической организации общества
ОК-4	Понимать и анализировать мировоззренческие, социально и личностно значимые философские проблемы
ОК-5	Владеть одним из иностранных языков на уровне бытового общения, а также переводить профессиональные тексты с иностранного языка

ОК-6	К кооперации с коллегами, работе в коллективе
ОК-7	Находить организационно-управленческие решения в нестандартных ситуациях и готов нести за них ответственность
ОК-8	Использовать нормативно-правовые документы в своей деятельности
ОК-9	Стремиться к саморазвитию, повышению своей квалификации и мастерства
ОК-10	Осознавать социальную значимость своей будущей профессии, обладать высокой мотивацией к выполнению профессиональной деятельности
ОК-11	Использовать основные положения и методы социальных, гуманитарных и экономических наук при решении социальных и профессиональных задач
ОК-12	Анализировать социально значимые проблемы и процессы
ОК-13	Использовать основные законы естественнонаучных дисциплин в профессиональной деятельности, применять методы математического анализа и моделирования, теоретического и экспериментального исследования
ОК-14	Понимать сущность и значение информации в развитии современного информационного общества, осознавать опасности и угрозы,

	возникающие в этом процессе, соблюдать основные требования информационной безопасности, в том числе защиты государственной тайны
ОК-15	Оформлять, представлять и докладывать результаты выполненной работы
ОК-16	Создавать и редактировать тексты профессионального назначения
ОК-17	Использовать для решения коммуникативных задач современные технические средства и информационные технологии
ОК-18	Владеть средствами самостоятельного, методически правильного использования методов физического воспитания и укрепления здоровья, быть способным к достижению должного уровня физической подготовленности для обеспечения полноценной социальной и профессиональной деятельности

Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Институт Физико-технический
Направление подготовки (специальность) Прикладная математика и информатика
Кафедра Высшей математики и математической физики

УТВЕРЖДАЮ:
Зав. кафедрой

_____ 09.06.17 Трифонов А.Ю.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Дипломной работы

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

Студенту:

Группа	ФИО
0В31	Булыгину Льву Эдуардовичу

Тема работы:

«Оценка релевантности текста для географической диверсификации компании»	
Утверждена приказом директора (дата, номер)	

Срок сдачи студентом выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Финансовые годовые отчеты мировых компаний в формате pdf. Полный список в приложении А.

Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	Используя различные методы обработки текста: 1) TF-IDF, 2) doc2vec построены модели на основе классификаторов: 1) наивный байесовский классификатор, 2) логистическая регрессия, 3) градиентный бустинг над решающими деревьями, 4) рекуррентные нейронные сети.

Консультанты по разделам выпускной квалификационной работы

(с указанием разделов)

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Верховская М.В.
Производственная и экологическая безопасность	Федорчук Ю.М.

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент каф. ВМиМФ	Семенов М.Е.	Кандидат ф-м. наук		10.03.17

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
0В31	Булыгин Лев Эдуардович		10.03.17

ЗАДАНИЕ ДЛЯ РАЗДЕЛА

«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
0В31	Булыгину Льву Эдуардовичу

Институт	ФТИ	Кафедра	ВММФ
Уровень образования	Бакалавр	Направление/специальность	Прикладная математика и информатика

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих	*1.Стоимость расходных материалов* *2.Стоимость расхода электроэнергии* *3.Норматив заработной платы*
2. Нормы и нормативы расходования ресурсов	*1.Тариф на электроэнергию* *2.Коэффициенты для расчета заработной платы*
3. Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования	*1.Отчисления во внебюджетные фонды (27,1%)* *2.Расчет дополнительной заработной платы (12%)*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Оценка коммерческого потенциала, перспективности и альтернатив проведения НИ с позиции ресурсоэффективности и ресурсосбережения	*1.Потенциальные потребители результатов исследования;* *2.Анализ конкурентных технических решений;* *3.SWOT – анализ.*
2. Планирование и формирование бюджета научных исследований	*1.Структура работ в рамках научного исследования;* *2.Определение трудоемкости выполнения работ и разработка графика проведения научного исследования;* *3.Бюджет научно - технического исследования (нти).*
3. Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования	*1.Определение интегрального финансового показателя разработки;* *2.Определение интегрального показателя ресурсоэффективности разработки;* *3.Определение интегрального показателя эффективности*
Перечень графического материала *(с точным указанием обязательных чертежей)***:**	

1. *Оценка конкурентоспособности технических решений* 2. *Матрица SWOT* 3. *Альтернативы проведения НИ* 4. *График проведения и бюджет НИ* 5. *Оценка ресурсной, финансовой и экономической эффективности НИ*
Дата выдачи задания для раздела по линейному графику

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Верховская Марина Витальевна	Доцент, кандидат экономических наук		06.02.17

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
0В31	Булыгин Лев Эдуардович		06.02.17

ЗАДАНИЕ ДЛЯ РАЗДЕЛА

«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
0В31	Булыгину Льву Эдуардовичу

Институт	ФТИ	Кафедра	ВММФ
Уровень образования	Бакалавр	Направление/специальность	Прикладная математика и информатика

Тема дипломной работы: Оценка релевантности текста для географической диверсификации компании

Исходные данные к разделу «Социальная ответственность»:

1.Целью данной работы является создание математической модели оценки релевантности текста с максимальной точностью.

2. Описание рабочего места на предмет возникновения:

- *вредных проявлений факторов производственной среды (освещение, шумы, вибрации, электромагнитные поля, ионизирующие излучения)*
- *опасных проявлений факторов производственной среды (механической природы, термического характера, электрической, пожарной и взрывной природы)*

Перечень вопросов, подлежащих исследованию, проектированию и разработке:

1. Анализ выявленных вредных факторов проектируемой производственной среды в следующей последовательности:

- *приводятся данные по оптимальным и допустимым значениям микроклимата на рабочем месте, перечисляются методы обеспечения этих значений; приводится расчет освещенности на рабочем месте;*
- *приводятся данные по реальным значениям шума на рабочем месте и мероприятия по защите персонала от шума, при этом приводятся значения ПДУ, средства коллективной защиты, СИЗ;*
- *приводятся данные по реальным значениям электромагнитных полей на рабочем месте, в том числе от компьютера или процессора, перечисляются СКЗ и СИЗ;*
- *приведение допустимых норм с необходимой размерностью (с ссылкой на соответствующий нормативно-технический документ);*
- *предлагаемые средства защиты (сначала коллективной защиты, затем – индивидуальные защитные средства)*

2. Анализ выявленных опасных факторов проектируемой произведённой среды в следующей последовательности

- *приводятся данные по значениям напряжения используемого оборудования, классификация помещения по электробезопасности, допустимые безопасные для человека значения напряжения, тока и заземления (в т.ч. статическое электричество, молниезащита - источники, средства защиты); перечисляются СКЗ и СИЗ;*
- *приводится классификация пожароопасности помещений, указывается класс пожароопасности помещения, перечисляются средства пожарообнаружения и принцип их работы, средства пожаротушения, принцип работы, назначение, маркировка;*
- *пожаровзрывобезопасность (причины, профилактические мероприятия).*

3. Охрана окружающей среды:

– *анализ воздействия при работе на ПЭВМ на атмосферу, гидросферу, литосферу;* – *наличие отходов (бумага, картриджи, компьютеры и т. д.);* – *методы утилизации отходов.*
4. Защита в чрезвычайных ситуациях: – *Приводятся возможные для Сибири ЧС; Возможные ЧС: морозы, диверсия* – *разрабатываются превентивные меры по предупреждению ЧС;* – *разработка действий в результате возникшей ЧС и мер по ликвидации её последствий*
5. Правовые и организационные вопросы обеспечения безопасности: – *Специальные (характерные для проектируемой рабочей зоны) правовые нормы трудового законодательства: СанПиН 2.2.2.542-96; СанПин 2.2.2.542-96; СанПиН 2.2.2/2.4.1340-03; СНиП-23-05-95; Сан.ПиН 2.2.2. 542 – 96; ГОСТ 12.1.036-96;ГОСТ 12.1.012-96; ГОСТ 12.1.004-76; ГОСТ 12.1.010-76; ГОСТ 12.1.013-78.*
Перечень графического материала: *1) Пути эвакуации* *2) План размещения светильников на потолке рабочего помещения*

Дата выдачи задания для раздела по линейному графику	10.03.17г

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор	Федорчук Юрий Митрофанович	д.т.н.		10.03.17г

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
0В31	Булыгин Лев Эдуардович		10.03.17

РЕФЕРАТ

Выпускная квалификационная работа содержит 93 страницы, 8 рисунков, 19 таблиц, 18 источников литературы, 5 приложений.

Ключевые слова: машинное обучение, анализ текста, географическая диверсификация, бинарная классификация, линейный классификатор, градиентный бустинг, TF-IDF, word2vec, doc2vec, оптимизация параметров обучения.

Объект исследования: 32 годовых финансовых отчетов мировых компаний за 2013-2016 г.

Цель работы: Создание математической модели оценки релевантности текста для географической диверсификации компании с приемлемой точностью.

Актуальность: Машинное обучение эффективно используется для автоматизации решения интеллектуальных задач, что позволяет снизить издержки, сократить объем рутинных операций.

Методы проведения исследования: теоретические (изучение литературы, обзор существующих методов и моделей анализа) и практическое применение методов машинного обучения для построения модели.

Полученные результаты: Сформулированы критерии релевантности: 1) явное указание, 2) контактная информация, 3) логический вывод из текста 4) логический вывод из числовой информации. С применением различных методов обработки текста: а) TF-IDF, б) word2vec, в) doc2vec построены модели на основе классификаторов: 1) наивный байесовский классификатор, 2) логистическая регрессия, 3) градиентный бустинг над решающими деревьями.

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ, СОКРАЩЕНИЯ, НОРМАТИВНЫЕ ССЫЛКИ

НОРМАТИВНЫЕ ССЫЛКИ

В настоящей работе использованы ссылки на следующие стандарты:

1. ГОСТ Р 1.5 – 2012 Стандартизация в Российской Федерации. Стандарты национальные в Российской Федерации. Правила построения, изложения, оформления и обозначения.
2. ГОСТ 7.1 – 2003 Система стандартов по информации, библиотечному и издательскому делу. Библиографическая ссылка.
3. ГОСТ 12.4.011-75 Система стандартов безопасности труда. Средства защиты работающих. Классификация.
4. ГОСТ 12.1.012-96 Вибрационная безопасность. Общие требования.
5. ГОСТ 12.1.036-81 Система стандартов безопасности труда. Шум. Допустимые уровни в жилых и общественных зданиях.
6. ГОСТ 12.0.002-80 Система стандартов безопасности труда. Термины и определения.
7. ГОСТ 12.1.038-82 Система стандартов безопасности труда. Электробезопасность. Предельно допустимые значения напряжений прикосновения и токов.
8. ГОСТ 12.1.004-91 Пожарная безопасность. Общие требования.
9. ГОСТ 12.1.010-76 Взрывобезопасность. Общие требования.

ОПРЕДЕЛЕНИЯ

В работе использованы термины с соответствующими определениями:

Географическая диверсификация компании – наличие филиалов компании в какой-либо географической точке. Под **географической точкой** будем понимать страну, область, субъект, город, городской адрес или какие-нибудь места, которые позволяют идентифицировать местность, например, Силиконовая долина.

Машинное обучение – это обучение какой-либо математической модели по прецедентам. В нашем случае **прецедентами** являются примеры релевантного текста и нерелевантного. Прецеденты, которые будут использоваться для обучения модели, называется **обучающей выборкой**, а прецеденты, по которым будет оцениваться точность выборки, **тестовой.**

Релевантный текст – текст, который может быть интересен для аналитика с точки зрения географический диверсификации.

Абзац (параграф) – текст, выделенный переносами строк.

Бинарная классификация – задача классификации объектов, в которой объекты могут принадлежать только двум классам. Наша задача является задачей бинарной классификации с двумя классами – 0 (релевантен) и 1 (нерелевантен).

Переобучение – нежелательное явление, возникающее при решении задач обучения по прецедентам, когда вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Машинное обучение – область научного знания, основным преимуществом которого является возможность восстановить неизвестную зависимость по некоторой выборке, что позволяет решать задачи с трудно формализуемыми правилами. Например, такими задачами являются распознавание изображений, речи или некоторых объектов.

Одной из важнейших задач машинного обучения является задача классификации объектов, которая и рассматривается в данной работе. В нашем случае классификации является бинарной, то есть объекты делятся на два класса: 0 (у объекта отсутствует некоторое свойство) и 1 (у объекта присутствует некоторое свойство).

Примеров задач бинарной классификации достаточно много. Например, задача «Титаник» [1] - по признаковому описанию людей (номер класса, возраст, титул, порт посадки и др.) определить выживет человек или нет, т.е. провести классификацию объектов на два класса: выжил и не выжил. Если обратиться в мир бизнеса, то популярной является задача об оценке кредитоспособности [18]. Аналогично задаче «Титаник», по признаковому описанию людей нужно определить, стоит человеку выдавать кредит или нет.

Анализ текста далеко продвинулся вперед благодаря технологиям преобразования текста в численные признаки, которые можно использовать в качестве входных данных для модели. Основными такими технологиями являются: TF-IDF[1], word2vec[2]и doc2vec[3]. Word2vec особенно хорошо способен понимать смысл слова, например: если из векторного представления слова «король» вычесть вектор слова «мужчина» и прибавить вектор слова «женщина», то результатом будет вектор слова «королева».

[1] TF – term frequency (от англ. частота слова), IDF – inverse document frequence (от англ. обратная частота слова в документе)
[2] Word To Vector (слово в вектор) – векторное представление слов
[3] Document ToVector (документ в вектор) – векторное представление документа

ЦЕЛЬ РАБОТЫ

Аналитику требуется из годовых отчетов компании определить список географических мест, в которых компания имеет филиалы. Однако, в отчетах имеется избыточная для аналитика информация, на ознакомление с которой требуется время. Именно от этой информации требуется избавиться, т.е. нужно выдавать аналитику лишь тот текст, который может быть для него полезен. Такой текст назовем *релевантным*, а текст, который не дает информации аналитику о географической диверсификации, назовем *нерелевантным*.

Сформулируем критерии релевантности текста следующими правилами:

1) Явное указание. Очевидно, текст является релевантным, если в нем указано местоположение компании.

Пример: «Our worldwide headquarters is located on a 35-acre office complex in Atlanta, Georgia» [2].

Перевод: Наша штаб-квартира находится в офисном комплексе площадью 35 акров в Атланте, Джорджия.

Пояснение: *Явное указание* на то, что компания имеет представительство в Атланте, Джорджия.

2) Контактная информация. В годовых отчетах упоминают сотрудников компании и их местоположение.

Пример: Dr. Paul Achleitner, Chairman, Frankfurt am Main, March 2014

Перевод: Доктор Пол Акляйтнер, Председатель, Франкфурт на Майне, Март 2014.

Пояснение: Пол Акляйтнер является председателем во Франкфурте на Майне, следовательно, *можно сделать вывод*, что компания имеет филиалы во Франкфурте на Майне.

3) Логический вывод из текста. Текст является релевантным, если из него *возможно сделать* логический вывод о том, что компания находится в какой-либо географической точке. Под географической точкой понимается страна, область, субъекты, города и улицы.

Пример: «The Bank of England’s mission is to promote the Good of the People of the United Kingdom by maintaining Monetary and Financial Stability»[4].

Перевод: Миссией Банка Англии является содействие на благо граждан Объединенного Королевства путем обеспечения денежной и финансовой стабильности.

Пояснение: Из этого текста *можно сделать вывод*, что объект (Банк Англии) функционирует в Объединенном Королевстве (географическая точка), следовательно, имеет филиалы там же. Пример может показаться тривиальным, так как название субъекта уже говорит о принадлежности к Англии, но мы не можем привязываться к названию объекта, поэтому пример является показательным.

4) Логический вывод из числовых показателей. В тексте приведены числовые показатели с упоминанием стран, из которых *следует вывод* о диверсификации.

Пример: 9 %, Europe (excl. Germany, UK), Middle East and Africa [3].

Перевод: 9% Европа (исключая Германию, Великобританию), Центральная Азия и Африка.

Пояснение: Числовые показатели не всегда являются релевантными, поскольку нужно знать контекст приведенной информации. В данном случае речь идет о распределении капитала, поэтому эти цифры могут иметь значение.

Как видно, правила 1 и 2 можно описать с помощью формальных алгоритмов, используя специфические словари, однако правила 3 и 4 трудно формализовать ввиду того, что они требуют некого логического вывода. Так как правила 3 и 4 трудно формализуемы, мы используем машинное обучение, которое позволит выявить закономерности из наших данных.

В итоге нам нужно построить математическую модель, на вход которой подается текст, для которого нужно определить его релевантность. На выходе получаем либо 0 (нерелевантен), либо 1 (релевантен). Также опционально на выходе есть возможность получить вероятность релевантности и поставить

порог, чтобы решить: текст релевантен или нет. По умолчанию, если вероятность больше 0.5, то текст является релевантным и модель выдает 1.

Для оценки точности модели мы будем использовать метрику *accuracy* (точность), которая считается по следующей формуле:

$$s = \frac{N_{correct}}{N}, \quad (1)$$

где $N_{correct}$– количество верных предсказаний модели на тестовой выборке, N – длина тестовой выборки .

Таким образом, мы можем сформулировать цель работы.

Цель работы: создание математической модели оценки релевантности текста с максимальной точностью, на вход которой подается текст, а на выходе получается бинарный ответ, где 0 означает, что текст нерелевантен, а 1 – текст релевантен.

ОБЗОР ЛИТЕРАТУРЫ

Анализ финансовых отчетов и анализ текста являются обширными задачами. Приведем некоторые примеры комбинации этих задач.

Balakrishnan R.,Qiu X.Y., Srinivasan [5] решают задачу поиска в отчетах компании повествовательной информации, которая может быть важна для инвестиционных решений. Для обработки текста использовались технологии TF-IDF и bagofwords (обычная частота слов). В итоге авторам удалось при помощи найденной информации построить доходный портфель. Сделан вывод о том, что повествовательная информация может использоваться инвесторами как фактор, предсказывающий избыточную доходность.

Tsai M. и Wang C., Srinivasan [6] называют текстовую информацию «soft information». Для извлечения признаков из текста они также используют TF-IDF и bagofwords. Для выявления связи между риском и текстом была использована регрессия и ранжирование слов.

В 2013 году выходит теоретическая работа Mikolov T., Sutskever I. и др. [7] «Distributed Representations of Words and Phrases and their Compositionality», в которой описывается принцип работы word2vec. В следующей работе в 2014

году Le .Q и Mikolov T. [8] уже описывается принцип работы doc2vec. Эксперименты проводятся на различных наборах в данных, в частности, на выборке от сайта IMDB [9], в которой нужно предсказать по тексту отзыва, является он положительным или отрицательным.

В 2002 году Sebastiani F. [10] описывает различные методы классификации текста и примеры их использования. Например, для фильтрации текста, для определения синонимов и иерархический категоризации веб-страниц.

1. ЛИНЕЙНЫЕ И НЕЛИНЕЙНЫЕ КЛАССИФИКАТОРЫ.

1.1. Основные понятия. Постановка задачи классификации.

Формальная постановка задачи классификации выглядит следующим образом:

Пусть X–множество описаний объектов, Y–конечное множество номеров классов. Существует неизвестная целевая зависимость $y^*: X \to Y$, значения которой известные только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$. Требуется построить алгоритм $a^*: X \to Y$, способный классифицировать произвольный объект $x \in X$.

Признаком называется отображение $f: X \to D_f$, где D_f – множество допустимых значений признака. Если заданы признаки $f_1, \dots, f_n$, то вектор $x=(f_1(x), \dots, f_n(x))$называется признаковым описанием объекта$x \in X$.

Признаковые описания допустимо отождествлять с самими объектами. При этом множество $X=D_{f1} \times \dots \times D_{fn}$ называют *признаковым пространством*.
В зависимости от типа множества D_f признаки делятся на следующие типы:

- бинарный признак: $D_f = \{0,1\}$;
- номинальный признак: D_f – конечное множество;
- порядковый признак: D_f – конечное упорядоченное множество;
- количественный признак: D_f - множество действительных чисел.

1.2. Линейный классификатор. Общее определение. Методы обучения.

1.2.1 Основные определения

Линейный классификатор – алгоритм классификации, основанный на построении линейной разделяющей поверхности. В случае двух классов разделяющей поверхностью является гиперплоскость, которая делит пространство признаков на два полупространства. В случае большего числа классов разделяющая поверхность кусочно-линейная.

В свою очередь, алгоритм классификации, не удовлетворяющий определению линейного классификатора, называется *нелинейным классификатором*.

Пусть объекты описываются n числовыми признаками

$$f_j: X \to R,\ j=1,2,\ldots,n.$$

Тогда пространство признаковых описаний объектов есть $X = R^n$. Пусть Y – конечное множество номеров классов. Положим $Y = \{-1,+1\}$. Тогда *линейным классификатором* называется алгоритм классификации $a: X \to Y$ вида [10]:

$$a(x,w) = sign\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right) = sign < x,w >. \quad (2)$$

1.2.2 Обучение линейного классификатора

1.2.2.1 Метод минимизации эмпирического риска

Суть метода в том, чтобы по заданной обучающей выборке пар (объект, класс) $X^m = \{(x_1,y_1),\ldots,(x_m,y_m)\}$ построить алгоритм $a: X \to Y$, который минимизирует функционал эмпирического риска [10]:

$$Q(w) = \sum_{i=1}^{m} [a(x_i,w) \neq y_i] \to min_w. \quad (3)$$

Таким образом, минимизируется количество неверных предсказаний алгоритма на обучающей выборке.

1.2.2.2 Запись функционала с помощью понятия отступа

Для обучающего объекта $x_i \in X^m$ введем *величину отступа* [10]:

$$M(x_i) = y_i < x_i,w >. \quad (4)$$

Значение отступа характеризует расстояние от объекта до границы классов. Чем меньше это значение, тем выше вероятность ошибки. Так как значение отступа $M(x_i)$ для объекта x_i отрицательно тогда и только тогда, когда алгоритм $a(x)$ допускает ошибку на объекте x_i. Тогда запишем значение функционала в следующем виде [10]:

$$Q(w) = \sum_{i=1}^{m} [M(x_i) < 0]. \quad (5)$$

1.2.2.3 Замена пороговой функции потерь

Поскольку задачу оптимизации гораздо удобнее решать, если функционал является непрерывной или гладкой функцией, наиболее известные методы обучения используют замену функции потерь ее различными аппроксимациями, т.е.

$$[M < 0] \leq L(M), \quad (6)$$

где L – аппроксимирующая функция. После замены функции потерь минимизируется не сам функционал эмпирического риска, а его верхняя оценка [10]:

$$Q(w) \leq \tilde{Q}(w) = \sum_{i=1}^{m} L\big(M(x_i)\big). \quad (7)$$

1.2.2.4 Регуляризация

Для борьбы с переобучением рекомендуется к функционалу (7) добавлять штрафное слагаемое, которое не будет допускать большие значения нормы вектора весов [11]:

$$\mathrm{Q(w)} \leq \widetilde{\mathrm{Q}}\mathrm{(w)} = \sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{L(M(x_i))} + \gamma \|\mathrm{w}\|^{p} \rightarrow \min_{\mathrm{w}} \quad (8)$$

Параметр регуляризации γ подбирается из априорных соображений, либо по скользящему контролю. Параметр p = 2 для применения градиентных методов минимизации и p = 1 для отсева неинформативных признаков.

1.2.3 Методы обучения линейных классификаторов

Методы различаются, в основном выбором функции *L(M)*, способом регуляризации, также численным методом решения задачи оптимизации.

1.2.3.1 Линейный дискриминант Фишера

По сути, это квадратичная аппроксимация при условии, что из каждого вектора признака вычитается вектор центра класса:

$$[M < 0] \leq (1 - M)^2. \quad (9)$$

Задача обучения решается методом наименьших квадратов.

1.2.3.2 Однослойный перцептрон

Аппроксимация выглядит следующим образом [10]:

$$[M < 0] \leq \frac{2}{1 + e^{\propto M}}, \quad (10)$$

где параметр α задается из априорных соображений. Задача оптимизации решается с помощью градиентных методов [10].

1.2.3.3 Метод опорных векторов

Метод опорных векторов соответствует кусочно-линейной аппроксимации [10]:

$$[M < 0] \leq (1 - M)_+, \quad (11)$$

где + обозначает положительные значения. Применяется регуляризация с квадратичной нормой. Задача оптимизации решается как задача квадратичного программирования.

1.3. Байесовский классификатор. Наивный байесовский классификатор. Логистическая регрессия.

Байесовский классификатор – класс алгоритмов классификации, основанный на принципе максимума апостериорной вероятности [11].

1.3.1 Постановка задачи

Пусть X–множество описаний объектов, Y–множество номеров классов. На множестве пар (объект, класс) множества $X \times Y$ определена вероятностная

мера P. Имеется конечная обучающая выборка независимых наблюдений $X^m=\{(x_1,y_1),\ldots,(x_m,y_m)\}$, полученных согласно вероятностной мере P.

Задача: построить алгоритм $a: X \to Y$, способный классифицировать объект $x \in X$.

Задача разделяется на две:

- Построение оптимального классификатора, если известны плотности классов;
- Оценка плотностей классов по обучающей выборке.

1.3.2 Построение классификатора при известных плотностях классов

Пусть для каждого класса $y \in Y$ известна априорная вероятность P_y того, что появится объект класса y и плотности распределения $p_y(x)$ каждого из классов, называемые *функциями правдоподобия классов*. Требуется построить алгоритм a(x), доставляющий минимальное значение *функционалу среднего риска*.

Средний риск – это математическое ожидание ошибки:

$$R(a) = \sum_{y \in Y} \sum_{s \in Y} \lambda_y P_y \, P_{(x,y)}\{a(x) = s | y\}, \tag{12}$$

где λ_y – цена ошибки или штраф за отнесение объекта класса y к какому-либо другому классу. Решение этой задачи дает следующая теорема [11]:

Теорема. Решением задачи является алгоритм

$$a(x) = argmax_{y \epsilon Y} \lambda_y P_y p_y(x). \tag{13}$$

1.3.3 Восстановление плотностей классов по обучающей выборке

Нужно по заданной подвыборке объектов класса y построить эмпирические оценки априорных вероятностей P_y и функций правдоподобия $p_y(x)$. В качестве оценки вероятностей берут относительную частоту каждого класса в обучающей выборке.

Существует три подхода восстановления функции правдоподобия:

- Параметрическое восстановление плотности при предположении, что плотности нормальные (линейный дискриминант Фишера [11]);
- Непараметрическое восстановление (метод парзеновского окна [11]);

- Разделение смеси распределений (expectation-maximization algorithm) [12].

1.3.3.1 Линейный дискриминант Фишера

В основе метода два предположения:

- Классы распределены по нормальному закону;
- Матрицы ковариаций классов равны.

Процесс классификации выглядит следующим образом:

1. Обучение
    a. Оценивание математических ожиданий μ_y;
    b. Вычисление общей ковариационной матрицы Σи ее обращение.
2. Классификация

Математическое ожидание и матрица ковариаций класса yоценивается по следующим формулам:

$$\mu_y = \frac{1}{l_y} \sum_{\substack{i=1 \\ y_i=y}}^{l} x_i \, \Sigma_y = \frac{1}{l_y} \sum_{\substack{i=1 \\ y_i=y}}^{l} (x_i - \mu_y)(x_i - \mu_y)^T, \tag{14}$$

$$\Sigma_y = \frac{1}{l_y} \sum_{\substack{i=1 \\ y_i=y}}^{l} (x_i - \mu_y)(x_i - \mu_y)^T, \tag{15}$$

где l – длина выборки, l_y–количество объектов класса y в обучающей выборке x.

Классификация происходит по следующей формуле:

$$a(x) = argmax_{y \epsilon Y} \left(\ln(\lambda_y P_y) - \frac{1}{2} \mu_y^T \Sigma^{-1} \mu_y + x^T \Sigma^{-1} \mu_y \right), \tag{16}$$

где λ_y – цена ошибки на объекте класса y.

1.3.3.2 Метод парзеновского окна

В основе подхода идея о том, что плотность выше в тех точках, рядом с которыми находится большое количество объектов выборки.

Парзеновская оценка плотности и классификатор имеют вид:

$$p_{y,h}(x) = \frac{1}{l_y V(h)} \sum_{i=1}^{l} [y_i = y] \, K\left(\frac{\rho(x, x_i)}{h} \right), \tag{17}$$

$$a(x; X^l, h) = argmax_{y \in Y} \lambda_y \sum_{i=1}^{l} [y_i = y] K\left(\frac{\rho(x, x_i)}{h}\right), \quad (18)$$

где h – ширина окна, $V(h)$ – некоторая функция от ширины окна, $\rho(x,x_i)$– расстояние от x до x_i в некотором метрическом пространстве, λ_y – цена правильного ответа на объекте класса y, $K(z)$–произвольная четная функция, называемая функцией ядра или окна.

1.3.3.3 Разделение смеси распределений. EM-алгоритм

В основе предположение о том, что искомая плотность распределения имеет вид смеси k нормальных распределений:

$$p(x) = \sum_{j=1}^{k} w_j p_j(x), \sum_{i=1}^{k} w_j = 1, w_j \geq 0, \quad (19)$$

где $p_j(x)$ – функция правдоподобия j-й компоненты смеси, w_j –ее априорная вероятность. Тогда плотность j-й компоненты смеси равна:

$$p_j(x) = N(x; \mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right) \quad (20)$$

Оценки параметров вычисляются по следующим формулам:

$$w_j = \frac{1}{m} \sum_{i=1}^{m} g_{ij}, \quad (21)$$

$$\mu_j = \frac{1}{m w_j} \sum_{i=1}^{m} g_{ij} x_i, \quad (22)$$

$$\sigma_j^2 = \frac{1}{m w_j} \sum_{i=1}^{m} g_{ij} (x_i - \mu_j)^2, j = 1, .., k, \quad (23)$$

где

$$g_{ij} = \frac{w_j N(x_i; \mu_j, \sigma_j)}{\sum_{s=1}^{k} w_s N(x_i; \mu_s, \sigma_s)}. \quad (24)$$

1.3.4 Наивный байесовский классификатор

Наивный байесовский классификатор – частный случай байесовского классификатора, основанный на предположении, что объекты $x \epsilon X$ описываются n статистически независимыми признаками:

$$x = (\xi_1, \dots, \xi_n) \equiv (f_1(x), \dots, f_n(x)). \quad (25)$$

Из предположения следует, что функции правдоподобия классов представимы в виде:

$$p_y(x) = p_{y1}(\xi_1) \cdot \dots \cdot p_{yn}(\xi_n), \quad (26)$$

где $p_{yj}(\xi_j)$ – плотность распределения j-го признака для класса y.

Таким образом, задача упрощается, поскольку оценить n одномерных плотностей гораздо легче, чем одну n-мерную плотность. Оценить эти плотности можно методами описанными выше.

1.3.5 Логистическая регрессия

Для обучения классификатора решается следующая оптимизационная задача:

$$Q(w) = \sum_{i=1}^{m} \ln(1 + \exp(-y_i < x_i, w >)) \to min_w. \quad (27)$$

После решения задачи оптимизации, классификация производится следующим образом:

$$a(x) = sign < x, w >. \quad (28)$$

Кроме того, можно оценить апостериорные вероятности принадлежности классам:

$$P\{y|x\} = \sigma(y < x, w >), \quad (29)$$

где

$$\sigma(z) = \frac{1}{1+e^{-z}}. \quad (30)$$

1.4. Бустинг моделей. Бустинг над решающими деревьями

1.4.1 Бустинг. Общее определение

Бустинг – процедура последовательного построения *композиции* алгоритмов машинного обучения, когда каждый следующий алгоритм

стремится компенсировать недостатки композиции всех предыдущих алгоритмов.

Будем искать финальный алгоритм классификации в виде композиции:

$$F_M(x) = \sum_{m=1}^{M} b_m h(x; a_m), b_m \in R, a_m \in R, \quad (31)$$

где M – количество алгоритмов, b_m – веса ответов моделей,a_m–вектор параметров для m-той модели, $h(x;a_m)$–ответ m-того алгоритма для объекта x.

Поскольку подбор полного набора параметров $\{a_m,b_m\}m=1,..,M$– трудоемкая задача, то будем считать, что нами уже построен классификатор F_{m-1} длины $m-1$. Тогда задача сводится к поиску пары оптимальных параметров $\{a_m,b_m\}$для классификатора длины m:

$$F_m(x) = F_{m-1}(x) + b_m h(x; a_m) \quad (32)$$

Вводим некоторую функцию потерь и минимизируем функционал ошибки:

$$Q = \sum_{i=1}^{N} L\big(y_i, F_m(x_i)\big) \to min. \quad (33)$$

Решить эту задачу можно с помощью градиентного спуска [13]. В итоге параметры подбираются линейным поиском:

$$a_m(x) = argmin \sum_{i=1}^{N} L(\nabla Q, h(x_i, a)), \quad (34)$$

$$b_m(x) = argmin \sum_{i=1}^{N} L\big(F_{m-1}(x_i) - bh(x_i, a_m)\big). \quad (35)$$

1.4.2 Бустинг над решающими деревьями

Рассмотрим в качестве базового семейства алгоритмов регрессионные решающие деревья из J вершин.

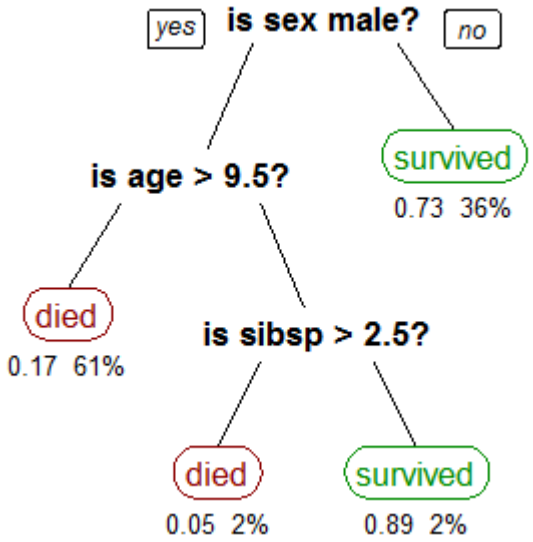


Рисунок. 1. Решающее дерево с 3 вершинами

Каждое решающее дерево имеет *J* листовых вершин, соответствующие *J* непересекающимся областям $\{R_j\}$ $j=1..J$, на которые разбивается пространство объектов *X*. Каждой листовой вершине соответствует некоторое значение регрессии b_j, которое будет ответом классификатора в случае попадания анализируемого объекта в соответствующую область. Запишем этот факт следующей формулой:

$$h\left(x, \{a_j, R_j\}_{j=1}^{J}\right) = \sum_{j=1}^{J} a_j I[x \in R_j], \tag{36}$$

где *I[A]* – индикатор события *A*. С учетом (32) запишем:

$$F_m(x) = F_{m-1}(x) + b_m \sum_{j=1}^{J} a_{jm} I[x \in R_j] = F_{m-1}(x) + \sum_{j=1}^{J} c_{jm} I[x \in R_j], c_{jm} = a_{jm} b_m. \tag{37}$$

Тогда оптимальные параметры можно найти по следующей формуле:

$$c_{jm} = argmin_c \sum_{x \in R_{jm}} L(y_i, F_{m-1}(x_i) + c). \tag{38}$$

2. ПОДХОДЫ К ПРЕОБРАЗОВАНИЮ ТЕКСТОВОЙ ИНФОРМАЦИИ В ЧИСЛЕННЫЕ ПРИЗНАКИ

2.1. Общая задача преобразования текста

Имеется корпус документов D, состоящий из документов d_i, $i=1..m$, где m– количество документов в корпусе D. В каждом документе d_iимеются слова t_j, $j=1..n_i$, где n_i–количество слов в документе d_i.

Задача: получить из корпуса документов D признаки $x_i \in X$.

2.2.TF-IDF. Мера важности слова [14]

TF (term frequency – частота слова) –отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова tiв пределах отдельного документа:

$$tf(t,d) = \frac{n_t}{\sum_k n_k}, \quad (39)$$

где n_t – число вхождений слова tв документ, а в знаменателе – общее число слов в документе d.

IDF (inverse document frequency – обратная частота документа) – инверсия частоты, с которой некоторое слово встречается в корпусе документов D. Учет IDFуменьшает вес широкоупотребительных слов. Оценка IDFсчитается следующим образом:

$$idf(t,D) = log\frac{|D|}{|\{d_i \in D | t \in d_i\}|}, \quad (40)$$

где $|D|$ - число документов в корпусе, в знаменателе – число документов из корпуса D, в которых встречается слово t.

Итоговая оценка TF-IDFвычисляется следующим образом:

$$tf\_idf(t,d,D) = tf(t,d) \times idf(t,D). \quad (41)$$

2.3.Word2Vec. Векторное представление слова [7]

2.3.1 Векторное представление слова

Основной идеей является представление слова в виде вектора. Например, в подходе «One-hot» векторное представление будет выглядеть:

Таблица 1. One-hot матрица

Слово	King	…	Queen	Book
King	1		0	0
…				
Queen	0		1	0
Book	0		0	1

Тогда векторное представление, например, слова *King = [1, 0, …, 0]*. Размерность этого вектора равна количеству слов в документе. Поскольку порядок слов в документе довольно высок, размерность вектора слов тоже становится высокой. Если в этом пространстве снизить размерность до трех, то можно получить компактное представление, например, *King = [0.9457, 0.5774, 0.2224]*. Задача понижения размерности решается с помощью SVD[4], которое имеет сложность $O(mn^2)$, что требует больших вычислений в случае, если *n*– велико.

2.3.2 Word2vec

Word2vec – программный инструмент анализа семантики естественных языков. В отличие от TF-IDF способен понимать семантический смысл слов. Основной идеей word2vec является отказ от подсчета количества слов, заменяя его предсказанием окружающих слов относительно каждого слова **[7].**

Тогда решим следующую задачу:

[4]SVD(англ. singular value decomposition) – сингулярное разложение матрицы

Предсказать окружающие слова в окне длиной m каждого слова. Под окном длины m понимается количество слов вокруг слова, которые нужно предсказать. Максимизируем следующий функционал:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m\leq j\leq m,\\ j\neq 0}} \log(p(w_{t+j}|w_t), \tag{42}$$

где θ – все переменные, которые мы оптимизируем, T–количество слов в документе, w_t–каждое слово из документа. Оптимизируемые вектора можно увидеть, если расписать следующее выражение:

$$p(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{W}\exp(u_w^T v_c)}, \tag{43}$$

где o – предсказываемое окружающее слово, c–слово, относительно которого предсказываются слова (т.н. центральное слово), u–вектор центрального слова, v – вектор окружающего слова. Тогда получаем для каждого слова два векторных представления: одно, в случае если слово центральное, второе, в случае если слово не является центральным.

2.4.Doc2Vec. Векторное представление документа.

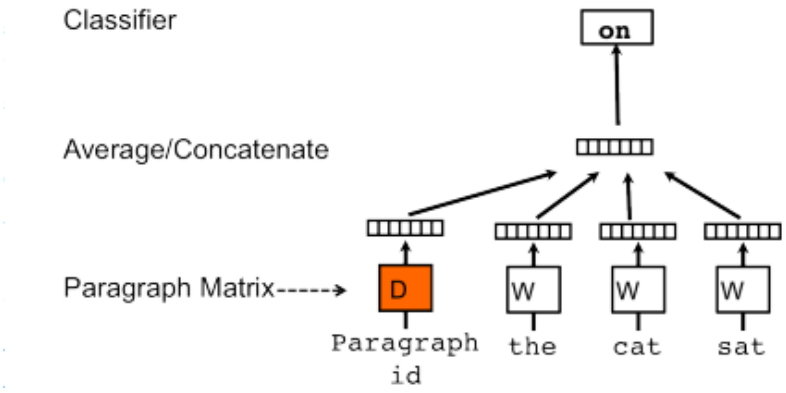


Рисунок 2. Принцип работы doc2vec

Подход аналогичен word2vec, но в качестве входных данных для модели также подается метка абзаца. В качестве метки может быть идентификатор абзаца или класс, к которому абзац относится.

3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

3.1. Сбор данных

Для сбора данных было использовано 30 годовых отчетов компании. (приложение А). Отчеты на английском языке, скачаны с официальных сайтов компаний. Текст из отчета получен с помощью функции pdfminer на Python, Далее все отчеты были разбиты на абзацы. В итоге получилось 7390 абзацев.

3.2. Подготовка данных к обучению

Перед тем, как обучать модель, данные нужно очистить. Были проделаны следующие действия:

- Удалены переносы строк;
- Удалены знаки пунктуации;
- Удалены числа;
- Слова были обрезаны до их основы;
- Удалены стоп-слова, т.е. слова, не имеющие семантического смысла, например, местоимение «I», частица «that».

3.3. Анализ данных

Посмотрим на количество объектов в каждом классе:

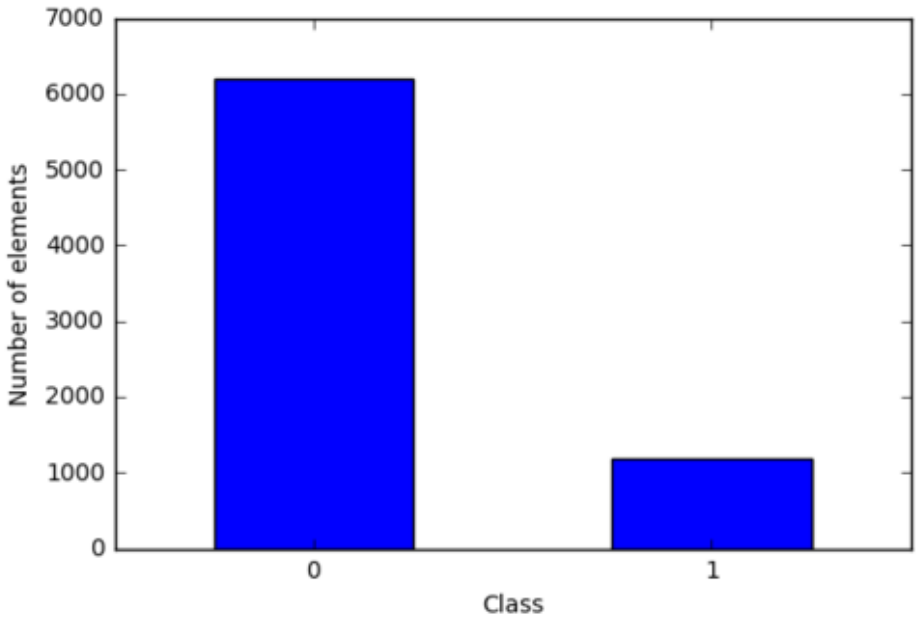


Рисунок 3. Распределение классов.

Приведем некоторые характеристики объектов в таблице 2.

Таблица 2. Характеристики абзацев из выборки

Характеристика	Значение
Среднее количество символов в абзаце	538
Минимальное количество символов в абзаце	4
Максимальное количество символов в абзаце	13261
Среднее количество слов в абзаце	78
Минимальное количество слов в абзаце	1
Максимальное количество слов в абзаце	972

3.4. Генерация количественных признаков

Сгенерируем следующие признаки:

- Количество слов в абзаце;
- Количество символов в абзаце;
- Количество символов в абзаце без пробелов.

3.5. Генерация признаков TF-IDF

Для генерации признаков TF-IDF использована функция TfidfVectorizer из библиотеки sklearn на языке Python [15]. Оценка TF вычислена по формуле (39). Оценка IDF вычислеана по формуле (40). Итоговая оценка TF-IDF определена по формуле (41). В итоге получена матрица размером 7390×25328, где 7390 – количество абзацев, 25328 – размер словаря.

3.6. Генерация признаков Word2Vec

Для обучения модели word2vec требуется большой корпус текстов, а в нашем распоряжении только 7390 абзацев, что недостаточно для обучения word2vec. Поэтому мы использовали векторное представление слов из модели word2vec от компании Google, обученной на текстах новостей [16]. Модель содержит векторные представления 3×10^6 слов и имеет размерность 300.

С помощью этой модели мы сгенерировали векторное представление абзаца, которое можно получить усреднением векторных представлений слов.

Пусть мы имеем абзац из n слов:

$$M = \{a_1, \dots, a_n\}, \quad (44)$$

где M – множество слов, из которых состоит абзац, причем слова могут повторяться; $a_1, \dots, a_n$ – слова.

Обозначим векторные представления слов $a_1, \dots, a_n$:

$$v_i = G(a_i), i = 1..n, \quad (45)$$

где G –обученная модель word2vec, которая по слову a_i возвращает векторное представление v_i. Тогда векторное представление V_{text} абзаца вычисляется следующим образом:

$$V = \sum_{i=1}^{n} v_i \ , V_{text} = \frac{V}{\sqrt{{V_1}^2 + \dots + {V_{300}}^2}}. \quad (46)$$

Это векторное представление использовано в качестве признаков. В случае, если у слова нет векторного представления, то оно не учитывается в векторном представлении абзаца.

3.7. Генерация признаков Doc2Vec

Для генерации признаков doc2vec использована функция Doc2Vec из библиотеки gensim на языке Python [17]. Сначала мы обучили модель doc2vec, подав на вход текст абзаца и идентификатор абзаца. В итоге для каждого абзаца получен вектор размерности 5. Код, который используется для генерации всех признаков, приведен в приложении Б**.**

3.8.Обучение моделей

Все сгенерированные признаки объединим в одну матрицу признаков. Таким образом, для каждого объекта будем иметь 25636 признаков:

3 количественных признака,

25328 признаков TF-IDF,

300 признаков word2vec и

5 признаков doc2vec.

В нашем примере выборка не сбалансирована, т.к. имеется 1 194 релевантных абзацев и 6 196 нерелевантных абзацев. Тогда, предположим, что

у нас есть наивный классификатор, который считает все абзацы нерелевантными, тогда его точность (1) на выборке будет равна:

$$s = \frac{N_{correct}}{N} = \frac{6196 - 1194}{6196} = 80{,}7\%.$$

Возьмем этот результат за базовое решение, т.е. классификаторы, имеющие точность меньше 80,7%, будем считать неудачными.

Разделим выборку на тренировочную и тестовую случайным образом по правилу: размер тренировочной выборки - 80%, тестовой – 20%.

В качестве алгоритмов построения моделей выберем:

- Наивный байесовский классификатор, использующий предположение (25) и строящий функции правдоподобия классов (26). Восстановление функции правдоподобия происходит с помощью линейного дискриминанта Фишера (14), (15). Тогда классификация происходит по формуле (16).
- Логистическая регрессия. Решается оптимизационная задача (27). Классификация производится по формуле (28).
- Градиентный бустинг над решающими деревьями. Решается оптимизационная задача (33). В итоге параметры модели будут найдены с помощью (34), (35).

Построим три модели, которые мы обучим на тренировочной выборке. С помощью них построим предсказания на тестовую выборку и посчитаем точность. Точность вычисляется по формуле (1).

Таблица 3. Результаты трех моделей, обученных на несбалансированной выборке

Классификатор	**Точность, %**
Наивный байесовский классификатор	38,22
Логистическая регрессия	87,68
Градиентный бустинг над решающими деревьями	88,70

Наивный байесовский классификатор хуже базового решения на 42,48%, поэтому можно считать его неудачным. Можно увидеть, что логистическая

регрессия дает точность 87,68%, градиентный бустинг дает прирост точности относительно логистической регрессии на 1,02%. В целом, получившиеся классификаторы точнее базового решения на 6,98% и 8% соответственно.

Теперь сбалансируем выборку в равных пропорциях, чтобы оценить обобщающую способность алгоритма более точно. Для этого отбросим случайным образом 6196 - 1194 = 5002 нерелевантных абзацев и обучим классификаторы заново. Кроме того, базовое решение теперь будет иметь точность 50%.

Таблица 4. Результаты трех моделей, обученных на сбалансированной выборке

Классификатор	**Точность, %**
Наивный байесовский классификатор	67,99
Логистическая регрессия	75,52
Градиентный бустинг над решающими деревьями	75,94

Все три классификатора лучше базового решения на 17,99%, 25,52% и 25,94%, что говорит о том, что классификаторы совершают обобщение.

Поскольку балансировка случайным образом отбрасывает некоторые объекты классов, проведем ее несколько раз. Если сделать балансировку и снова обучить классификатор 20 раз, затем взять усредненные значения точности, то ситуация изменится.

Таблица 5. Усредненные результаты моделей,
обученных на 20 разных сбалансированных выборках

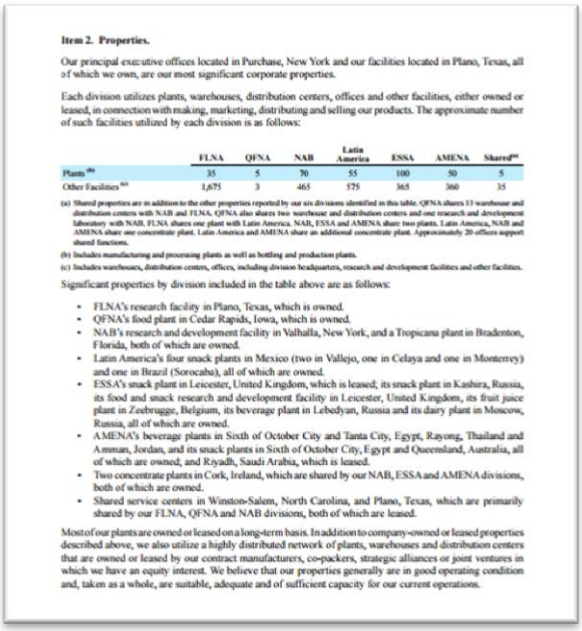
Классификатор	**Точность, %**
Наивный байесовский классификатор	69,51
Логистическая регрессия	76,51
Градиентный бустинг над решающими деревьями	74,50

В первых двух случаях градиентный бустинг точнее линейных классификаторов. Однако, в среднем, логистическая регрессия точнее градиентного бустинга, поэтому можно сделать вывод, что нелинейный

классификатор не всегда точнее линейного. Листинг программы приведем в приложении В.

3.9 Иллюстративный пример работы модели

Для примера возьмем годовой отчет компании PepsiCo за 2016 год (номер 12 в приложении А). Разобьем его на абзацы и оставим только те абзацы, которые имеют географическую информацию.



Item 2. Properties.

Our principal executive offices located in Purchase, New York and our facilities located in Plano, Texas, all of which we own, are our most significant corporate properties.

Each division utilizes plants, warehouses, distribution centers, offices and other facilities, either owned or leased, in connection with making, marketing, distributing and selling our products. The approximate number of such facilities utilized by each division is as follows:

	FLNA	QFNA	NAB	Latin America	ESSA	AMENA	Shared[a]
Plants[b]	35	5	70	55	100	50	5
Other Facilities[c]	1,675	3	465	575	365	360	35

(a) Shared properties are in addition to the other properties reported by our six divisions identified in this table. QFNA shares 13 warehouse and distribution centers with NAB and FLNA. QFNA also shares two warehouse and distribution centers and one research and development laboratory with NAB. FLNA shares one plant with Latin America. NAB, ESSA and AMENA share two plants. Latin America, NAB and AMENA share one concentrate plant. Latin America and AMENA share an additional concentrate plant. Approximately 20 offices support shared functions.

(b) Includes manufacturing and processing plants as well as bottling and production plants.

(c) Includes warehouses, distribution centers, offices, including division headquarters, research and development facilities and other facilities.

Significant properties by division included in the table above are as follows:

- FLNA's research facility in Plano, Texas, which is owned.
- QFNA's food plant in Cedar Rapids, Iowa, which is owned.
- NAB's research and development facility in Valhalla, New York, and a Tropicana plant in Bradenton, Florida, both of which are owned.
- Latin America's four snack plants in Mexico (two in Vallejo, one in Celaya and one in Monterrey) and one in Brazil (Sorocaba), all of which are owned.
- ESSA's snack plant in Leicester, United Kingdom, which is leased; its snack plant in Kashira, Russia, its food and snack research and development facility in Leicester, United Kingdom, its fruit juice plant in Zeebrugge, Belgium, its beverage plant in Lebedyan, Russia and its dairy plant in Moscow, Russia, all of which are owned.
- AMENA's beverage plants in Sixth of October City and Tanta City, Egypt, Rayong, Thailand and Amman, Jordan, and its snack plants in Sixth of October City, Egypt and Queensland, Australia, all of which are owned, and Riyadh, Saudi Arabia, which is leased.
- Two concentrate plants in Cork, Ireland, which are shared by our NAB, ESSA and AMENA divisions, both of which are owned.
- Shared service centers in Winston-Salem, North Carolina, and Plano, Texas, which are primarily shared by our FLNA, QFNA and NAB divisions, both of which are leased.

Most of our plants are owned or leased on a long-term basis. In addition to company-owned or leased properties described above, we also utilize a highly distributed network of plants, warehouses and distribution centers that are owned or leased by our contract manufacturers, co-packers, strategic alliances or joint ventures in which we have an equity interest. We believe that our properties generally are in good operating condition and, taken as a whole, are suitable, adequate and of sufficient capacity for our current operations.

Рисунок 4. Пример страницы из отчета PepsiCo

В итоге получили 130 абзацев, для которых построим вышеуказанные признаки. Подаем на вход модели вычисленные признаки, получаем 17 релевантных абзацев из 130. Приведем несколько примеров этих абзацев:

Таблица 6. Примеры релевантных абзацев.

Абзац	**Географические места**
Creating well-paying jobs, along with the promise of long, successful careers, for hardworking men and women from a variety of backgrounds, including not just MBAs and scientists, but also truck drivers, farmers and factory workers with all kinds ofskills. But were also doing our part to be a good neighbor in a number of other ways, from PepsiCo Gives Back, a day of service when hundreds of associates volunteered in the New York City area, to our efforts supporting the people of Flint, Michigan in the wake of the water crisis. Working with civic leaders and other partners, we donated roughly 6.5million bottles of water and opened help centers where Flint residents could get everything from lead mitigating foods, to mental and physical health services, to personal care items and otheressentials. That same spirit, that same commitment to being good neighbors, was felt around the world over the past yearfrom Mexico, where Quaker launched a PeanOat malnutrition prevention trial with more than 1,000low-income children, to the UK, where we are supporting Magic Breakfast, a charity offering more than 30,000children a healthy start to their day, to	Нью-Йорк; г. Флинт, штат Мичиган; г. Мехико; Великобритания.

Pakistan, where we tripled the number of girls benefiting from our I am PepsiCo mentoring and scholarshipprogram. Of course, we cannot and should not try to solve all of the worlds problems. But we also know we can make a difference. And by doing so, we are not only fostering the kind of widespread public support upon which the success of our companylike any consumer goods companydepends, we are also lending a hand to people who need it, meeting our responsibilities as fellow citizens of the countries we callhome.	
See Off-Balance-Sheet Arrangements in Our Financial Results Our Liquidity and Capital Resources in Item 7. Managements Discussion and Analysis of Financial Condition and Results of Operations for more information on our independent bottlers. Our Competition Our beverage, food and snack products are in highly competitive categories and markets and compete against products of international beverage, food and snack companies that, like us, operate in multiple geographies, as well as regional, local and private label manufacturers, economy brands and other competitors. In many countries in which our products are sold, including the United States, The Coca-Cola Company is our primary beverage competitor. Other beverage, food and snack competitors include, but are not limited to, DPSG, International, Inc., Monster Beverage Corporation, Kellogg Company, The Kraft Heinz Company, Nestl S.A., Red Bull GmbH and Snyders-Lance, Inc. Many of our food and snack products hold significant leadership positions in the food and snack industry in the United States and worldwide. In 2016, we and The Coca-Cola Company represented approximately 24% and 20%, respectively, of the U.S. liquid refreshment beverage category by estimated retail sales in measured channels, according to Information Resources, Inc. However, The Coca-Cola Company has significant carbonated soft drink (CSD) share advantage in many markets outside the United States. Our beverage, food and snack products compete primarily on the basis of brand recognition and loyalty, taste, price, value, quality, product variety, innovation, distribution, advertising, marketing and promotional activity, packaging, convenience, service and the ability to anticipate and effectively respond to consumer preferences and trends, including increased consumer focus on health and wellness. Success in this competitive environment is dependent on effective promotion of existing products, effective introduction of new products and the effectiveness of our advertising campaigns, marketing programs, product packaging, pricing, increased efficiency in production techniques, new vending and dispensing equipment and brand and trademark development and protection. We believe that the strength of our brands, innovation and marketing, coupled with the quality of our products and flexibility of our distribution network, allows us to compete effectively. Research and Development We engage in a variety of research and development activities and invest in innovation globally with the goal of meeting changing consumer demands and preferences and accelerating sustainable growth. These activities principally involve: development of new ingredients, flavors and products; reformulation and improvement in the quality and appeal of existing products; improvement and modernization of manufacturing processes, including cost reduction; improvements in product quality, safety and integrity; development of, and improvements in, dispensing equipment, packaging technology, package design and portion sizes; efforts focused on identifying opportunities to transform, grow and broaden our product portfolio, including by developing products with improved nutrition profiles that reduce sodium, saturated fat or added sugars, including through the use of sweetener alternatives and flavor modifiers and innovation in existing sweeteners, and by offering more products with positive nutrition including whole grains, fruits and vegetables, dairy, protein and hydration; and improvements in energy efficiency and efforts focused on reducing our impact on the environment. Our research centers are located around the world, including in Brazil, China, India, Mexico, Russia, the United Arab Emirates, the United Kingdom and the United States, and leverage nutrition science, food science, engineering and consumer insights to meet our strategy to continue to develop nutritious and convenient beverages, foods and snacks. In 2016, we continued to refine our beverage, food and snack portfolio to meet changing consumer demands by reducing added sugars in many of our beverages and saturated fat and sodium in many of our foods and snacks, and by developing a broader portfolio of product choices, including: continuing to expand our 7	США; Бразилия; Китай; Индия; Мексика; Россия; ОАЭ; Великобритания.

Отметив географические точки из таблицы 6 на карте, получим следующее:

Рисунок 5. Географические точки из некоторых релевантных абзацев годового отчета компании PepciCo 2016

3.10. Выводы

Была собрана выборка из 30 годовых отчетов, разделенная на 7390 абзацев, которые прошли предварительную обработку

Для каждого абзаца были построены количественные признаки, признаки TF-IDF, признаки word2vec и признаки doc2vec. В итоге каждый объект выборки имел 25 636 признаков.

Для сбалансированной и несбалансированной выборок построены модели: наивный байесовский классификатор, логистическая регрессия и градиентный бустинг.

В качестве наилучшей модели выбрана модель градиентного бустинга, обученная на несбалансированной выборке с точностью 88,70%.

В качестве примера работы модели был выбран отчет компании PepsiCo за 2016 год. Отчет был разбит на 3420 абзацев, из них 130 абзацев содержали географические точки, далее из 130 абзацев получено 17 релевантных абзацев.

4. СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ

На данный момент безопасность работников на рабочем месте является важным вопросом. Задачами являются охрана здоровья работников, защита от различных видов травм и заболеваний.

Электронно-вычислительные машины (ЭВМ) используются все чаще в различных сферах: на производстве, в исследовательских центрах, в образовании, бизнесе и т.д. Однако, как известно, вредное воздействие компьютеров на человека является основной причиной некоторых заболеваний. Поэтому важно знать о вредном воздействии ЭВМ на организм и мерах защиты.

4.1. Описание рабочего места

В данном разделе рассмотрены вопросы, связанные с организацией рабочего места в соответствии с нормами производственной санитарии, техники производственной безопасности и охраны окружающей среды.

В данной работе рассмотрена проектировка рабочего места и помещения, в котором оно находится.

Под проектированием рабочего места понимается целесообразное пространственное размещение в горизонтальной и вертикальной плоскостях функционально взаимосвязанных средств производства, необходимых для осуществления трудового процесса.

При проектировании рабочих мест должны быть учтены освещенность, температура, влажность, давление, шум, наличие вредных веществ, электромагнитных полей и другие санитарно-гигиенические требования к организации рабочих мест.

При проектировании рабочей зоны необходимо уделить внимание охране окружающей среды, а в частности, организации безотходного производства.

Также необходимо учитывать возможность чрезвычайных ситуаций. Так как рабочая зона находится в городе Томске, наиболее типичной ЧС является

мороз. Так же, в связи с нестабильной ситуаций в мире, одной из возможных ЧС может быть диверсия.

Результатами разработки данного раздела будут являться достижение следующих целей:

- выявление и изучение вредных и опасных производственных факторов при работе с ЭВМ;
- оценка условий труда;
- определение способов снижения действия вредных факторов до безопасных пределов или, по возможности, полного их исключения;
- рассмотрение вопросов техники пожарной безопасности и охраны окружающей среды.

4.2. Анализ опасных и вредных факторов

Вредным называется производственный фактор, воздействие которого на работающего в определенных условиях приводит к заболеванию или снижения работоспособности. При изменении уровня и времени воздействия вредные производственные факторы могут стать опасными.

Опасными считаются производственные факторы, воздействие которых на работающего в конкретных условиях может привести к травмам, а также к другим внезапным резким ухудшениям здоровья.

При работе с ЭВМ пользователь (оператор, программист) подвергается воздействию опасных и вредных производственных факторов:

1. электромагнитных полей;
2. электростатических полей;
3. шуму и вибрации;
4. микроклимат в помещении;
5. освещенность рабочей зоны;
6. психофизиологические факторы.

Эти факторы могут привести к ухудшению здоровья пользователя, а также к профессиональным заболеваниям. Кроме того, вынужденная неудобная рабочая поза, длительное сосредоточенное наблюдение, из которого 20% приходится на непосредственное наблюдение за экраном ВДТ, вызывают повышенное напряжение мышц зрительного аппарата, а в комплексе с неблагоприятными производственными факторами обуславливают развитие общего утомления и снижение работоспособности.

Отрицательное воздействие ЭВМ на человека носит комплексный характер комбинации вредных и опасных производственных факторов:

1. монитор компьютера является источником: электромагнитного поля (ЭМП); электростатического поля; рентгеновского излучения; вредного действия светового потока и отраженного света;
2. значительной нагрузке подвергается зрительный аппарат в результате несовершенства способов создания изображения на экране монитора;
3. работа компьютера сопровождается акустическими шумами, включая ультразвук;
4. несоблюдение эргономических параметров, обеспечивающих безопасность приёмов работы пользователя ЭВМ: гигиенических и психофизиологических, антропометрических и эстетических может повлечь снижение эффективности действий человека.

Характеристика помещения, где была разработана бакалаврская работа: ширина комнаты составляет b= 4 м, длина a= 6 м, высота H = 2,8 м. Тогда площадь помещения будет составлять $S=ab$=24 м^2, объем $V=abh$=73м^3. В помещении имеется окно, через которое осуществляется вентиляция помещения. В помещении отсутствует принудительная вентиляция, т.е. воздух поступает и удаляется через дверь и окно, вентиляция является естественной. В зимнее время помещение отапливается, что обеспечивает достаточное, постоянное и равномерное нагревание воздуха. В помещении используется комбинированное освещение – искусственное и естественное. Искусственное

освещение создается люминесцентными лампами типа ЛБ. Рабочая поверхность имеет высоту 0,75 м. Конструкция стола соответствует нормам СН 245-78. Стол оборудуется специальными ящиками с необходимыми для работы предметами. Электроснабжение сети переменного напряжения 220В. Помещение без повышенной опасности в отношении поражения человека электрическим током по ГОСТ 12.1.013-78.

Компьютер, расположенный на рабочей поверхности высотой 0,77 м, обладает следующими характеристиками: процессор AMDA8, оперативная память 8 Гб, система Microsoft Windows 8.1, частота процессора – 2,00 ГГц, PnP 15,6-й дюймовый монитор с разрешением 1366 на 768 точек и частотой 60 Гц.

Наиболее правильная организация рабочего места позволяет значительно снять напряженность в работе, уменьшить неблагоприятные чрезмерные нагрузки на организм и, как следствие, повысить производительность труда.

Место для работы на компьютере и взаиморасположение всех его элементов должно соответствовать антропометрическим, физическим и психологическим требованиям. При устройстве рабочего места человека, работающего за ПК необходимо соблюсти следующие основные условия: наилучшее местоположение оборудования и свободное рабочее пространство.

Основными элементами рабочего места являются стол и стул, т.к.рабочим положением является положение сидя. Рациональная планировка рабочего места определяет порядок и местоположение предметов, в особенности тех, которые для работ необходимы чаще.

Основные зоны досягаемости рук в горизонтальной плоскости показаны на рисунке 4.1.

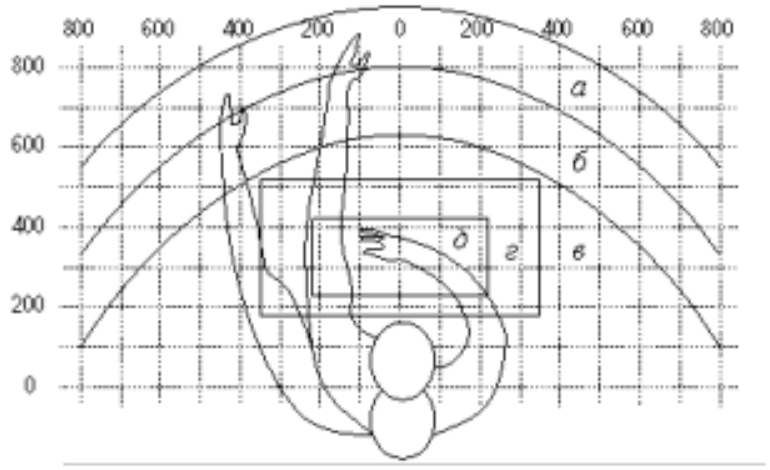


Рисунок 4.1 – Зоны досягаемости рук в горизонтальной плоскости: а – зона максимальной досягаемости; б – зона досягаемости пальцев при вытянутой руке; в – зона легкой досягаемости ладони; г – оптимальное пространство для грудой работы; д – оптимальное пространство для тонкой работы.

В соответствии с этим, принимается следующее оптимальное размещение предметов труда и документации в зонах досягаемости:

1. дисплей размещается в зоне **а** (в центре);
2. системный блок размещается в предусмотренной нише стола;
3. клавиатура – в зоне **г/д;**
4. манипулятор «компьютерная мышь» - в зоне **в** справа;
5. сканер в зоне **а/б** (слева);
6. принтер находится в зоне **а** (справа);
7. документация, необходимая при работе в зоне **в,** а в выдвижных ящиках стола – литература, иногда используемая.

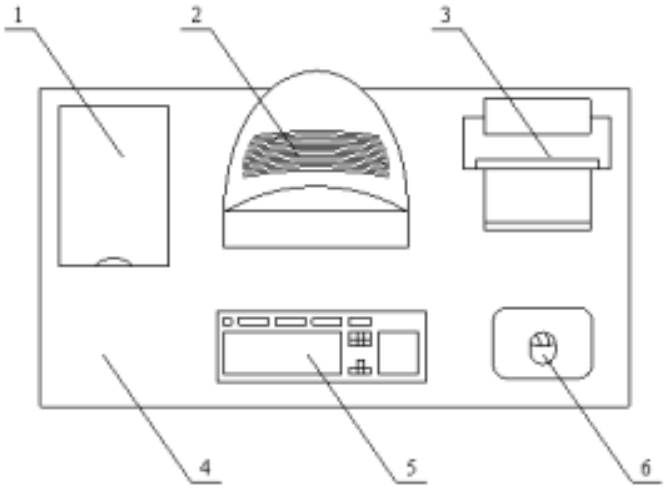


Рисунок 4.2 – Пример размещения основных и периферийных составляющих ПК
на рабочем столе: 1 – сканер, 2 – монитор, 3 – принтер, 4 – поверхность рабочего стола, 5 – клавиатура, 6 – манипулятор типа «мышь».

При проектировании письменного стола должны быть учтены следующие требования.

Высота рабочей поверхности стола рекомендуется в пределах 680–800 мм. Высота рабочей поверхности, на которую устанавливается клавиатура, должна быть 650 мм. Рабочий стол должен быть шириной не менее 700 мм и длиной не менее 1400 мм. Должно иметься пространство для ног высотой не менее 600 мм, шириной — не менее 500 мм, глубиной на уровне колен — не менее 450 мм и на уровне вытянутых ног — не менее 650 мм.

Рабочее кресло должно быть подъёмно-поворотным и регулируемым по высоте и углам наклона сиденья и спинки, а также расстоянию спинки до переднего края сиденья. Рекомендуется высота сиденья над уровнем пола 420–550 мм. Конструкция рабочего кресла должна обеспечивать: ширину и глубину поверхности сиденья не менее 400 мм.

Монитор должен быть расположен на уровне глаз оператора на расстоянии 500–600 мм. Согласно нормам угол наблюдения в горизонтальной плоскости должен быть не более 45º к нормали экрана. Лучше если угол обзора будет составлять 30º. Кроме того должна быть возможность выбирать уровень

контрастности и яркости изображения на экране. Должна предусматриваться возможность регулирования экрана.

Рабочие места с компьютерами должны размещаться так, чтобы расстояние от экрана одного монитора до тыла другого было не менее 2,0 м, а расстояние между боковыми поверхностями мониторов - не менее 1,2 м.

Общие требования к организации и оборудованию рабочих мест с ПЭВМ даны в СанПиН 2.2.2/2.4.1340-03. Все параметры рабочего стола удовлетворяют нормативным требованиям.

Для внутренней отделки интерьера помещений, должны использоваться диффузно отражающие материалы с коэффициентом отражения для потолка - 0,7- 0,8; для стен - 0,5 - 0,6; для пола - 0,3 - 0,5.

Для прекращения неблагоприятного воздействия вредных факторов при работе с ВДТ и ПЭВМ определены санитарно-гигиенические требования к обеспечению безопасных условий труда. Последствия воздействия этих факторов на организм оператора ЭВМ зависят от их интенсивности, продолжительности и режимов действия.

4.3. Микроклимат в помещении

Микроклимат производственных помещений – это климат внутренней среды помещений, который определяется действующими на организм человека сочетаниями температур воздуха и поверхностей, относительной влажности воздуха, скорости движения воздуха и интенсивности теплового излучения. Показатели микроклимата должны обеспечивать сохранение теплового баланса человека с окружающей средой и поддержание оптимального или допустимого теплового состояния организма.

Оптимальные микроклиматические при воздействии на человека в течение рабочей смены обеспечивают сохранение теплового состояния организма и не вызывают отклонений в состоянии здоровья. Допустимые микроклиматические условия могут приводить к незначительным дискомфортным тепловым ощущениям. Возможно временное (в течение рабочей смены) снижение работоспособности, без нарушения здоровья.

Нормы оптимальных и допустимых показателей микроклимата при работе с ЭВМ устанавливает СанПиН 2.2.2/2.4.1340-03 . Все категории работ разграничиваются на основе интенсивности энергозатрат организма в ккал/ч (Вт). Работа, производимая сидя и сопровождающаяся незначительным физическим напряжением, относится к категории Iа – работа с интенсивностью энергозатрат до 120 ккал/ч (до 139 Вт). Для данной категории допустимые нормы микроклимата представлены в таблице 4.1.

Таблица 4.1 – Допустимые нормы микроклимата в рабочей зоне производственных помещений.

Сезон года	Категория тяжести выполняемых работ	Темпаратура, C^0		Относительная влажность, %		Скорость движения воздуха, м/сек	
		Фактическое значение	Допустимое значение	Фактическое значение	Допустимое значение	Фактическое значение	Допустимое значение
Холодный	Iа	(22÷24)	(19÷24)	55	(15÷75)	0,1	≤0,1
Теплый	Iа	(23÷25)	(20÷28)	55	(15÷75)	0,1	≤0,2

Анализируя таблицу 4.1, можно сделать вывод, что в рассматриваемом помещении параметры микроклимата соответствуют нормам СанПиН. Допустимый уровень микроклимата помещения обеспечивается системой водяного центрального отопления и естественной вентиляцией.

В производственных помещениях, где допустимые нормативные величины микроклимата поддерживать не представляется возможным, необходимо проводить мероприятия по защите работников от возможного перегревания и охлаждения. Это достигается различными средствами: применением систем местного кондиционирования воздуха; использованием

индивидуальных средств защиты от повышенной или пониженной температуры; регламентацией периодов работы в неблагоприятном микроклимате и отдыха в помещении с микроклиматом, нормализующим тепловое состояние; сокращением рабочей смены и др.

Профилактика перегревания работников в нагревающем микроклимате включает следующие мероприятия: нормирование верхней границы внешней 46 термической нагрузки на допустимом уровне применительно к 8-часовой рабочей смене; регламентация продолжительности воздействия нагревающей среды (непрерывно и за рабочую смену) для поддержания среднесменного теплового состояния на оптимальном или допустимом уровне; использование специальных СКЗ и СИЗ, уменьшающих поступление тепла извне к поверхности тела человека и обеспечивающих допустимое тепловое состояние работников. Защита от охлаждения осуществляется посредством одежды, изготовленной в соответствии с требованиями ГОСТ 29335—92 и 29338—92 "Костюмы мужские и женские для защиты от пониженных температур. Технические условия".

4.4. Освещенность рабочей зоны

Свет является естественным условием жизни человека. Правильно спроектированное и выполненное освещение обеспечивает высокий уровень работоспособности, оказывает положительное психологическое действие на человека и способствует повышению производительности труда. На рабочей поверхности должны отсутствовать резкие тени, которые создают неравномерное распределение поверхностей с различной яркостью в поле зрения, искажает размеры и формы объектов различия, в результате повышается утомляемость и снижается производительность труда.

Существует три вида освещения: естественное – за счёт солнечного излучения, искусственное – за счёт источников искусственного света и совмещенное – освещение, включающее в себя как естественное, так и искусственное освещения.

Оценка освещенности рабочей зоны проводится в соответствии с СанПиН 2.2.2/2.4.1.1340-03.

В данном рабочем помещении используется комбинированное освещение: искусственное и естественное. Искусственное освещение создается люминесцентными лампами типа ЛД.

Расчёт общего равномерного искусственного освещения горизонтальной рабочей поверхности выполняется методом коэффициента светового потока, учитывающим световой поток, отражённый от потолка и стен. Длина помещения a = 6 м, ширина b = 4 м, высота H =2,8 м. Высота рабочей поверхности над полом h_p = *0,75 м*. Интегральным критерием оптимальности расположения светильников является величина λ, которая для люминесцентных светильников с защитной решёткой лежит в диапазоне 1,1–1,3.

Выбираем лампу дневного света ЛД-40, световой поток которой равен *ФЛД* = 2300 Лм.

Выбираем светильники с люминесцентными лампами типа ОДОР-2-40. Этот светильник имеет две лампы мощностью 40 Вт каждая, длина светильника равна 1227 мм, ширина – 265 мм.

На первом этапе определим значение индекса освещенности *i*.

$$i = \frac{S}{(a+b)h}, \tag{4.1}$$

где S – площадь помещения;h–расчетная высота подвеса светильника, м;aи b–длина и ширина помещения, м.

Высота светильника над рабочей поверхностью h:

$$h = H - h_p - h_c = 2{,}8 - 0{,}75 - 0{,}3 = 1{,}55, \tag{4.2}$$

где H – высота помещения, м; h_p–высота рабочей поверхности, м.

В результате проведенных расчетов, индекс освещенности *I*равен

$$i = \frac{S}{(a+b)h} = \frac{24}{(4+6)\cdot 1{,}55} = 1{,}5. \tag{4.3}$$

Расстояние между соседними светильниками или рядами определяется по формуле:

$$L = \lambda \cdot h = 1{,}1 \cdot 1{,}55 = 1{,}6 \text{ м}. \quad (4.4)$$

Число рядов светильников в помещении:

$$Nb = \frac{b}{L} = \frac{4}{1{,}6} = 2{,}5 \approx 3. \quad (4.5)$$

Число светильников в ряду:

$$Na = \frac{a}{L} = \frac{6}{1{,}6} = 3{,}75 \approx 4. \quad (4.6)$$

Общее число светильников:

$$N = Na \cdot Nb = 4 \cdot 3 = 12. \quad (4.7)$$

Учитывая, что в каждом светильнике установлено две лампы, общее число ламп в помещении N=24.

Расстояние от крайних светильников или рядов до стены определяется по формуле:

$$l = \frac{L}{3} = \frac{1{,}6}{3} = 0{,}53 \text{ м} \quad (4.8)$$

Размещаем светильники в три ряда. План помещения и размещения светильников с люминесцентными лампами представлен в приложении А.

Световой поток лампы определяется по формуле:

$$\Phi = \frac{E_H \cdot S \cdot K_з \cdot Z}{N \cdot \eta}, \quad (4.9)$$

где $E_н$ – нормируемая минимальная освещенность по СНиП 23-05-95, лк; S– площадь освещаемого помещения, м^2; $K_з$ – коэффициент запаса, учитывающий загрязнение светильника (источника света, светотехнической арматуры, стен и пр., т.е. отражающих поверхностей), наличие в атмосфере цеха дыма, пыли;Z– коэффициент неравномерности освещения, отношение $E_{ср}/E_{min}$. Для люминесцентных ламп при расчетах берется равным 1,1; N – число ламп в помещении; η – коэффициент использования светового потока.

Данное помещение относится к типу помещения со средним выделением пыли, в связи с этим $K_з$ = 1,5; состояние потолка – свежепобеленный, поэтому значение коэффициента отражения потолка $\rho_п$=70; состояние стен – побеленные бетонные стены, поэтому значение коэффициента отражения стен $\rho_с$ = 50.

Коэффициент использования светового потока, показывающий какая часть светового потока ламп попадает на рабочую поверхность, для светильников типа ОДОР с люминесцентными лампами при ρ_n=70%, ρ_c = 50% и индексе помещения i=1,5 равен η=0,47.

Нормируемая минимальная освещенность при использовании ЭВМ и одновременной работе с документами должна быть равна 600лк.

$$\Phi = \frac{E_H \cdot S \cdot K_з \cdot Z}{N \cdot \eta} = \frac{600 \cdot 24 \cdot 1{,}5 \cdot 1{,}1}{24 \cdot 0{,}47} = 2106 \text{ Лм} \qquad (4.10)$$

Для люминесцентных ламп с мощностью 40 Вт и напряжением сети 220В, стандартный световой поток ЛД равен 2300 Лм.

$$-10\% \leq \frac{\Phi_{ЛД} - \Phi_{л.расч}}{\Phi_{ЛД}} \cdot 100\% \leq 20\%,$$

$$\frac{2300 - 2106}{2300} \cdot 100\% = 8{,}43\%, \qquad (4.11)$$

$$-10\% \leq 8{,}43\% \leq 20\%.$$

Таким образом необходимый световой поток светильника не выходит за пределы требуемого диапазона.

4.5. Электромагнитное поле

ЭМП обладает способностью биологического, специфического и теплового воздействия на организм человека, что может повлечь следующие последствия: биохимические изменения в клетках и тканях; нарушения условно- рефлекторной деятельности, снижение биоэлектрической активности мозга, изменения межнейронных связей, отклонения в эндокринной системе; вследствие перехода ЭМП в тепловую энергию может наблюдаться повышение температуры тела, локальный избирательный нагрев тканей и так далее.

Согласно СанПиН 2.2.2.542-96:

1. Напряженность электромагнитного поля на расстоянии 50 см вокруг ВДТ по электрическое составляющей должна быть не более:
    - в диапазоне частот 5 Гц - 2 кГц – 25 В/м;
    - в диапазоне частот 2 кГц/400 кГц – 2,5 В/м.

2. Плотность магнитного потока должна быть не более:
    - в диапазоне частот 5 Гц - 2 кГц – 250 нТл;
    - в диапазоне частот 2 кГц/400 кГц – 25 нТл.

Защите человека от опасного воздействия электромагнитного излучения осуществляется следующими способами:

1. Применение СКЗ
    - защита временем;
    - защита расстоянием;
    - снижение интенсивности излучения непосредственно в самом источнике излучения;
    - экранирование источника;
    - защита рабочего места от излучения;
2. Применение средств индивидуальной защиты (СИЗ), которые включают в себя:
    - Очки и специальная одежда, выполненная из металлизированной ткани (кольчуга). При этом следует отметить, что использование СИЗ возможно при кратковременных работах и является мерой аварийного характера. Ежедневная защита обслуживающего персонала должна обеспечиваться другими средствами;
    - Вместо обычных стекол используют стекла, покрытые тонким слоем золота или диоксида олова (SnO_2).

Экранирование источника излучения и рабочего места осуществляется специальными экранами по ГОСТ 12.4.154.

4.6. Электростатическое поле

Электризация заключается в следующем: нейтральные тела, в нормальном состоянии не проявляющие электрических свойств, при условии отрицательных контактов или взаимодействий становятся электрозаряженными. Опасность возникновения статического электричества

проявляется в возможности образования электрической искры и вредном воздействии его на человеческий организм, и не только в случае непосредственного контакта с зарядом, но и за счет действий электрического поля, которое возникает при заряде. При включенном питании компьютера на экране дисплея накапливается статическое электричество. Электрический ток искрового разряда статического электричества мал и не может вызвать поражение человека. Тем не менее, вблизи экрана электризуется пыль и оседает на нем. В результате чего искажается резкость восприятия информации на экране. Кроме того, пыль попадает на лицо работающего и в его дыхательные пути. Основные способы защиты от статического электричества следующие: заземление оборудования, увлажнение окружающего воздуха. Также целесообразно применение полов из антистатического материала.

4.7. Электробезопасность

Электробезопасность представляет собой систему организационных и технических мероприятий и средств, обеспечивающих защиту людей от вредного и опасного воздействия электрического тока, электрической дуги, электромагнитного поля и статистического электричества.

Электроустановки классифицируют по напряжению: с номинальным напряжением до 1000 В (помещения без повышенной опасности), до 1000 В с присутствием агрессивной среды (помещения с повышенной опасностью) и свыше 1000 В (помещения особо опасные).

В отношении опасности поражения людей электрическим током различают:

1. Помещения без повышенной опасности, в которых отсутствуют условия, создающие повышенную или особую опасность.
2. Помещения с повышенной опасностью, которые характеризуются наличием в них одного из следующих условий, создающих повышенную опасность: сырость, токопроводящая пыль, токопроводящие полы (металлические, земляные, железобетонные, кирпичные и т.п.), высокая температура, возможность

одновременного прикосновения человека к имеющим соединение с землей металлоконструкциям, технологическим аппаратам, с одной стороны, и к металлическим корпусам электрооборудования - с другой.

3. Особо опасные помещения, которые характеризуются наличием оборудования свыше 1000 В и одного из следующих условий, создающих особую опасность: особой сырости, химически активной или органической среды, одновременно двух или более условий повышенной опасности. Территории размещения наружных электроустановок в отношении опасности поражения людей электрическим током приравниваются к особо опасным помещениям.

Помещение, где была разработана бакалаврская работа, принадлежит к категории помещений без повышенной опасности по степени вероятности поражения электрическим током, вследствие этого к оборудованию предъявляются следующие требования:

- экран монитора должен находится на расстоянии не менее 50 см от пользователя (расстояния от источника);
- применение приэкранных фильтров, специальных экранов.

Защитное заземление — это преднамеренное электрическое соединение с землей или ее эквивалентом металлических нетоковедущих частей, которые могут оказаться под напряжением.

Сопротивление заземления – основной показатель заземляющего устройства, определяющий его способность выполнять свои функции и определяющий его качество в целом.

Сопротивление заземления зависит от площади электрического контакта заземлителя (заземляющих электродов) с грунтом (“стекание” тока) и удельного электрического сопротивления грунта, в котором смонтирован этот

заземлитель (“впитывание” тока). Согласно ПЭУ номинальное сопротивление заземления должно быть не более 4 Ом.

К основным электрозащитным средствам в электроустановках напряжением до 1000 В относятся:

- изолирующие штанги;
- изолирующие и электроизмерительные клещи;
- диэлектрические перчатки; изолированный инструмент.

Работать со штангой разрешается только специально обученному персоналу в присутствии лица, контролирующего действия работающего. При операциях с изолирующей штангой необходимо пользоваться дополнительными изолирующими защитными средствами – диэлектрическими перчатками и изолирующими основаниями (подставками, ковриками) или диэлектрическими ботами.

Изолирующие клещи применяют в электроустановках до 35 кВ для операций под напряжением с плавкими вставками трубчатых предохранителей, а также для надевания и снятия изолирующих колпаков на ножи однополюсных разъединителей. Изолирующие клещи выполняют из пластмассы.

При пользовании изолирующими клещами оператор должен надевать диэлектрические перчатки и быть изолированным от пола или грунта; при смене патронов трубчатых предохранителей он должен быть в очках. Клещи нужно держать в вытянутых руках.

К дополнительным изолирующим электрозащитным средствам относятся диэлектрические перчатки, боты, резиновые коврики и дорожки, изолирующие подставки на фарфоровых изоляторах и переносные заземления.

Перед началом работы следует убедиться в отсутствии свешивающихся со стола или висящих под столом проводов электропитания, в целостности вилки и провода электропитания, в отсутствии видимых повреждений аппаратуры и рабочей мебели, в отсутствии повреждений и наличии заземления приэкранного фильтра.

4.8. Производственный шум и вибрация

Вентиляция производственных помещений предназначена для уменьшения запыленности, задымленности и очистки воздуха от вредных выделений производства, а также для сохранности оборудования. Она служит одним из главных средств оздоровления условий труда, повышения производительности и предотвращения опасности профессиональных заболеваний. Система вентиляции обеспечивает снижение содержания в воздухе помещения пыли, газов до концентрации не превышающей ПДК. Проветривание помещения проводят, открывая форточки. Проветривание помещений в холодный период года допускается не более однократного в час, при этом нужно следить, чтобы не было снижения температуры внутри помещения ниже допустимой. Воздухообмен в помещении можно значительно сократить, если улавливать вредные вещества в местах их выделения, не допуская их распространения по помещению. Для этого используют приточно-вытяжную вентиляцию. Кратность воздухообмена не ниже 3.

Предельно допустимый уровень (ПДУ) шума – это уровень фактора, который при ежедневной (кроме выходных дней) работе, но не более 40 часов в неделю в течение всего рабочего стажа, не должен вызывать заболеваний или отклонений в состоянии здоровья, обнаруживаемых современными методами исследований в процессе работы или в отдаленные сроки жизни настоящего и последующих поколений. Соблюдение ПДУ шума не исключает нарушения здоровья у сверхчувствительных лиц.

Допустимый уровень шума ограничен ГОСТ 12.1.003-83 и СанПиН 2.2.4/2.1.8.10-32-2002. Уровень шума на рабочем месте математиков-программистов и операторов видеоматериалов не должен превышать 50дБА, а в залах обработки информации на вычислительных машинах - 65дБА.

При значениях выше допустимого уровня необходимо предусмотреть СКЗ и СИЗ.

1. СКЗ

- устранение причин шума или существенное его ослабление в источнике образования;
- изоляция источников шума от окружающей среды средствами звуко- и виброизоляции, звуко- и вибропоглощения;
- применение средств, снижающих шум и вибрацию на пути их распространения;

2. СИЗ

- применение спецодежды, спецобуви и защитных средств органов слуха: наушники, беруши, антифоны.

Защита от шумов – заключение вентиляторов в защитный кожух и установление их внутри корпуса ЭВМ. Для снижения уровня шума стены и потолок помещений, где установлены компьютеры, могут быть облицованы звукопоглощающими материалами с максимальными коэффициентами звукопоглощения в области частот 63 - 8000 Гц.

Вибрация оборудования на рабочих местах не должна превышать допустимых величин, установленных ГОСТ 12.1.012-96.

4.9. Психофизиологические факторы и опасные факторы

Значительное умственное напряжение и другие нагрузки приводят к переутомлению функционального состояния центральной нервной системы, нервно-мышечного аппарата рук. Нерациональное расположение элементов рабочего места вызывает необходимость поддержания вынужденной рабочей позы. Длительный дискомфорт вызывает повышенное позвоночное напряжение мышц и обуславливает развитие общего утомления и снижение работоспособности.

При длительной работе за экраном дисплея появляется выраженное напряжение зрительного аппарата с появлением жалоб на неудовлетворительность работы, головные боли, усталость и болезненное ощущение в глазах, в пояснице, в области шеи, руках.

Режим труда и отдыха работника: при вводе данных, редактировании программ, чтении информации с экрана непрерывная продолжительность работы не должна превышать 4-х часов при 8-часовом рабочем дне. Через каждый час работы необходимо делать перерыв на 5-10 минут, а через два часа на 15 минут.

С целью снижения или устранения нервно-психологического, зрительного и мышечного напряжения, предупреждение переутомления необходимо проводить комплекс физических упражнений и сеансы психофизической разгрузки и снятия усталости во время регламентируемых перерывов, и после окончания рабочего дня.

4.10. Охрана окружающей среды

Охрана окружающей среды – это комплексная проблема и наиболее активная форма её решения - это сокращение вредных выбросов промышленных предприятий через полный переход к безотходным или малоотходным технологиям производства.

С точки зрения потребления ресурсов компьютер потребляет сравнительно небольшое количество электроэнергии, что положительным образом сказывается на общей экономии потребления электроэнергии в целом.

Основными отходами являются черновики бумаги и отработавшие люминесцентные лампы. Бумагу направляют на утилизацию, а люминесцентные лампы собирают и направляют на утилизацию в соответствующую организацию.

При выполнении бакалаврской работы никакого ущерба окружающей среде нанесено не было.

4.11. Защита в чрезвычайных ситуациях

В Томске преобладает континентально-циклонический климат. Природные явления (землетрясения, наводнения, засухи, ураганы и т. д.) отсутствуют.

Возможными ЧС могут быть сильные морозы и диверсия.

Для Сибири в зимнее время года характерны морозы. Достижение критически низких температур приведет к авариям систем теплоснабжения и жизнеобеспечения, приостановке работы, обморожениям и даже жертвам среди населения. В случае переморозки труб должны быть предусмотрены запасные обогреватели. Их количества и мощности должно хватать для того, чтобы работа на производстве не прекратилась.

Чрезвычайные ситуации, возникающие в результате диверсий, возникают все чаще. Зачастую такие угрозы оказываются ложными. Но случаются взрывы и в действительности.

Для предупреждения вероятности осуществления диверсии предприятие необходимо оборудовать системой видеонаблюдения, круглосуточной охраной, пропускной системой, надежной системой связи, а также исключения распространения информации о системе охраны объекта, расположении помещений и оборудования в помещениях, системах охраны, сигнализаторах, их местах установки и количестве. Должностные лица раз в полгода проводят тренировки по отработке действий на случай экстренной эвакуации.

4.12. Правовые и организационные вопросы обеспечения безопасности:

Специальные (характерные для проектируемой рабочей зоны) правовые нормы трудового законодательства:

- СанПиН 2.2.2.542-96 «Гигиенические требования к видеодисплейным терминалам, персональным электронно-вычислительным машинам и организации работы»;
- СанПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы»;
- СНиП-23-05-95 «ЕСТЕСТВЕННОЕ И ИСКУССТВЕННОЕ ОСВЕЩЕНИЕ»;
- ГОСТ 12.1.036-96 «ССБТ. Шум. Допустимые уровни в жилых и общественных зданиях»;

- ГОСТ 12.1.012-96 «ССБТ. Вибрационная безопасность. Общие требования»;
- ГОСТ 12.1.004-76 «Пожарная безопасность. Общие требования (с Изменением N 1)»;
- ГОСТ 12.1.010-76 «Система стандартов безопасности труда (ССБТ). Взрывобезопасность. Общие требования (с Изменением N 1)»;
- ГОСТ 12.1.013-78 «ССБТ. Строительство. Электробезопасность. Общие требования».

4.13. Выводы и рекомендации

Проанализировав условия труда на рабочем месте, где была разработана бакалаврская работа, можно сделать вывод, что помещение удовлетворяет необходимым нормам и в случае соблюдения техники безопасности и правил пользования компьютером работа в данном помещении не приведет к ухудшению здоровья работника.

Само помещение и рабочее место в нем удовлетворяет всем нормативным требованиям. Кроме того, действие вредных и опасных факторов сведено к минимуму, т.е. микроклимат, освещение и электробезопасность соответствуют требованиям, предъявленным в соответствующих нормативных документах.

Относительно рассмотренного вопроса об экологической безопасности можно сказать, что деятельность помещения не представляет опасности окружающей среде.

Важно добавить, что монитор компьютера служит источником ЭМП – вредного фактора, который отрицательно влияет на здоровье работника при продолжительной непрерывной работе и приводит к снижению работоспособности. Поэтому во избежание негативного влияния на здоровье необходимо делать перерывы при работе с ЭВМ и проводить специализированные комплексы упражнений для глаз.

5 ОЦЕНКА КОММЕРЧЕСКОГО ПОТЕНЦИАЛА И ПЕРСПЕКТИВНОСТИ ПРОВЕДЕНИЯ НАУЧНЫХ ИССЛЕДОВАНИЙ С ПОЗИЦИИ РЕСУРСОЭФФЕКТИВНОСТИ И РЕСУРСОСБЕРЕЖЕНИЯ

5.1 Потенциальные потребители результатов исследования

Для анализа потребителей результатов исследования необходимо рассмотреть целевой рынок и провести его сегментирование.

Целевой рынок – сегменты рынка, на котором будет продаваться в будущем разработка. В свою очередь, сегмент рынка – это особым образом выделенная часть рынка, группы потребителей, обладающих определенными общими признаками.

Сегментирование – это разделение покупателей на однородные группы, для каждой из которых может потребоваться определенный товар (услуга).

В зависимости от категории потребителей (коммерческие организации, физические лица) необходимо использовать соответствующие критерии сегментирования. Например, для коммерческих организаций критериями сегментирования могут быть: месторасположение, отрасль, выпускаемая продукция, размер и др. Для физических лиц: возраст, пол, национальность, образование, уровень дохода, социальная принадлежность, профессия.

Потенциальные потребители результатов исследования:

Услуги по предоставлению ПО:

- российские брокерские компании;
- иностранные брокерские компании;
- российские частные инвесторы;
- иностранные частные инвесторы;
- банки;
- консалтинговые фирмы.

Услуги по внедрению ПО в бизнес-модель:

- российские брокерские компании;
- иностранные брокерские компании;

- российские частные инвесторы;
- иностранные частные инвесторы;
- банки;
- консалтинговые фирмы.

	Виды использования сформированных портфелей		
Виды потребителей		Предоставление ПО	Внедрение ПО в бизнес-модель
	Российские брокерские компании		
	Иностранные брокерские компании		
	Российские частные инвесторы		
	Иностранные частные инвесторы		
	Банки		
	Консалтинговые фирмы		

Фирма А - □, Фирма Б - ■

Рисунок 5.1. Карта сегментирования рынка услуг по виду предоставляемых услуг

5.2 Анализ конкурентных технических решений

Детальный анализ конкурирующих разработок, существующих на рынке, необходимо проводить систематически, поскольку рынки пребывают в

постоянном движении. Такой анализ помогает вносить коррективы в научное исследование, чтобы успешнее противостоять своим соперникам. Важно реалистично оценить сильные и слабые стороны разработок конкурентов.

Анализ конкурентных технических решений с позиции ресурсоэффективности и ресурсосбережения позволяет провести оценку сравнительной эффективности научной разработки и определить направления для ее будущего повышения.

Основными конкурентами являются организации, деятельность которых связанна с использованием вычислительной техники и информационных технологий и последующим написанием программного обеспечения.

Таблица 5.1 – Оценочная карта для сравнения конкурентных технических решений

Критерии оценки	**Вес крите-рия**	**Баллы**			**Конкуренто-способность**		
		$Б_ф$	$Б_{к1}$	$Б_{к2}$	$К_ф$	$К_{к1}$	$К_{к2}$
Технические критерии оценки ресурсоэффективности							
1. Повышение производительности труда пользователя	0,08	5	4	4	0,4	0,32	0, 32
2. Удобство в эксплуатации (соответствует требованиям потребителей)	0,09	4	5	4	0,36	0,45	0, 36
3. Помехоустойчивость							
4. Энергоэкономичность							
5. Надежность	0,08	4	4	4	0,32	0,32	0, 32
6. Уровень шума							
7. Безопасность	0,08	4	5	4	0,32	0,4	0, 32
8. Потребность в ресурсах памяти	0,04	5	4	3	0,2	0,16	0, 12
9. Функциональная мощность (предоставляемые возможности)							
10. Простота эксплуатации	0,08	4	5	4	0,32	0,4	0, 32
11. Качество интеллектуального интерфейса	0,09	4	4	3	0,36	0,36	0, 27
12. Возможность подключения в сеть ЭВМ							
Экономические критерии оценки эффективности							

1. Конкурентоспособность продукта	0,09	4	4	3	0,36	0,36	0,27
2. Уровень проникновения на рынок	0,08	4	5	4	0,32	0,4	0,032
3. Цена	0,09	5	3	4	0,45	0,27	0,36
4. Предполагаемый срок эксплуатации							
5. Послепродажное обслуживание	0,08	5	4	4	0,4	0,32	0,32
6. Финансирование научной разработки	0,07	5	5	4	0,35	0,35	0,28
7. Срок выхода на рынок	0,05	4	5	5	0,2	0,25	0,25
8. Наличие сертификации разработки							
Итого	**1**				**4,36**	**4,36**	**3,83**

Критерии для сравнения и оценки ресурсоэффективности и ресурсосбережения, приведенные в таблице, подбираются, исходя из выбранных объектов сравнения с учетом их технических и экономических особенностей разработки, создания и эксплуатации.

Позиция разработки и конкурентов оценивается по каждому показателю экспертным путем по пятибалльной шкале, где 1 – наиболее слабая позиция, а 5 39 – наиболее сильная. Веса показателей, определяемые экспертным путем, в сумме должны составлять 1.

Анализ конкурентных технических решений определяется по формуле:

$$K = \sum \text{В}_i \text{Б}_i, \qquad (5.1)$$

где K – конкурентоспособность научной разработки или конкурента; B_i–вес показателя (в долях единица); $Б_\text{i}$ – балл i-го показателя.

Основываясь на знаниях о конкурентах, можно объяснить, что большинство моделей учитываются только ретроспективные (предварительно предусмотренные) данные, в то время как воздействие может оказываться и внешними факторами, которые не рассматриваются, но вносят в полученные

результаты и их точность. Поэтому необходимо учитывать и анализировать внешние факторы для построения более точной модели.

5.3 SWOT-анализ

SWOT – Strengths (сильные стороны), Weaknesses (слабые стороны), Opportunities (возможности) и Threats (угрозы) – представляет собой комплексный анализ научно-исследовательского проекта. SWOT-анализ применяют для исследования внешней и внутренней среды проекта. Разработанная для данного исследования матрица SWOT представлена в Таблице 5.2.

Таблица 5.2 – SWOT-анализ

Внутренняя среда Внешняя среда	Сильные стороны научно-исследовательского проекта: С1. Модель универсальна для различных предприятий. С2. Задача географической диверсификации облегчается для аналитиков.	Слабые стороны научно-исследовательского проекта: С1. При появлении в отчетах слов, с которыми модель не работала на тренировочной выборке, модель может повести себя непредсказуемо. С2. Данных, готовые к обучению, нет. Нужно вручную размечать их, что довольно трудоемко и рутинно.
Возможности: В1. Невысокий уровень конкуренции. В2. Данные для обучения модели всегда в открытом доступе.	Поскольку данных много, можно бесконечно улучшать качество и универсальность модели.	Чтобы снизить вероятность неадекватного поведения модели, можно ее постоянно

		дообучать данными, однако нужно отдать задачу разметки данных на аутсорсинг.
Угрозы: У1. Аналогичную разработку может создать любая IT-компания У2. Низкий спрос у потребителя, поскольку задача очень специфична.	Клиентов можно привлекать лично из-за специфичности задачи. В случае появления конкурентов нужно сравнивать основные метрики своей и чужой модели (точность модели, обобщающая способность и т.д.).	Постоянная работа с данными повышает уровень теоретических и практических знаний о предмете, позволяет применять и исследовать новые методы построения модели.

Выводы:

SWOT-анализ используется для оценки факторов и явлений, влияющих на деятельность компании, а также на возникновение кризисных ситуаций. Для SWOT-анализа актуальны не все существующие на рынке возможности, а только те, которые можно использовать в данном случае. Преимущество SWOT-анализа заключается в том, что аналитическая работа не зациклена только на финансовом состоянии или на анализе конкурентов, а связывает разнообразные факторы внешней и внутренней среды воедино.

5.4 Планирование научно-исследовательских работ

5.4.1 Структура работ в рамках научного исследования

Планирование комплекса предполагаемых работ осуществляется в следующем порядке:

- определение структуры работ в рамках научного исследования;
- определение участников каждой работы;

- установление продолжительности работ;
- построение графика проведения научных исследований.

Для выполнения научных исследований сформирована рабочая группа, в состав которой входят руководитель и инженер. По каждому виду запланированных работ устанавливается соответствующая должность исполнителей. В данном разделе был составлен перечень этапов и работ в рамках проведения научного исследования, а также проведено распределение исполнителей по видам работ. Примерный порядок составления этапов и работ, распределение исполнителей по данным видам работ приведен в Таблице 5.3.

Таблица 5.3 – Комплекс работ по разработке проекта

Основные этапы	№ Раб	Содержание работ	Должность Исполнителя
Разработка технического задания	1	Составление и утверждение научного задания	Руководитель
Выбор направления исследований	2	Подбор и изучение материалов по теме	Инженер
	3	Выбор направления исследований	Инженер
	4	Календарное планирование работ по теме	Руководитель
Теоретические и экспериментальные исследования	5	Корректно и некорректно поставленные задачи	Инженер
	6	Несовместные и плохо обусловленные СЛАУ и их сингулярный анализ	Инженер
	7	Предварительная обработка данных	Инженер
	8	Построение математической	Инженер

		модели	
Обобщение и оценка результатов	9	Оценка эффективности полученных результатов	Руководитель
Оформление отчета по ВКР	10	Составление пояснительной записки к ВКР	Инженер
	11	Оформление пояснительной записки к ВКР по ГОСТу	Инженер

5.4.2 Определение трудоемкости выполнения работ и разработка графика проведения научного исследования

Трудовые затраты в большинстве случаев образуют основную часть стоимости разработки, поэтому важным моментом является определение трудоемкости работ каждого из участников научного исследования.

Трудоемкость выполнения научного исследования оценивается экспертным путем в человеко-днях и носит вероятностный характер, т.к. зависит от множества трудно учитываемых факторов. Для определения ожидаемого (среднего) значения трудоемкости используется следующая формула:

$$t_{\text{ож}\,i} = \frac{3t_{\min i} + 2t_{\max i}}{5}, \qquad (5.2)$$

где $t_{ожi}$ – ожидаемая трудоемкость выполнения i-й работы, человеко-дни;t_{mini} – минимально возможная трудоемкость выполнения заданной i-й работы (оптимистическая оценка: в предположении наиболее благоприятного стечения обстоятельств), человеко-дни;t_{naxi} – максимально возможная трудоемкость выполнения заданной i-й работы (пессимистическая оценка: в предположении наиболее неблагоприятного стечения обстоятельств), человеко-дни;

Рассчитаем значение ожидаемой трудоемкости работы.

Установление длительности работ в рабочих днях осуществляется по формуле:

$$t_{pi} = \frac{t_{\text{ож}\,i}}{\text{Ч}_i}, \tag{5.3}$$

где t_{pi} − трудоемкость работы, человеко-дни; $Ч_i$ – численность исполнителей, выполняющих одновременно одну и ту же работу на данном этапе, чел.

При выполнении дипломных работ студенты в основном становятся участниками сравнительно небольших по объему научных тем. Поэтому наиболее удобным и наглядным является построение ленточного графика проведения научных работ в форме диаграммы Ганта.

Диаграмма Ганта – горизонтальный ленточный график, на котором работы по теме представляются протяженными во времени отрезками, характеризующимися датами начала и окончания выполнения данных работ.

Для удобства построения графика, длительность каждого из этапов работ из рабочих дней следует перевести в календарные дни. Для этого необходимо воспользоваться формулой:

$$T_{ki} = T_{pi} \cdot K_{\text{кал}}, \tag{5.4}$$

где T_{ki} − продолжительность выполнения работы в календарных днях; T_{pi} − продолжительность выполнения работы в рабочих днях; $K_{кал}$ – коэффициент календарности, предназначен для перевода рабочего времени в календарное.

Коэффициент календарности определяется по формуле:

$$K_{\text{кал}} = \frac{T_{\text{кал}}}{T_{\text{кал}} - T_{\text{пр}} - T_{\text{вых}}}, \tag{5.5}$$

где $T_{кал}$ – календарное число дней в году; $T_{пр}$, $T_{вых}$ – число праздничных и выходных дней в году.

Рассчитанные значения в календарных днях по каждой работе необходимо округлить до целого числа.

Вычислим коэффициент календарности:

$$K_{\text{кал}} = \frac{T_{\text{кал}}}{T_{\text{кал}} - T_{\text{пр}} - T_{\text{вых}}} = \frac{365}{365 - 10 - 104} = 1{,}45.$$

Таблица 5.4 – Временные показатели осуществления комплекса работ

№ работы	Продолжительность работ			Исполнители	tpi, человеко-дни	tki, человеко-дни
	tmin i, человеко-дни	tmax i, человеко-дни	тож i, человеко-дни			
1	1	3	2	Б, Р	1	1
2	14	18	16	Б	16	23
3	7	12	9	Б	9	13
4	3	6	4	Б, Р	2	3
5	2	5	3	Б	3	4
6	10	16	12	Б	12	17
7	5	7	6	Б	6	9
8	3	5	4	Б, Р	2	3
9	5	11	7	Б	7	10
10	4	7	5	Б	5	7

Календарный план-график выполнения работ представим в виде таблицы:

Таблица 5.5 – Календарный план-график выполнения работ

Календарный план-график выполнения работ по теме														
№ работы	Наименование работы	Исполнители	t_{ki}, дни	Продолжительность выполнения работ, дни										
				Март			Апрель				Май			
				1	23	13	3	4	17	9	3	10	7	
1	Составление и утверждение ТЗ	Б	1											
		Р												
2	Подбор и изучение материалов по теме	Б	23											

3	Анализ исходных данных	Б	13			■							
4	Выбор метода выполнения работы	Б	3				■						
		Р					■						
5	Календарное планирование работ по теме	Б	4					■					
6	Применение выбранного метода к данным	Б	17						■				
7	Анализ результатов работы	Б	9							■			
8	Определение целесообразности проведения НИР	Б	3								■		
		Р									■		
9	Составление пояснительной записки к ВКР	Б	10									■	
10	Оформление пояснительной записки к ВКР по ГОСТу	Б	7										■

Руководитель ■ , Бакалавр ■.

5.5. Бюджет научно-технического исследования

При планировании бюджета НТИ должно быть обеспечено полное и достоверное отражение всех видов расходов, связанных с его выполнением. В

процессе формирования бюджета НТИ используется следующая группировка затрат по статьям:

- материальные затраты НТИ;
- затраты на специальное оборудование для научных (экспериментальных) работ;
- основная заработная плата исполнителей темы;
- дополнительная заработная плата исполнителей темы;
- отчисления во внебюджетные фонды (страховые отчисления);
- затраты научные и производственные командировки;
- контрагентные расходы; − накладные расходы.

5.5.1. Затраты на материалы

Данная статья отражает стоимость всех материалов, используемых при разработке проекта, включая расходы на их приобретение и доставку. Транспортные расходы принимаются в пределах 3-5% от стоимости материалов. В материальные затраты, помимо вышеуказанных, включаются дополнительно затраты на канцелярские принадлежности, диски, картриджи и т.п. Однако их учет ведется в данной статье только в том случае, если в научной организации их не включают в расходы на использование оборудования или накладные расходы.

Расчет затрат на материалы производится по форме, приведенной в таблице 5.6.

Таблица 5.6 – Материальные затраты

Наименование	Единица измерения	Количество	Цена за ед., руб.	Затраты на материалы ($З_м$), руб.
Бумага	Пачка	1	250	250
Картридж для принтера	Шт	1	2500	2500
Канцелярские	Шт	1	300	300

принадлежности				
Компьютер	Шт	1	30 000	0
Итого			33 050	3050

5.5.2. Основная заработная плата

Величина расходов по заработной плате определяется исходя из трудоемкости выполняемых работ и действующей системы окладов и тарифных ставок. В состав основной заработной платы включается премия, выплачиваемая 72 ежемесячно из фонда заработной платы в размере 20 – 30 % от тарифа или оклада.

Статья включает основную заработную плату работников, непосредственно занятых выполнением НТИ, (включая премии, доплаты) и дополнительную заработную плату:

$$З_{зп} = З_{осн} + З_{доп}, \quad (5.6)$$

где $З_{осн}$ – основная заработная плата; $З_{доп}$ – дополнительная заработная плата.

Основная заработная плата ($З_{осн}$) руководителя (лаборанта, инженера) от предприятия (при наличии руководителя от предприятия) рассчитывается по следующей формуле:

$$З_{осн} = З_{дн} \cdot T_р, \quad (5.7)$$

где $З_{осн}$ – основная заработная плата одного работника; $T_р$ – продолжительность работ, выполняемых научно-техническим работником, раб. дн.; $З_{дн}$ – среднедневная заработная плата работника, руб.

Среднедневная заработная плата рассчитывается по формуле:

$$З_{дн} = \frac{З_м}{Г_д}, \quad (5.8)$$

где $З_м$ – месячный должностной оклад работника, руб.; $Г_д$ – количество рабочих дней в месяце.

Месячный должностной оклад работника:

$$З_м = З_{тс} \cdot (1 + k_{пр} + k_д)k_р, \quad (5.9)$$

где $З_{тс}$ – заработная плата по тарифной ставке, руб.; $k_{пр}$ – премиальный коэффициент, равный 0,3 (т.е. 30% от $З_{тс}$); $k_д$ – коэффициент доплат и надбавок составляет примерно 0,2 – 0,5 (в НИИ и на промышленных предприятиях – за расширение сфер об- служивания, за профессиональное мастерство, за вредные условия: 15- 20 % от $З_{тс}$); $k_р$ – районный коэффициент, равный 1,3 г.Томск.

Пример расчета заработной платы для руководителя:

$$З_м = З_{тс} \cdot (1 + k_{пр} + k_д) k_р$$
$$= 23264{,}86 \cdot (1 + 0{,}3 + 0{,}2) \cdot 1{,}3$$
$$= 45366{,}48 \text{ руб.}$$
$$З_{осн} = З_{дн} \cdot T_р = 2160{,}31 \cdot 7 = 15122{,}16 \text{ руб.}$$

Таблица 5.7 – Расчет основной заработной платы

Исполнители	$З_{тс}$, руб.	$k_р$	$З_м$, руб.	$З_{дн}$, руб.	$Т_р$, дни	$З_{осн}$, руб.
Руководитель	23264,86	1,3	45366,48	2160,31	7	15122,16
Бакалавр	3300	0	3300	157	96	15072
					ИТОГО:	30194,16

5.5.3. Дополнительная заработная плата

Затраты по дополнительной заработной плате исполнителей темы учитывают величину предусмотренных Трудовым кодексом РФ доплат за 74 отклонение от нормальных условий труда, а также выплат, связанных с обеспечением гарантий и компенсаций (при исполнении государственных и общественных обязанностей, при совмещении работы с обучением, при предоставлении ежегодного оплачиваемого отпуска и т.д.).

Расчет дополнительной заработной платы ведется по следующей формуле:

$$З_{доп} = k_{доп} \cdot З_{осн}, \quad (5.10)$$

где $k_{доп}$ – коэффициент дополнительной заработной платы (на стадии проектирования принимается равным 0,12 – 0,15).

Таблица 5.8 – Расчет дополнительной заработной платы

Исполнители	Основная ЗП, руб.	Дополнительная ЗП, руб.
Руководитель (доцент)	15122,16	1814,66
Бакалавр	15072	1808,64
	ИТОГО:	3623,3

5.5.4. Отчисления во внебюджетные фонды

Отчисления во внебюджетные фонды являются обязательными по установленным законодательством Российской Федерации нормам органам государственного социального страхования (ФСС), пенсионного фонда (ПФ) и медицинского страхования (ФФОМС) от затрат на оплату труда работников.

Величина отчислений во внебюджетные фонды определяется исходя из следующей формулы:

$$З_{внеб} = k_{внеб} \cdot (З_{осн} + З_{доп}), \quad (5.11)$$

где $k_{внеб}$ – коэффициент отчислений на уплату во внебюджетные фонды (пенсионный фонд, фонд обязательного медицинского страхования и пр.).

В соответствии с Федеральным законом от 24.07.2009 №212-ФЗ установлен размер страховых взносов равный 30%. На основании пункта 1 ст.58 закона №212-ФЗ для учреждений осуществляющих образовательную и научную деятельность водится пониженная ставка – 27,1%.

Отчисления во внебюджетные фонды представлены в таблице 5.9.

Таблица 5.9 – Отчисления во внебюджетные фонды

Исполнители	Основная ЗП, руб.	Дополнительная ЗП, руб.
Руководитель	15122,16	1814,66
Бакалаавр	15072	1808,64
Коэффициент отчислений во внебюджетные фонды	0,271	-
ИТОГО:		4098,1

5.5.5. Расчет затрат на научные и производственные командировки

Затраты на научные и производственные командировки исполнителей определяются в соответствии с планом выполнения темы и с учетом действующих норм командировочных расходов различного вида и транспортных тарифов. В данном дипломном проекте таких затрат нет.

5.5.6. Контрагентные расходы

Контрагентные расходы включают в себя затраты, связанные с выполнением каких-либо работ по теме сторонними организациями. Расчет величины этой группы расходов зависит от планируемого объема работа и определяется из условий договоров с контрагентами или субподрядчиками. Контрагентные расходы составляют 10% от основной и дополнительной заработной платы. В данном дипломном проекте таких затрат нет.

5.5.7. Накладные расходы

Накладные расходы учитывают прочие затраты организации, не попавшие в предыдущие статьи расходов: печать и ксерокопирование материалов исследования, оплата услуг связи, электроэнергии, почтовые и телеграфные расходы, размножение материалов и т.д. Их величина определяется по следующей формуле:

$$З_{накл} = k_{нр} \cdot (З_{осн} + З_{доп} + З_{внеб} + З_{мат}), \qquad (5.12)$$

где knp – коэффициент, учитывающий накладные расходы.

$$З_{накл} = 40965{,}56 \cdot 0{,}16 = 6554{,}49 \text{ руб.}$$

5.5.8. Формирование бюджета затрат НТИ

Рассчитанная величина затрат научно-исследовательской работы (темы) является основой для формирования бюджета затрат проекта, который при формировании договора с заказчиком защищается научной организацией в качестве нижнего предела затрат на разработку научно-технической продукции. Определение бюджета затрат на научно-исследовательский проект приведен в таблице 5.10.

Таблица 5.10 – Расчет бюджета затрат НТИ

Наименование статьи	Сумма, руб.
1. Материальные затраты НТИ	3050
2. Затраты по основной заработной плате исполнителей темы	30194,16
3. Затраты по дополнительной заработной плате исполнителей темы	3623,3
4. Отчисления во внебюджетные фонды	4098,1
5. Расчет затрат на научные и производственные командировки	0
6. Контрагентные расходы	0
7. Накладные расходы	6554,49
8. Бюджет затрат НТИ	47520,05

5.6 Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования

Определение эффективности происходит на основе расчета интегрального показателя эффективности научного исследования. Его нахождение связано с определением двух средневзвешенных величин: финансовой эффективности и ресурсоэффективности.

Интегральный показатель финансовой эффективности научного исследования получают в ходе оценки бюджета затрат трех (или более) вариантов исполнения научного исследования. Для этого наибольший интегральный показатель реализации технической задачи принимается за базу расчета (как знаменатель), с которым соотносится финансовые значения по всем вариантам исполнения.

Интегральный финансовый показатель разработки определяется как:

$$I_{\text{финр}}^{\text{исп.}i} = \frac{\Phi_{pi}}{\Phi_{max}}, \qquad (5.13)$$

где $I^{исп.i}_{финр}$ – интегральный финансовый показатель разработки; $Ф_{pi}$ – стоимость i-го варианта исполнения; $Ф_{max}$ – максимальная стоимость исполнения научно-

исследовательского проекта (в т.ч. аналоги). За максимально возможную стоимость исполнения примем 100000 руб.

Полученная величина интегрального финансового показателя разработки отражает соответствующее численное увеличение бюджета затрат разработки в разах (значение больше единицы), либо соответствующее численное удешевление стоимости разработки в разах (значение меньше единицы, но больше нуля).

В нашем случае вариант исполнения научного исследования один. Поэтому интегральный финансовый показатель равен 1.

Интегральный показатель ресурсоэффективности вариантов исполнения объекта исследования можно определить следующим образом:

$$I_{pi} = \sum a_i b_i, \qquad (5.14)$$

где I_{pi} – интегральный показатель ресурсоэффективности для i-го варианта исполнения разработки; a_i – весовой коэффициент i-го варианта исполнения разработки; b^a_i, b^p_i – бальная оценка i-го варианта исполнения разработки, устанавливается экспертным путем по выбранной шкале оценивания; n – число параметров сравнения.

Расчет интегрального показателя ресурсоэффективности представлен в таблице 5.11.

Таблица 5.11 – Расчет интегрального показателя ресурсоэффективности

Критерии	Весовой коэффициент параметра	Оценка	Оценка макс.
Адекватность (статическая значимость)	0,2	5	5
Возможность применения любым предприятием	0,15	3	5
Требует наличия исторических данных	0,25	5	5

Простота применения	0,15	4	5
Конкурентоспособность (с другими моделями)	0,25	4	5
ИТОГО	1	4,3	5

$$I_{p-\text{исп}1} = 5 \cdot 0{,}2 + 3 \cdot 0{,}15 + 5 \cdot 0{,}25 + 4 \cdot 0{,}15 + 4 \cdot 0{,}25 = 4{,}3;$$

$$I_{p-\text{исп}max} = 5 \cdot 0{,}2 + 5 \cdot 0{,}15 + 5 \cdot 0{,}25 + 5 \cdot 0{,}15 + 5 \cdot 0{,}25 = 5;$$

Интегральный показатель эффективности вариантов исполнения разработки ($I_{испi}$) определяется на основании интегрального показателя ресурсоэффективности и интегрального финансового показателя по формуле:

$$I_{\text{исп}i} = \frac{I_{p-\text{исп}i}}{I_{\text{финр}}^{\text{исп}i}}, \quad (5.15)$$

Сравнение интегрального показателя эффективности вариантов исполнения разработки позволит определить сравнительную эффективность проекта и выбрать наиболее целесообразный вариант из предложенных. Так как исследование выполнено в одном варианте исполнения, рассчитаем интегральный показатель эффективности относительно максимально возможного варианта. Сравнительная эффективность разработки представлена в табл. 5.12.

Таблица 5.12 – Сравнительная эффективность разработки

Показатели	Исп. 1	Исп. max
Интегральный финансовый показатель разработки	0,44	1
Интегральный показатель ресурсоэффективности разработки	4,3	5
Интегральный показатель эффективности	9,77	5
Сравнительный показатель эффективности	1,954	

Сравнение значений интегральных показателей эффективности позволяет понять и выбрать более эффективный вариант решения поставленной

в бакалаврской работе технической задачи с позиции финансовой и ресурсной эффективности.

5.7 Выводы

В процессе выполнения части работы по финансовому менеджменту, ресурсоэффективности и ресурсосбережению был проведен анализ разрабатываемого исследования.

Во-первых, оценен коммерческий потенциал и перспективность проведения исследования. Полученные результаты говорят о потенциале и перспективности на уровне выше среднего.

Во-вторых, проведено планирование НИР, а именно: определена структура и календарный план работы, трудоемкость и бюджет НТИ. Результаты соответствуют требованиям к ВКР по срокам и иным параметрам.

В-третьих, определена эффективность исследования в разрезах ресурсной, финансовой, бюджетной, социальной и экономической эффективности.

В ходе проделанной работы было произведенно определение структуры работы в рамках научного исследования, определение участников каждой работы установлены продолжительности работ, построен график проведения научных исследований. Произведен расчет материальных затрат, основной и дополнительной заработной платы, внебюджетных отчислений. Полученные результаты и общая эффективность позволяют сделать вывод, что с точки зрения использования ресурсов, затрат бюджета и экономики, наше исполнение является оптимальным. Исполнение подразумевает выполнение работы с руководителем, с использованием ручки и бумаги средней стоимости, а также выполнение расчетов на ноутбуке DEXP.

6. ЗАКЛЮЧЕНИЕ

В результате работы был проведен сбор и предварительная обработка корпуса документов: 32 годовых финансовых отчета мировых компаний за 2013-2016 (приложение А). Автоматизированная обработка документов включала извлечение текстовой информации из pdf файлов: из каждого файла извлечен текст, текст таблиц и рисунков. Для извлечения текстовой информации из инфографики использованы методы компьютерного зрения. Получена выборка из 7 390 абзацев, элементы выборки подготовлены для машинного обучения: удалены переносы строк, знаки пунктуации, числа (слова, содержащие только цифры), стоп-слова, слова обрезаны до их основы.

На основании сформулированных критериев релевантности текста: 1) явное указание, 2) контактная информация, 3) логический вывод из текста 4) логический вывод из числовой информации, осуществлена бинарная классификация элементов выборки: 1 194 абзаца (16%) релевантные, метка – «1».

С использованием машинных методов обучения: 1) мера важности слова (TF-IDF), 2) векторное представление слова (word2vec), 3) векторное представление документа (doc2vec) для элементов выборки сгенерированы признаки: 25 633 признака, из них 25 328 признаков TF-IDF, 300 признаков word2vec и 5 признаков doc2vec. На основании описательной статистики выборки полученное множество признаков дополнено тремя количественными признаками (количество символов (с пробелами и без пробелов), слов в абзаце).

В результате работы для сбалансированной и несбалансированной выборок оценены параметры для следующих классификаторов: наивный байесовский классификатор, логистическая регрессия и градиентный бустинг над решающими деревьями. Последняя модель показала наибольшую точность оценки релевантности текста (88,70%) и была выбрана в качестве оптимальной модели. Для данной модели приведен иллюстративный графический пример.

Выражаю благодарность компании ООО «Эко-Томск» в лице Михаила Ожгибесова и Андрея Орлова за задачу, помощь и сотрудничество.

7. СПИСОК ПУБЛИКАЦИЙ СТУДЕНТА

1. Semenov M. E. , Bulygin L. E. , Koroleva E. A. , Tursunov D. A. A project teams creation based on communities detection // CEUR Workshop Proceedings. - 2016 - Vol. 1710. - p. 303-314.
2. Semenov M. E. , Bulygin L. E. Investigating the Plastic Behavior in Face-Centered Cubic Metals with Strain Rate Jumps // Key Engineering Materials . - 2016 - Vol. 683. - p. 100-105.

8. СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. Titanic: Machine Learning from Disaster // [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/c/titanic - дата доступа: 22.05.2017.
2. 2016 Annual Report on Form 10-K Coca-Cola Company// [Электронный ресурс]. – Режим доступа: http://www.coca-colacompany.com/content/dam/journey/us/en/private/fileassets/pdf/investors/2016-AR-10-K.pdf - дата доступа: 22.05.2017.
3. 2016 Annual Report Bank of England // [Электронный ресурс]. – Режим доступа: http://www.bankofengland.co.uk/publications/Documents/annualreport/2016/boereport.pdf - дата доступа: 22.05.2017.
4. 2014 Annual Report Deutsche Bank // [Электронный ресурс]. – Режим доступа: https://www.db.com/ir/en/download/Deutsche_Bank_Annual_Report_2014_entire.pdf - дата доступа: 22.05.2017.
5. R. Balakrishnan, X.Y. Qiu, P. Srinivasan. On the predictive ability of narrative disclosures in annual reports // [Электронный ресурс]. – Режим доступа: http://www.sciencedirect.com/science/article/pii/S0377221709004822 - дата доступа: 22.05.2017.
6. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. Distributed Representations of Words and Phrases and their Compositionality // [Электронный ресурс]. – Режим доступа: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf - дата доступа: 22.05.2017.
7. Q. Le, T. Mikolov. Distributed Representations of Sentences and Documents // [Электронный ресурс]. – Режим доступа: https://cs.stanford.edu/~quoc le/paragraph_vector.pdf - дата доступа: 22.05.2017.
8. Large Movie Review Dataset // [Электронный ресурс]. – Режим доступа: http://ai.stanford.edu/~amaas/data/sentiment/ - дата доступа: 22.05.2017.
9. F. Sebastiani. Machine Learning in Automated Text Categorization // [Электронный ресурс]. – Режим доступа: http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf - дата доступа: 22.05.2017.
10. К. Воронцов. Линейные методы классификации и регрессии: метод стохастического градиента // [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/images/5/53/Voron-ML-Lin-SG.pdf - дата доступа: 22.05.2017.

11. К. Воронцов. Байесовская теория классификации и методы восстановления плотности // [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/images/c/c1/Voron-ML-Bayes1-slides.pdf - дата доступа: 22.05.2017.
12. К. Воронцов. Разделение смеси распределений. EM-алгоритм для классификации и кластеризации // [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/images/c/c6/Voron-ML-Bayes2-slides.pdf - дата доступа: 22.05.2017.
13. К. Воронцов. Композиции классификаторов // [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/images/c/cd/Voron-ML-Compositions-slides.pdf - дата доступа: 22.05.2017.
14. K. S. Jones. A statistical interpretation of term specificity and its application in retrieval // [Электронный ресурс]. – Режим доступа: https://ai2-s2-pdfs.s3.amazonaws.com/4f09/e6ec1b7d4390d23881852fd7240994abeb58.pdf - дата доступа: 22.05.2017.
15. SciKit learn TfidfVectorizer // [Электронный ресурс]. – Режим доступа: http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html - дата доступа: 22.05.2017.
16. Pre-trained Google News word2vec model // [Электронный ресурс]. – Режим доступа: https://github.com/mmihaltz/word2vec-GoogleNews-vectors - дата доступа: 22.05.2017.
17. Deep learning with paragraph2vec // [Электронный ресурс]. – Режим доступа: https://radimrehurek.com/gensim/models/doc2vec.html - дата доступа: 22.05.2017.
18. Give me some credit // [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/c/GiveMeSomeCredit - дата доступа: 22.05.2017.

ПРИЛОЖЕНИЕ А. Используемые отчеты в выборке

N	Компания	Год
1.	Bank of England	2014
2.	Deutsche Bank	2013
3.	Deutsche Bank	2014
4.	European Central Bank	2015
5.	Aeroflot	2015
6.	Coca Cola	2015
7.	Daimler	2015
8.	FastRetailing	2015
9.	Ford	2015
10.	Gazprombank	2015
11.	Google	2008
12.	PepsiCo	2015
13.	PhillipMorris	2015
14.	Starbucks	2015
15.	M&T Bank	2011
16.	WorldBankGroup	2014
17.	Bank of Russia	2004
18.	Apple	2015
19.	Atrium	2015
20.	Auchan	2015
21.	BritishAmercanTobacco	2015
22.	GeneralElectric	2015
23.	LG	2015
24.	LOreal	2015
25.	MasterCard	2015
26.	Metrogroup	2014-2015
27.	NBC	2015
28.	Nestle	2015
29.	Gillete	2016
30.	P&G	2016
31.	SPT	2015
32.	Visa	2015

ПРИЛОЖЕНИЕ Б. Листинг программы для генерации признаков

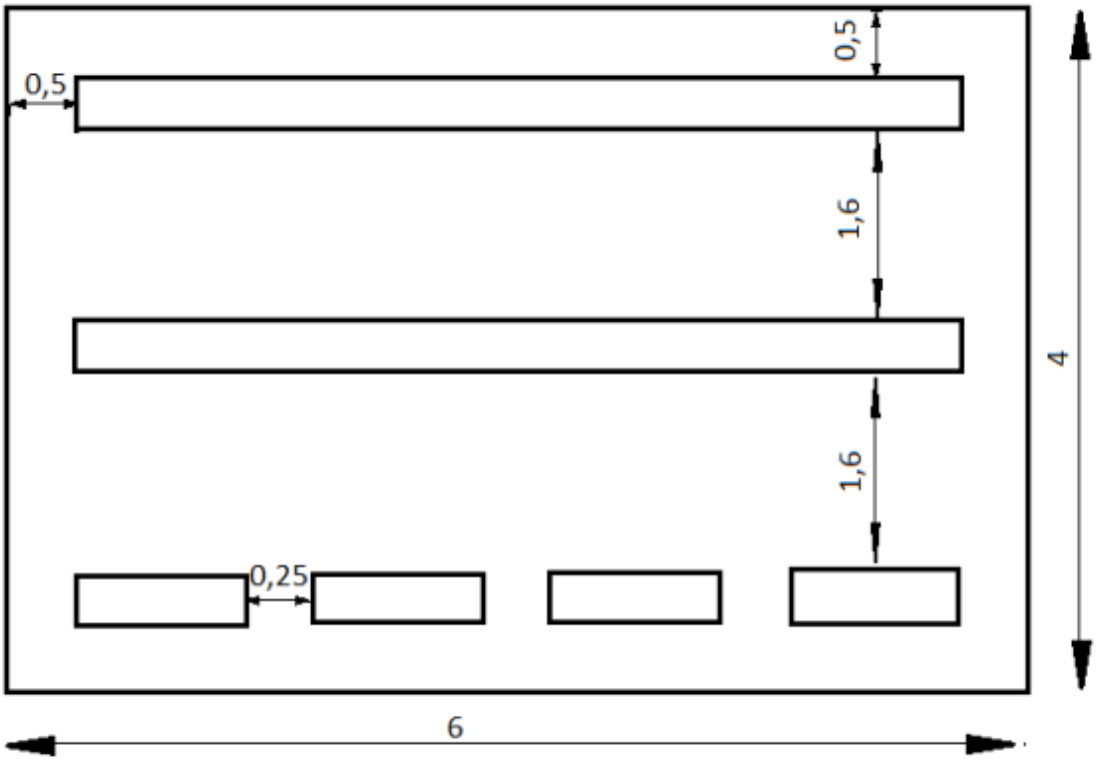
```
import pandas as pd
import string
from stemming.porter2 import stem
from nltk.corpus import stopwords
import re
%matplotlib inline
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
cachedStopWords = stopwords.words('english')
from nltk import word_tokenize
import gensim
from gensim.models import Doc2Vec
from gensim.models.doc2vec import TaggedDocument
from tqdm import tqdm
df = pd.read_csv('sample_without_dubl.csv', sep=';')
def ClearText(text):

    text = str(text)

    # Remove line breaks
    text = text.replace("\n", "")
    text = text.replace("\r", "")

    # Remove punctuation and numbers
    text = text.translate(str.maketrans('','',string.punctuation))
    text = text.translate(str.maketrans('','','1234567890'))

    # Remove multiple spaces
    text = re.sub(' +',' ',text)

    # Remove spaces at the beginning and end of the line
    text = text.rstrip().lstrip()

    # Split by space
    text = text.split(' ')

    # Remove stopwords
    text = [ x for x in text if x not in cachedStopWords ]

    # Stem each word
    text = [ stem(x).lower() for x in text ]

    # Return CLEANEST TEXT IN THE WORLD
    return ' '.join(text)

def sent2vec(s):
    words = str(s)
    words = word_tokenize(words)
    words = [w for w in words if w.isalpha()]
    M = []
    for w in words:
        try:
            M.append(model[w])
        except:
            continue
```

```
    M = np.array(M)
    v = M.sum(axis=0)
    return v / np.sqrt((v ** 2).sum())
# Label distribution
df.target.value_counts().plot(kind="bar", rot=0)
# Characters
len_char = df.text.apply(lambda x: len(''.join((str(x)))))

print(np.mean(len_char))
print(np.max(len_char))
print(np.min(len_char))
# Words
len_words = df.text.apply(lambda x: len(str(x).split()))
print(np.mean(len_words))
print(np.max(len_words))
print(np.min(len_words))
data = df.text.apply(lambda x: ClearText(x))
features = pd.DataFrame()
# n_features
features['len_words'] = data.apply(lambda x: len(str(x).split()))
features['len_char'] = data.apply(lambda x: len(''.join((str(x)))))
features['len_char_ws'] = data.apply(lambda x:
len(''.join((str(x).replace(' ','')))))
# TFIDF features
vectorizer = TfidfVectorizer(min_df=1)

tfidf = vectorizer.fit_transform(data)
# Word2Vec features
model = gensim.models.KeyedVectors.load_word2vec_format('GoogleNews-
vectors-negative300.bin.gz', binary=True)

vectors = np.zeros((data.shape[0], 300))

for i, q in tqdm(enumerate(data.values)):
    vectors[i, :] = sent2vec(q)

docs = pd.DataFrame()
docs['text'] = data
docs['id'] = df.index.values

docs_tagged = docs.apply(lambda r:
TaggedDocument(words=word_tokenize(r['text']), tags=[r['id']]), axis=1)
doc2vec_model = Doc2Vec(docs_tagged.values, workers=4, size=5, iter=20,
dm=1)

len(doc2vec_model.docvecs)
tfidf_df = pd.DataFrame(tfidf.toarray())
headers = ['w'+str(i) for i in range(0,300)]
w2v = pd.DataFrame(vectors, columns=headers)
headers = ['d'+str(i) for i in range(0,5)]
d2v = pd.DataFrame(list(doc2vec_model.docvecs), columns=headers)
all_data = pd.concat([features, tfidf_df, w2v, d2v], axis=1,
join='inner')
all_data.to_csv('features.csv', index=False)
```

ПРИЛОЖЕНИЕ В. Листинг программы для обучения моделей

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
import xgboost as xgb
from sklearn.cross_validation import train_test_split
from sklearn.metrics import accuracy_score
# Unbalanced sample
X = pd.read_csv('features.csv')
data = pd.read_csv('sample_without_dubl.csv', sep=';')
y = data['target']
del data
# Balanced sample
X = pd.read_csv('features.csv')
data = pd.read_csv('sample_without_dubl.csv', sep=';')
X['target'] = data['target']
del data
X_pos = X[X['target'] == 1]
X_neg = X[X['target'] == 0].sample(n = len(X_pos), random_state=42)
X = X_pos.append(X_neg)
y = X['target']
del X['target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state = 42)
LR = LogisticRegression()
X_train_impute = X_train.fillna(0).replace(np.nan,0).replace(np.inf,0)
X_test_impute = X_test.fillna(0).replace(np.nan,0).replace(np.inf,0)
LR.fit(X_train_impute,y_train)
y_pred = LR.predict(X_test_impute)
print(accuracy_score(y_test,y_pred))
GNB = GaussianNB()
X_train_impute = X_train.fillna(0).replace(np.nan,0).replace(np.inf,0)
X_test_impute = X_test.fillna(0).replace(np.nan,0).replace(np.inf,0)
GNB.fit(X_train_impute,y_train)
y_pred = GNB.predict(X_test_impute)
print(accuracy_score(y_test,y_pred))
dtrain = xgb.DMatrix(data=X_train, label=y_train)
dtest = xgb.DMatrix(data=X_test, label=y_test)
param = {'max_depth':2, 'eta':0.2}
param['nthread'] = 4
param['eval_metric'] = 'error'
evallist  = [(dtrain,'train'),(dtest,'test')]
plst = param.items()
num_round = 100
bst1 = xgb.train( plst, dtrain, num_round, evallist )
bst1.save_model('xgboost_balance.model')
bst = xgb.Booster({'nthread':4}) #init model
bst.load_model('xgboost_balance.model') # load data
num_round = 100
bst1 = xgb.train( plst, dtrain, num_round, evallist, xgb_model=bst )
bst1.save_model('xgboost_balance.model')
y_prob = bst1.predict(dtest)
y_pred = []

for x in y_prob:
    if x>=0.5:
        s = 1
    else:
        s = 0
    y_pred.append(s)
accuracy_score(y_test, y_pred)
```

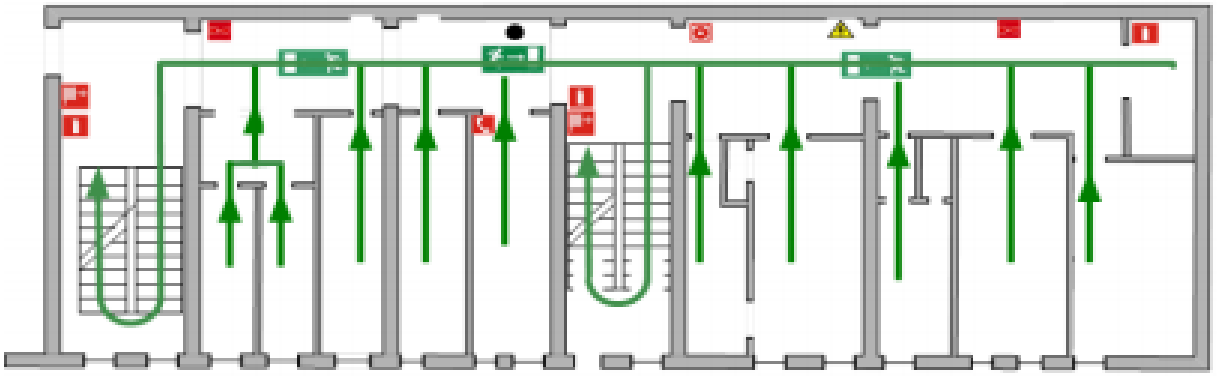
ПРИЛОЖЕНИЕ Г. План помещения и размещения светильников с люминесцентными лампами.



ПРИЛОЖЕНИЕ Д. План эвакуации в случае пожара.

План эвакуации в случае пожара

ПЛАН ЭВАКУАЦИИ
2-го этажа



Направление движения
Кнопка включения АПС
Электрощит
Телефон
Вы находитесь здесь
Пожарный кран
Огнетушитель
Извещатель

Действия при пожаре Сохранять спокойствие			
1	Сообщить по телефону		• Адрес объекта • Место возникновения пожара • Свою фамилию
2	Эвакуировать людей		• Ориентироваться по знакам направления движения • Взять с собой пострадавших
3	По возможности принять меры по тушению пожара		• Использовать средства противопожарной защиты • При необходимости обеспечить помещение

Ответственный за эвакуацию и включение системы оповещения

_____