TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа информационных технологий и робототехники
Направление подготовки 09.03.02 «Информационные системы м технологии»
Отделение школы (НОЦ) информационных технологий

БАКАЛАВРСКАЯ РАБОТА

Тема работы
Применение нейросетевых алгоритмов для определения авторства текста на основе различных метрик и методов лингвистического анализа

УДК 004.032.26:82.05:808.1

Студент

Группа	ФИО	Подпись	Дата
8И5А	Демиденко Людмила Руслановна		

Руководитель

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ	Цапко Сергей Геннадьевич	к.т.н.		

КОНСУЛЬТАНТЫ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
ассистент ОСГН ШБИП	Шулинина Юлия Игоревна			

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
ассистент ООД	Немцова Ольга Александровна			

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ	Цапко Ирина Валериевна	к.т.н		

Томск – 2019 г.

РЕЗУЛЬТАТЫ ОБУЧЕНИЯ (КОМПЕТЕНЦИИ ВЫПУСКНИКОВ)

по направлению 09.03.02 «Информационные системы и технологии»

Код результатов	Результат обучения (выпускник должен быть готов)
Профессиональные компетенции	
Р1	Применять базовые и специальные естественнонаучные и математические знания для комплексной инженерной деятельности по созданию, внедрению и эксплуатации геоинформационных систем и технологий, а также информационных систем и технологий в бизнесе
Р2	Применять базовые и специальные знания в области современных информационных технологий для решения инженерных задач.
Р3	Ставить и решать задачи комплексного анализа, связанные с созданием геоинформационных систем и технологий, информационных систем в бизнесе, с использованием базовых и специальных знаний, современных аналитических методов и моделей.
Р4	Выполнять комплексные инженерные проекты по созданию информационных систем и технологий, а также средств их реализации (информационных, методических, математических, алгоритмических, технических и программных).
Р5	Проводить теоретические и экспериментальные исследования, включающие поиск и изучение необходимой научно-технической информации, математическое моделирование, проведение эксперимента, анализ и интерпретация полученных данных, в области создания геоинформационных систем и технологий, а также информационных систем и технологий в бизнесе.
Р6	Внедрять, эксплуатировать и обслуживать современные геоинформационные системы и технологии, информационные системы и технологии в бизнесе, обеспечивать их высокую эффективность, соблюдать правила охраны здоровья, безопасность труда, выполнять требования по защите окружающей среды.
Универсальные компетенции	
Р7	Использовать базовые и специальные знания в области проектного менеджмента для ведения комплексной инженерной деятельности.
Р8	Осуществлять коммуникации в профессиональной среде и в обществе в целом. Владеть иностранным языком (углублённый английский язык), позволяющем работать в иноязычной среде, разрабатывать документацию, презентовать и защищать результаты комплексной инженерной деятельности.
Р9	Эффективно работать индивидуально и в качестве члена команды, состоящей из специалистов различных направлений и квалификаций,
Р10	Демонстрировать личную ответственность за результаты работы и готовность следовать профессиональной этике и нормам ведения комплексной инженерной деятельности.
Р11	Демонстрировать знания правовых, социальных, экологических и культурных аспектов комплексной инженерной деятельности, а также готовность к достижению должного уровня физической подготовленности для обеспечения полноценной социальной и профессиональной деятельности.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа информационных технологий и робототехники
Направление подготовки 09.03.02 «Информационные системы м технологии»
Отделение школы (НОЦ) информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП

_____ _____ _____
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Бакалаврской работы

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

Студенту:

Группа	**ФИО**
8И5А	Демиденко Людмиле Руслановне

Тема работы:

Применение нейросетевых алгоритмов для определения авторства текста на основе различных метрик и методов лингвистического анализа	
Утверждена приказом директора (дата, номер)	№3654/с от 13.05.2019 г.

Срок сдачи студентом выполненной работы:	04.06.2019

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	Нейронная сеть, позволяющая идентифицировать автора по отрывку из его текста. Нейронная сеть должна определять автора с точностью более 90% при минимальном количестве входных символов.

Перечень подлежащих исследованию, проектированию и разработке вопросов	Изучение теории по теме исследования; Подбор корпуса текстов; Обработка корпуса текстов; Проектирование и реализация нейросети; Тестирование нейросети на двух различных метриках; Анализ результатов работы, сравнение метрик; Расчет ресурсоэффективности и ресурсосбережения и анализ вредных производственных факторов.
Перечень графического материала	Презентация в формате *.pptx на 17 слайдах

Консультанты по разделам выпускной квалификационной работы

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	ассистент ОСГН ШБИП Шулинина Юлия Игоревна
Социальная ответственность	ассистент ООД Немцова Ольга Александровна

Названия разделов, которые должны быть написаны на русском и иностранном языках:

Введение
Обзор предметной области
Анализ собранных данных
Реализация системы
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение
Социальная ответственность
Заключение
Заключение (английский)

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ	Цапко Сергей Геннадьевич	к.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И5А	Демиденко Людмила Руслановна		

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа информационных технологий и робототехники
Направление подготовки 09.03.02 «Информационные системы м технологии»
Отделение школы (НОЦ) информационных технологий
Период выполнения (осенний / весенний семестр 2018/2019 учебного года)

Форма представления работы:

Бакалаврская работа

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

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	4.06.2019

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
22.02.19	Подбор и изучение литературы	
01.03.19	Анализ предметной области	
28.03.19	Подбор корпуса текстов	
04.04.19	Реализация программного кода для обработки корпуса текстов	
07.04.19	Обработка корпуса текстов	
11.04.19	Проектирование нейронной сети	
18.04.19	Разработка нейронной сети	
09.05.19	Тестирование нейронной сети	
29.05.19	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	
29.05.19	Социальная ответственность	

Составил преподаватель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ	Цапко Сергей Геннадьевич	к.т.н.		

СОГЛАСОВАНО:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ	Цапко Ирина Валериевна	к.т.н		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8И5А	Демиденко Людмиле Руслановне

Школа	ИШИТР	**Отделение школы (НОЦ)**	ОИТ
Уровень образования	бакалавриат	**Направление/специальность**	09.03.02 Информационные системы и технологии

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих	Оклад инженера – 21760 руб. Оклад руководителя – 33664 руб.
2. Нормы и нормативы расходования ресурсов	Премиальный коэффициент 30%; Коэффициент доплат и надбавок 20%; Районный коэффициент 30%; Коэффициент дополнительной заработной платы 12%; Накладные расходы 16%.
3. Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования	Коэффициент отчислений на уплату во внебюджетные фонды 30%
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Оценка коммерческого потенциала, перспективности и альтернатив проведения НИ с позиции ресурсоэффективности и ресурсосбережения	Анализ конкурентных технических решений
2. Планирование и формирование бюджета научных исследований	Формирование плана и графика разработки: - определение структуры работ; - определение трудоемкости работ; - разработка графика Ганта. Формирование бюджета затрат на научное исследование: - материальные затраты; -затраты на специальное оборудование; - заработная плата (основная и дополнительная); - отчисления на социальные цели; - накладные расходы.
3. Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования	- Определение потенциального эффекта исследования
Перечень графического материала *(с точным указанием обязательных чертежей)*:	
1. Оценочная карта конкурентных технических решений *2. Матрица SWOT* *3. График Гантта* *4. Расчет бюджета затрат*	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент ОСГН ШБИП	Шулинина Ю.И.			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И5А	Демиденко Людмила Руслановна		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8И5А	Демиденко Людмиле Руслановне

Школа	ИШИТР	Отделение (НОЦ)	ОИТ
Уровень образования	бакалавриат	Направление/специальность	09.03.02 Информационные системы и технологии

Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Система, позволяющая пользователю определить автора по его тексту.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	Требования к организации и оборудованию рабочих мест с ПЭВМ.
2. Производственная безопасность: 2.1. Анализ выявленных вредных и опасных факторов 2.2. Обоснование мероприятий по снижению воздействия	1.Отклонение показателей микроклимата 2. Превышение уровня шума 3.Отсутствие или недостаток естественного света 4.Недостаточная освещенность рабочей зоны 5.Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека.
3. Экологическая безопасность:	Влияние объекта исследования на окружающую среду; мероприятия по защите окружающей среды.
4. Безопасность в чрезвычайных ситуациях:	Основные и типичные чрезвычайные ситуации в офисном помещении; установка общих правил поведения и рекомендаций во время ЧС.

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент ООД	Немцова О.А.			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И5А	Демиденко Людмила Руслановна		

РЕФЕРАТ

Пояснительная записка содержит 79 страниц, 13 рисунков, 19 таблиц, 15 источников и 2 приложения.

Данная работа посвящена разработке нейросетевого алгоритма, позволяющего определить автора по его тексту. Был собран корпус текстов русской литературы начиная от 19 века, заканчивая нашим временем. Всего корпус состоит более чем из 10 млн слов. Также были разработаны 2 нейросети, использующие 2 разные метрики для представления текста в векторном виде. Проведено сравнение данных метрик.

Ключевые слова: нейросеть, метрика.

Объектом исследования является разработка нейронной сети.

Цель работы – разработка нейронной сети, идентифицирующей автора по его тексту.

Сеть была реализована на языке python с использованием библиотек tensorflow, sklearn, gensim, pymorphy2.

Данная сеть может быть интегрирована в веб-ресурс или десктопное приложение.

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ, СОКРАЩЕНИЯ И НОРМАТИВНЫЕ ССЫЛКИ

Нейросеть – система соединенных и взаимодействующих между собой простых процессоров (искусственных нейронов).

Нейрон – структурная единица искусственной нейронной сети и представляет собой аналог биологического нейрона.

ML – машинное обучение.

Фреймворк – программные продукты, которые упрощают создание и поддержку технически сложных или нагруженных проектов.

TF – фреймворк TensorFlow.

ОБЪЕКТ И МЕТОД ИССЛЕДОВАНИЯ

Объектом исследования в данной работе является печатный текст. Предметом исследования – характеристики данных текстов и методы определения авторства текста.

Методами данного исследования стали методы математической статистики, вычислительного эксперимента и искусственного интеллекта.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

То, как человеческий мозг обрабатывает информацию, коренным образом отличается от методов цифровой обработки. Работа человеческого мозга сравнима со сложным, нелинейным устройством с параллельными вычислениями. Учеными доказано, что структуру мозга составляют огромное количество нервных клеток или нейронов. Каждый нейрон способен создавать от 10 до 100 тысяч структурных связей. Эти нейроны и связи между ними образуют нейронную сеть, которая передает нервные сигналы, от количества которых зависит активность мозга.

Построенная таким образом нейронная сеть способна выполнять задачи разной сложности за минимальные промежутки времени. Например, на зрительное распознавание объекта у мозга уходит от 100 до 200 мс. В то же время даже у самого быстродействующего и мощного компьютера на эту задачу уйдет гораздо больше времени.

Нейронные сети начали создавать еще в начале XX века. Однако начали распространяться они только в конце того же века. И несмотря на то, что до сих пор нейросети проигрывают мозгу человека, они уже применяются для решения многих задач. Их главное отличие от привычных нам вычислительных устройств заключается в том, что нейросети не программируются для некой конкретной задачи, а обучаются выполнять эту задачу.

Нейросети могут выполнять различные задачи. Целый блок связан с обработкой текстовой информации. Например, существуют задачи определения автора текста, его пола, профессии, национальности, уровня образования автора и т.д.

Идентификация авторства текста приобретает свою актуальность в связи с переходом от рукописного письма к печатному набору. В спорной ситуации, например, при криминалистическом исследовании печатного текста, методы определения авторства по почерку теряют свою силу. К тому же идентификация автора по почерку может выявить только исполнителя, а

не автора текста. Идентификация авторства текста охватывает большой спектр целей: от отыскания автора необходимой статьи в интернете или запоминающегося отрывка художественного произведения до достаточно серьёзных военных и криминалистических целей.

1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

1.1. Идентификация автора текста

Данной проблемой отечественные и зарубежные ученые занимаются уже около 120 лет. За это время было предложено множество различных методов идентификации, начиная от подсчета слов в текстах, заканчивая применением нейронных сетей и искусственного интеллекта.

В основном у экспертов, занимающихся данной проблемой, пользуются популярностью методы, в основе которых стоит гипотеза о том, что каждый автор характеризуется набором уникальных стилистических приемов и характерными языковыми особенностями (лексическими, грамматическими, фразеологическими), которые прослеживаются в большинстве или всех его произведениях. Также ученые ориентируются на автобиографическую информацию и «любимые» слова автора.

Особенностью данных методов является то, что у автора может не быть таких уникальных особенностей. Либо, если эти особенности ярко выражены, существует вероятность подмены авторского стиля.

Также существует ряд методов анализа текста, основанный на неконтролируемых человеком характеристиках, которые являются общими для всех авторов.

Вопросами определения авторства текста в России занимались такие ученые, как Морозов Н.А., Шевелев О.Г., Марков А.А., Рогов А.А., Хмелев Д.В., Хетсо Г.

Наиболее цитируемыми зарубежными авторами являются Farringdon J.M., Morton A.Q., Efron B. Mendenhall T.C..

Единого мнения о том, какой набор характеристик будет давать лучший результат, пока нет. Тем более, что большая часть литературы написана на основе англоязычных авторов и текстах, в то время как про русские тексты литературы сравнительно мало. Так как каждый язык имеет

свои особенности, то логично предположить, что и методы идентификации автора текста должны быть разными для разных языков.

1.2. Метрика TF-IDF

Для работы с текстом в компьютере необходимо представить текст в числовом виде. Для этого существуют различные метрики, позволяющие представить текст в виде вектора числовых значений, такие как one-hot encoding, мешок слов, Global Vectors и др. В данной работе были выбраны метрики TF-IDF и word2vec, так как их часто рекомендуют в литературе.

Метрика TF-IDF – это простой и удобный способ оценить важность термина для какого-либо документа относительно всех остальных документов [1]. В ее основе лежит следующий принцип: слово имеет большую значимость для документа в том случае, если оно часто встречается в данном документе и редко встречается во всех остальных.

Достоинства метрики:

1. Простота расчета,
2. Слова, встречающиеся во многих документах, такие как предлоги, союзы, местоимения, будут иметь низкий вес TF-IDF, а слова, уникальные для данного текста, например, неологизмы, термины, признаки времени, – высокий.

Данная метрика состоит из двух частей:

TF (term frequency) – частотность термина, которая измеряет, насколько часто термин встречается в документе. Так как в длинных текстах вероятность встретить несколько раз один и тот же термин выше, чем в коротких, то здесь применяют относительные величины – делят количество раз, когда нужный термин встретился в тексте, на общее количество слов в тексте.

$$tf(t,d) = \frac{n_t}{\sum n_k} \quad (1)$$

где n_t – количество вхождений слова t в документ d,

$\sum n_k$ – общее число слов в документе.

IDF (inverse document frequency) – обратная частотность документов. Учет данной метрики уменьшает вес широкоупотребимых слов, таких как предлоги, союзы, местоимения, которые не влияют на смысл текста.

$$idf(t, D) = \frac{|D|}{|\{d_i \in D | \ t \in d_i\}|} \quad (2)$$

где $|D|$ – количество документов в коллекции,

$|\{d_i \in D | \ t \in d_i\}|$ – число документов из коллекции D, в которых встречается слово t, при условии, что $n_t \neq 0$.

И, наконец, найти метрику TF-IDF можно по формуле 3.

$$\text{TfIdf}(t, d, D) = tf(t, d) \times idf(t, D) \quad (3)$$

1.3. Метрика word2vec

Второй метрикой, использованной в работе, была метрика word2vec.

Word2vec – набор алгоритмов, рассчитывающих векторное представление слова. Векторные представления находятся исходя из предположения, что слова, которые находятся в похожих контекстах, скорее всего будут значить похожие вещи (быть семантически близкими). Более формально задача стоит так: максимизация косинусной близости между векторами слов (скалярное произведение векторов), которые появляются рядом друг с другом, и минимизация косинусной близости между векторами слов, которые не появляются друг рядом с другом. Рядом друг с другом в данном случае значит в близких контекстах [2].

Например, слова «ученые» и «исследователи» часто встречаются в похожих контекстах, вроде «Ученые выявили причины старения» или «Исследователи выявили причины старения». Алгоритмы word2vec анализируют данные контексты и приходят к заключению, что слова «ученые» и «исследователи» близки по смыслу, а значит, их векторное представление будет близко. Однако для того, чтобы алгоритмы оценивали семантическую близость слов верно, их необходимо обучать на большом корпусе текстов. Часто такие словари обучают на Википедии.

Основные задачи, решающиеся с помощью word2vec:

- Кластеризация слов по принципу их семантической близости,
- Выявление семантической близости слов,
- Исправление опечаток,
- Транслитерация,
- В поисковых сервисах.

1.4. Нейронная сеть и формальный нейрон

Одна из задач данного исследования – проектирование и реализация нейронной сети, определяющей автора текста.

Нейронная сеть – последовательность связанных нейронов, единиц, получающих и передающих информацию. Сами по себе они не играют важной роли: нейроны имеют значение только в выстроенной из них цепи [3].

Устройство нейросети позаимствовано из биологии. Подражая устройству человеческого мозга, машина обучается выполнять задачи, которые невозможно выполнить обычным программированием.

Самыми распространенными задачами для нейросети являются:

- классификация — распределение данных по параметрам;
- предсказание — возможность предсказывать следующий шаг;
- распознавание — в настоящее время, самое широкое применение нейронных сетей. Используется в Google, когда вы ищете фото или в камерах телефонов, когда оно определяет положение вашего лица и выделяет его и многое другое [4].

Как уже было сказано, нейросеть состоит из нейронов. Рассмотрим их строение. Нейроны состоят из входов (синапсов), сумматора и функции активации (рис. 1).

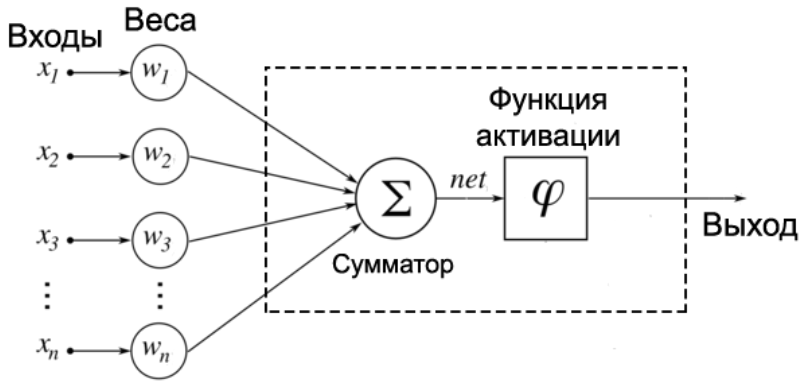

Рисунок 1 – Строение формального нейрона

Поступившие на входы сигналы умножаются на свои веса. Сигнал первого входа x_1 умножается на соответствующий этому входу вес w_1. В итоге получаем $x_1 w_1$. И так до n-ого входа. В итоге на последнем входе получаем $x_n w_n$.

Далее сигналы идут на сумматор. Он агрегирует все входные сигналы во взвешенную сумму – число, которое характеризует поступивший на нейрон сигнал в целом.

$$x_1 w_1 + x_2 w_2 + \cdots + x_n w_n = \sum_{i=1}^{n} x_i w_i \quad (4)$$

где x_i – сигнал i-го входа,

w_i – вес i-го входа.

После сумматора взвешенная сумма попадает в функцию активации. Функция активации — это способ нормализации входных данных. То есть, если на входе будет большое число, пропустив его через функцию активации, можно получить выход в нужном диапазоне. Функций активации достаточно много самые распространенные: линейная, сигмоид и гиперболический тангенс. Главные их отличия — это диапазон значений.

1.5. Язык программирования Python

Для реализации нейросети был выбран язык Python, так как он является одним из самых популярных языков на данный момент. Это интерпретируемый, объектно-ориентированный язык программирования. Он чрезвычайно прост и содержит небольшое число ключевых слов, вместе с тем очень гибок и выразителен. Это язык более высокого уровня, чем Pascal,

C++ и, естественно C, что достигается, в основном, за счет встроенных высокоуровневых структур данных (списки, словари, тьюплы).

Недостатки языка:

- скорость исполнения кода;
- проблемы с совместимостью версий.

Достоинства языка:

- предельно простой синтаксис;
- большое количество справочной литературы;
- множество доступных сред разработки, сервисов и фреймворков [5].

Одно из самых больших преимуществ языка – широкий спектр библиотек. Самыми популярными библиотеками, связанными непосредственно с машинным обучением и с подготовкой данных для обучения являются Scikit-learn (работа с классическими алгоритмами машинного обучения), Tensorflow и Theano (глубокое обучение), Pandas (извлечение и подготовка данных) и Matplotlib (визуализация данных).

1.6. Морфологический анализатор pymorphy2

Для того чтобы грамотно считать метрики текста, нужно, чтобы все слова были использованы в начальной форме. Для этого применялась библиотека pymorphy2.

Pymorphy2 написан на языке Python (работает под 2.7 и 3.3+).

Возможности анализатора:

- приводить слово к нормальной форме (например, «люди» → «человек» или «гулял» → «гулять»),
- ставить слово в нужную форму. Например, ставить слово во множественное число, менять падеж слова и т.д.,
- возвращать грамматическую информацию о слове (число, род, падеж, часть речи и т.д.).

При работе используется словарь OpenCorpora; для незнакомых слов строятся гипотезы. Библиотека достаточно быстрая: в настоящий момент скорость работы - от нескольких тыс слов/сек до > 100тыс слов/сек (в зависимости от выполняемой операции, интерпретатора и установленных пакетов); потребление памяти - 10…20Мб; полностью поддерживается буква ё.

В исходном словаре из OpenCorpora около 400тыс. лексем и 5млн отдельных слов. Он представляет собой файл, в котором слова объединены в лексемы следующим образом:

1

енот NOUN,anim,masc sing,nomn

енота NOUN,anim,masc sing,gent

еноту NOUN,anim,masc sing,datv

енота NOUN,anim,masc sing,accs

енотом NOUN,anim,masc sing,ablt

еноте NOUN,anim,masc sing,loct

енотами NOUN,anim,masc plur,nomn

енотов NOUN,anim,masc plur,gent

енотам NOUN,anim,masc plur,datv

енотов NOUN,anim,masc plur,accs

енотами NOUN,anim,masc plur,ablt

енотах NOUN,anim,masc plur,loct

Сначала указывается номер лексемы, затем перечисляются формы слова и соответствующая им грамматическая информация (тег). Для существительного часть речи (NOUN), одушевленность или неодушевленность (anim), род (masc), число (sing или plur) и падеж.

Две основные операции, которые умеет делать морфологический анализатор - разбор и склонение слов. Если у нас есть словарь с лексемами, и мы хотим разобрать/просклонять словарное слово, то эти операции очень простые:

- разобрать слово - найти его в словаре и вернуть приписанную ему грамматическую информацию;
- просклонять слово - найти слово в словаре, определить его лексему, а затем найти в лексеме нужное слово с запрошенными грамматическими характеристиками.

Если все, что нужно - разбор словарных слов, то можно загрузить все слова и их грамматическую информацию в память “как есть”, либо сохранить в какую-то БД общего назначения. С этим есть 2 проблемы:

- в словаре OpenCorpora для русского языка около 400тыс. лексем и 5млн отдельных слов; если все загрузить в питоний list, то потратим примерно 2Гб оперативной памяти (в dict - еще больше);
- хотелось бы, чтоб операции по анализу и склонению слов осуществлялись быстро - в том числе для слов, отсутствующих в словаре, и слов, записанных каким-то “упрощенным” способом (например, с буквой е вместо ё) [6].

1.7. Tensorflow

Для работы с нейросетью применялся framework TensorFlow и Tf.Keras.

TensorFlow – ML-framework от Google, который предназначен для проектирования, создания и изучения моделей глубокого обучения. TensorFlow можно использовать для того, чтобы производить численные вычисления. Они будут производятся с помощью data-flow графов. В этих графах вершины представляют собой математические операции, в то время как ребра представляют собой данные, которые обычно представляются в виде многомерных массивов или тензоров, которые сообщаются между этими ребрами.

При написании программ в TensorFlow ключевым объектом всех операций является tf.Tensor. Этот объект представляет собой частично определенное вычисление, которое в конечном итоге выдаст какое-либо

значение. Программы TensorFlow сначала строят граф объектов tf.Tensor, детализирует как каждый тензор будет вычисляться на других доступных тензорах, а затем запускает построение этого графа для получения желаемых результатов.

Объект tf.Tensor имеет следующие параметры:

- тип данных (например, float32, int32, или string)
- форму (shape)

Каждый элемент в тензоре имеет одинаковый тип данных, и этот тип всегда известен. Размерность, которая определяется количеством размерностей и размером каждого массива, может быть известна частично. Большинство операций производят тензоры полностью известных размерностей, если эти входные размерности также известны, но в некоторых случаях узнать размерность тензора можно только в режиме graph execution.

Достоинства TensorFlow:

- визуализация – TensorFlow поставляется с полным набором инструментов визуализации, которые упрощают понимание, отладку и оптимизацию приложений. Благодаря поддержке различных стилей (от изображений и аудио до гистограмм и графиков) появляется возможность быстро и просто создавать обширные глубокие нейронные сети;
- разработка мобильных приложений – TensorFlow Mobile имеет сокращенный объем кода и математические инструменты для уменьшения размеров модели. TensorFlow Mobile подходит для Android, а также идеален для ситуаций, когда доступ к сети нестабильный или дорогой;
- документация – с TensorFlow вы получаете доступ к обширной документации и учебным пособиям, которые могут помочь ускорить разработки в области искусственного интеллекта. TensorFlow также имеет большое и чрезвычайно активное сообщество пользователей, которые регулярно добавляют код и устраняют проблемы в GitHub.

1.8. Tf.keras

Keras - это высокоуровневый API для создания моделей глубокого обучения. Он используется для быстрого создания прототипов, сложных исследований, а также для создания приложений. Далее приведены три ключевые преимущества Keras API.

- Простота в использовании. Keras имеет простой интерфейс, оптимизированный для большинства распространенных задач глубокого обучения. Также он дает конкретные подсказки как быстро исправить возможные ошибки.
- Модульность. Модели Keras строятся при помощи объединения нескольких простых модулей, каждый из которых может быть настроен независимо, и не накладывает каких либо значительных ограничений.
- Легко расширить модель. Ты можешь создавать свои собственные модули, чтобы свободно выражать свои идеи для исследования. Создавай новые слои, функции потерь и разрабатывай современные модели глубокого обучения.
- Импорт tf.keras. Tf.keras - это реализация спецификации Keras API в TensorFlow. Это высокоуровневый API для построения моделей глубокого обучения с первоклассной поддержкой функционала TensorFlow, например eager execution, конвееры tf.data и алгоритмы оценки Estimators. tf.keras делает TensorFlow простым в использовании, не теряя в гибкости и производительности.

1.9. Обзор аналогов

В настоящее время существует не так много программных продуктов, позволяющих определить автора текста. Однако данные продукты отличаются подходами к выполнению данной задачи.

Например, система «Лингвоанализатор» использует такие методы распознавания, как марковские цепи и информационная энтропия. Анализ текста основан на статистическом анализе частот встречаемости графем. Для

успешного распознания текста системе требуется от 40 до 100 тысяч символов, а точность распознавания составляет 84 – 89 % [7]. В базе данной системы находятся 129 авторов, однако, многие авторы представлены 1-2 произведениями, что может не дать достаточной точности определения данных авторов. Также стоит отметить, что данной базе представлены только русские писатели-фантасты. Например, система некорректно определила авторство текста Льва Толстого.

Система «Автороведˮ основана на применении нейронных сетей, метода опорных векторов, а также статистическом анализе наиболее часто встречающихся триграмм и слов русского языка. Требуемый объем текста для данной системы – от 20 до 25 тысяч символов, точность распознавания 95 – 98 %. Данной системы нет в открытом доступе, однако она внедрена в воинской части 51952 и Центре Технологий Безопасности ТУСУР. Внедрение показало положительный результат, состоящий в повышении точности идентификации автора, снижении временных затрат на эксперименты за счет автоматизации процесса и применяемых подходов [8].

Система «Атрибуторˮ является онлайн лингвистическим процессором для машинного сравнения текстов и их классификации по параметрам индивидуального авторского стиля. Произведения подбирались так, чтобы тексты разных писателей имели как можно больше различий, а тексты одного писателя имели максимальные сходства [9]. На данный момент система обучена сравнивать только тексты романов. Для атрибуции достаточно примерно 20 – 25 тысяч символов. В базе системы хранятся данные о 103 русских писателях 19 и 20 веков. Согласно отзывам сайта, точность распознавания автора около 60%. По состоянию на апрель 2019 года при попытке определить автора текста открывается страница с ошибкой сервера.

2. АНАЛИЗ СОБРАННЫХ ДАННЫХ

2.1. Формирование корпуса

Для обучения и тестирования нейросети был сформирован корпус текстов русской литературы. Для этого были скачаны полные собрания сочинений русских авторов, удалены произведения, написанные в соавторстве, в каждом произведении удалены введения, примечания и фамилии автора.

Таблица 1. Содержимое корпуса

Автор	**Кол-во слов, тыс**	**Автор**	**Кол-во слов, тыс**	**Автор**	**Кол-во слов, тыс**
Стогов	892	Мамин-Сибиряк	1444	Мережковский	2624
Трифонов	895	Лесков	1466	Аксенов	2777
Бурносов	958	Прилепин	1533	Устинова	2918
Куприн	964	Ишков	1617	Семенов	3012
Платонов	981	Земляной	1625	Звездная	3781
Пелевин	994	Астафьев	1697	Ливадный	3836
Салтыков-Щедрин	1023	Достоевский	1739	Каменистый	3919
Булгаков	1074	Толстой	1755	Дворецкая	4493
Сорокин	1117	Стругацкие	1851	Булычев	4624
Тургенев	1130	Перумов	1924	Бушков	4756
Волоконский	1149	Акунин	2020	Маринина	4962
Казанцев	1184	Лукьяненко	2186	Донцова	5007
Пришвин	1202	Эренбург	2212	Белянин	5200
Макс Фрай	1421	Силлов	2529	Проханов	5219
Чехов	1427	Демина	2542	**ВСЕГО:**	**101679**

2.2. Анализ корпуса

Для максимально корректного обучения нейронной сети были выбраны авторы, писавшие в разных жанрах и в разное время. Из 44 авторов самыми популярными жанрами стали фантастика и реализм (рис. 2).

Рисунок 2 – Распределение авторов по жанрам

Также была составлена гистограмма распределения авторов по времени работы (рис. 3). Многие авторы работали на рубеже веков, поэтому они были добавлены в оба века. Стоит заметить, что из всех авторов 22 работают по сей день.

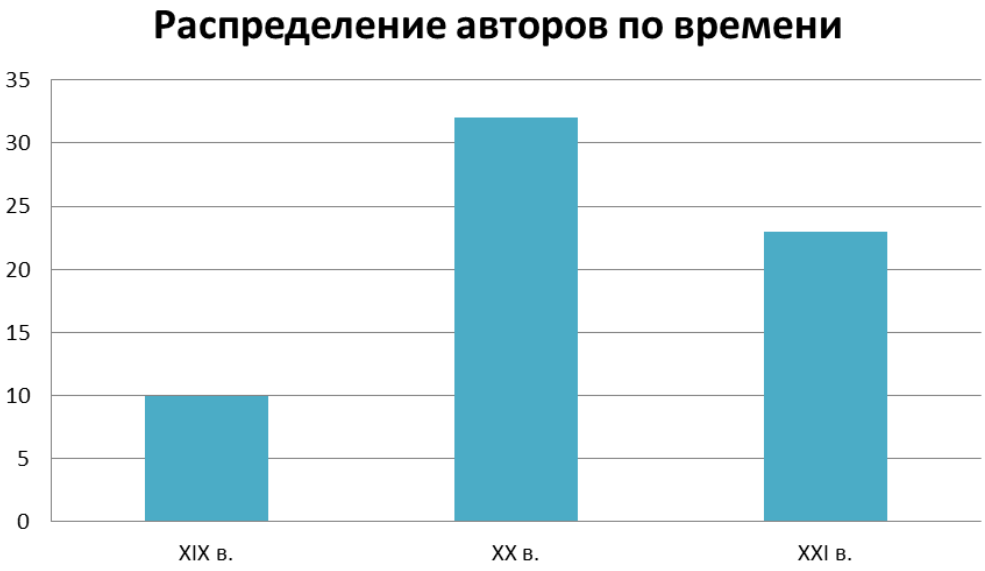

Рисунок 3 – Распределение авторов по времени

3. РЕАЛИЗАЦИЯ СИСТЕМЫ

3.1. ПО, отвечающее за подготовку данных

Для того чтобы на основе собранного корпуса текстов рассчитать выбранные метрики, тексты должны быть подготовлены следующим образом:

- удалены все знаки препинания;
- удалены все цифры и иностранные буквы;
- все слова приведены в начальную форму.

Для обработки данных таким образом было спроектировано и реализовано приложение на языке python.

Данное приложение реализует класс Normalizer, имеющий методы norm() – приводит слово в нормальную форму с использованием библиотеки pymorphy2, prepareText() – удаляет ненужные знаки, и normText() – реализует считывание текста из файла и запись отредактированного текста в новый файл. Также данное приложение для каждого автора в отдельности формирует из текстов файлы по wordsAmount слов в каждом.

Диаграмма последовательностей для процесса нормализации текста представлена на рисунке 4. Код приложения представлен в Приложении 1.

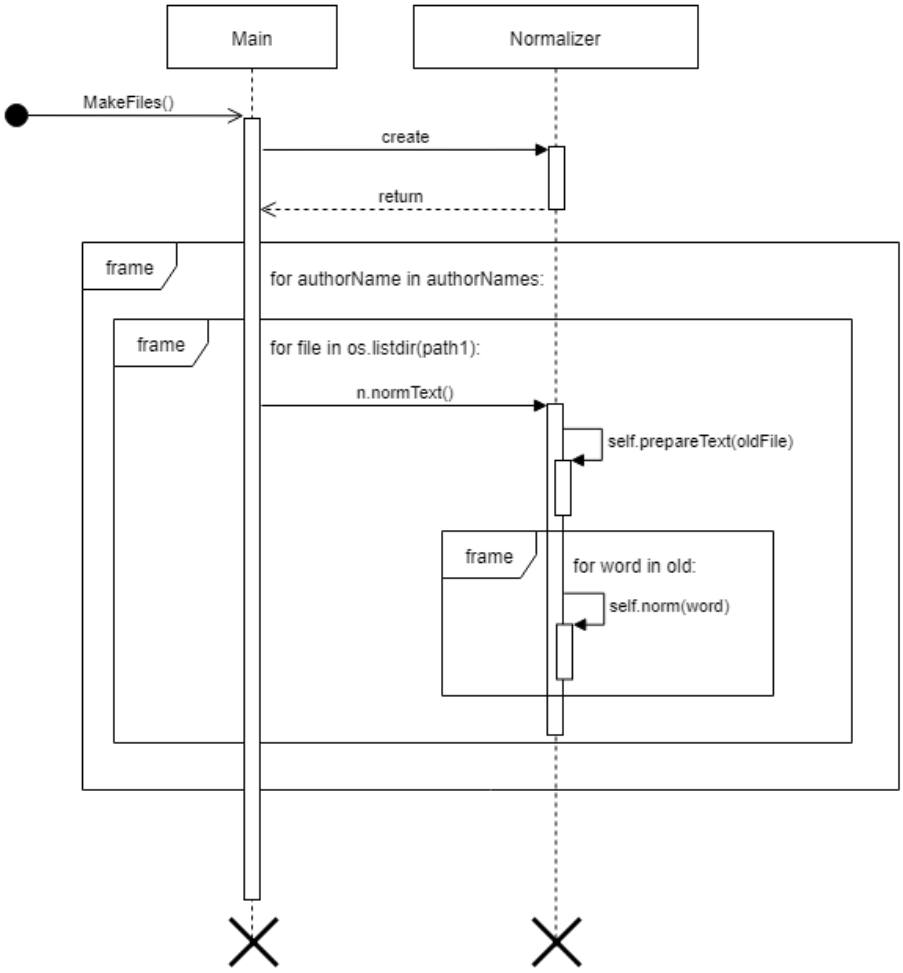

Рисунок 4 – Диаграмма последовательностей для процесса нормализации текста

На рисунках 5 и 6 представлен пример фрагментов исходного и полученного текстов.

— Вы нас не подбросите до Соловца?

Второй, с рыжей бородой и без усов, тоже улыбался, выглядывая из-за его плеча. Положительно, это были приятные люди.

— Давайте садитесь, — сказал я. — Один вперед, другой назад, а то у меня там барахло, на заднем сиденье.

— Благодетель! — обрадованно произнес горбоносый, снял с плеча ружье и сел рядом со мной.

Бородатый, нерешительно заглядывая в заднюю дверцу, сказал:

— А можно я здесь немножко того?..

Я перегнулся через спинку и помог ему расчистить место, занятое спальным мешком и свернутой палаткой. Он деликатно уселся, поставив ружье между колен.

— Дверцу прикройте получше, — сказал я.

Рисунок 5 – Фрагмент исходного текста

смуглый горбоносый лицо и спросить улыбаться вы мы не подбросить до соловец два с рыжеть борода и без усов тоже улыбаться выглядывать из за он плечо положительно это быть приятный человек давать садиться сказать я один вперёд другой назад а то у я там барахло на задний сидение благодетель обрадованно произнести горбоносый снять с плечо ружьё и сесть ряд с я бородатый нерешительно заглядывать в задний дверца сказать а можно я здесь немножко тот я перегнуться через спинка и помочь он расчистить место занятой спальная мешок и свёрнутый палатка он деликатно усесться поставить ружьё между колено дверца прикрыть хороший сказать я весь идти как обычно машина тронуться горбоносый

Рисунок 6 – Фрагмент отредактированного текста

Так как объем текстов достаточно большой, было важно добиться высокой скорости работы программы. В таблицах 2 и 3 представлены статистические данные по работе программы с разными входными данными.

Таблица 2 – Статистические данные по работе программы на базе текстов А. Беляева

Александр Беляев		
Время работы программы, с	Количество знаков в файле	Скорость работы программы, зн/с
28	401000	14321.4
29	479000	16517.2
34	488000	14352.9
41	523000	12756.1
43	616000	14325.6
43	540000	12558.1
46	657000	14282.6

Таблица 3 – Статистические данные по работе программы на базе текстов братьев Стругацких

Братья Стругацкие		
Время работы программы, с	Количество знаков в файле	Скорость работы программы, зн/с
45	815000	18111.11
47	876000	18638.3
49	909000	18551.02
53	998000	18830.19
54	990000	18333.33
54	994000	18407.41
60	994000	16566.67
65	1150000	17692.31
69	1057000	15318.84
87	1496000	17195.4

Было выяснено, что скорость работы программы зависит от количества знаков, которые необходимо удалить.

Средняя скорость обработки текстов составила 16236,7 знаков в секунду. В среднем тексты обрабатываются за 48 секунд.

3.2. Нейросеть, основанная на метрике TF-IDF

Нейросеть в данном проекте является сверточной и состоит из 2 слоев: входного и выходного. Основные параметры слоев представлены в таблице 4.

Таблица 4 – Основные параметры нейронной сети

Слой	Количество нейронов	Функция активации
Входной	75	Relu
Выходной	44	Softmax

На вход в качестве обучающих и тестовых данных нейросеть принимает вектор, в котором для каждого файла с текстом есть свой вектор, состоящий из метрики tf-idf. Так как количество авторов в корпусе равно 44, файлов для каждого автора в зависимости от количества слов в файле от 500 и больше, а уникальных слов в корпусе 325829, то в лучшем случае на вход нейросети поступал вектор векторов размером 22000×325829. Количество слов в файлах в данном исследовании варьировалось от 10000 до 1. Если предположить, что в файле все 10000 слов уникальны, то можно посчитать максимальную степень разреженности матрицы. Для этого необходимо разделить количество ненулевых элементов матрицы на общее количество элементов. В данном случае получаем:

$$\frac{10000 * 22000}{22000 * 325829} = 0{,}03$$

Это значит, что при самом благоприятном раскладе в нашей матрице доля значащих цифр будет составлять только 3% от всей матрицы. Остальное – нули. Чтобы избежать ненужной траты места, было предложено решение

хранить матрицу в разреженом виде. А для того, чтобы нейросеть смогла работать с такой матрицей, был использован batch_generator. Данная функция берет часть входного вектора и приводит его в обычный вид. Дальше нейросеть обучается по этой части, после чего берет следующую часть вектора и обучается на ней.

Также для данной нейросети есть возможность сохранять данные об обученной сети в файл и восстанавливать сеть из файла, что позволяет существенно сократить время на повторном обучении сети.

Для того чтобы оценить точность работы нейросети, из обучающей выборки случайным образом были выбраны 10% файлов и перемещены в отдельную папку. Они составили валидационную выборку. Данные файлы не использовались в процессе обучения, а значит, на них можно проверить работу сети. Валидационная выборка также хранится в разреженном виде и восстанавливается в нормальный вид посредством batch_generator.

3.3. Нейросеть, основанная на метрике word2vec

Данная нейросеть состоит из трех слоев: входного, выходного и скрытого. Количество нейронов на скрытом слое в литературе советуют определять как среднее геометрическое между количеством нейронов во входном и выходном слое.

Таблица 5 – Основные параметры нейронной сети

Слой	Количество нейронов	Функция активации
Входной	75	Relu
Скрытый	55	Relu
Выходной	44	Softmax

Как было сказано выше, word2vec – семейство алгоритмов, которые обучаются на большом количестве данных. После обучения они способны вернуть векторное представление слов. Для того чтобы векторные преставления были как можно точнее, сеть нужно обучать очень тщательно.

Однако в свободном доступе есть уже предобученные модели. Такая модель была скачана с сайта rusvectores.org. Она была создана в январе 2019 года, содержит в себе 270 миллионов слов и была обучена на национальном корпусе русского языка. С помощью библиотеки gensim модель была загружена в программу. Сложностью стало то, что в словаре хранятся слова с тегами, отражающими их части речи (не «мама», а «мама_NOUN»). Поэтому был использован MorphAnalyzer из библиотеки pymorphy2, который может автоматически определить часть речи. Для того чтобы найти векторное представление текста, суммировались векторные представления слов, входящих в него. Данные векторные представления использовались в качестве обучающей, тестовой и валидационной выборки.

3.4. Результаты работы сети

Был проведен анализ работы сети на разных входных данных. Для этого в качестве входных данных были использованы кусочки текста, размером 10000, 5000, 4000, 3000, 2000, 1000, 100, 50, 25, 10, 1 слов.

На рисунках 7 и 8 представлен график поведения точности тестовой, обучающей и валидационной выборки в зависимости от количества входных слов для метрики TF-IDF, а на рисунках 9 и 10 – для word2vec. По оси X на данных графиках расположены количество слов во входном тексте, а по оси Y – точность нейросети.

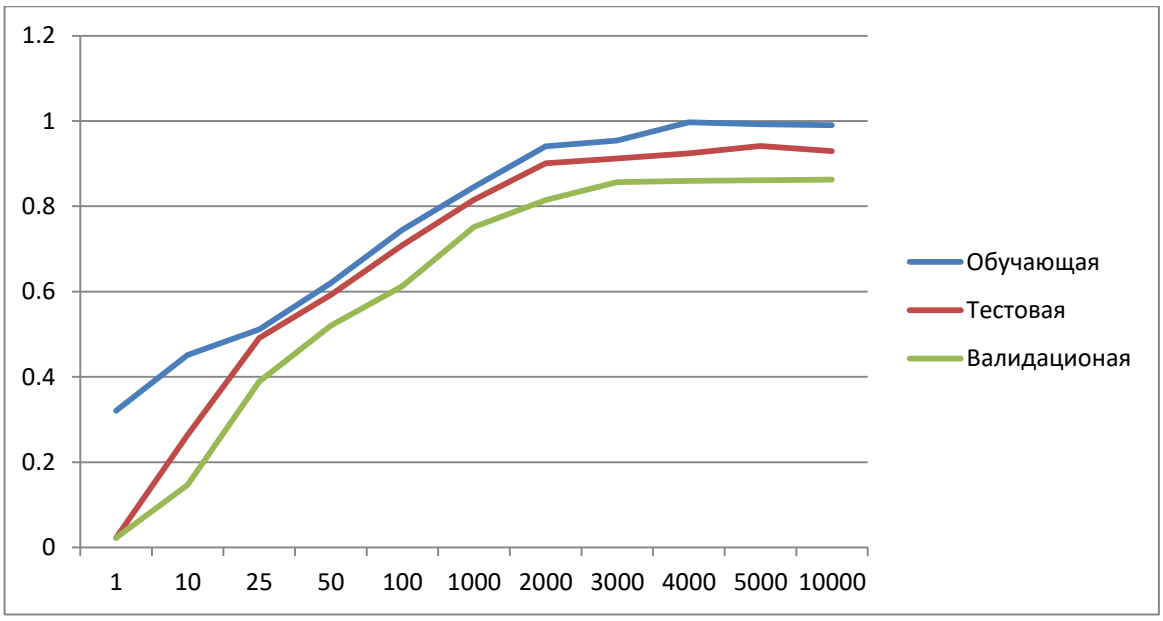

Рисунок 7 – Точность на различных выборках для метрики TF-IDF

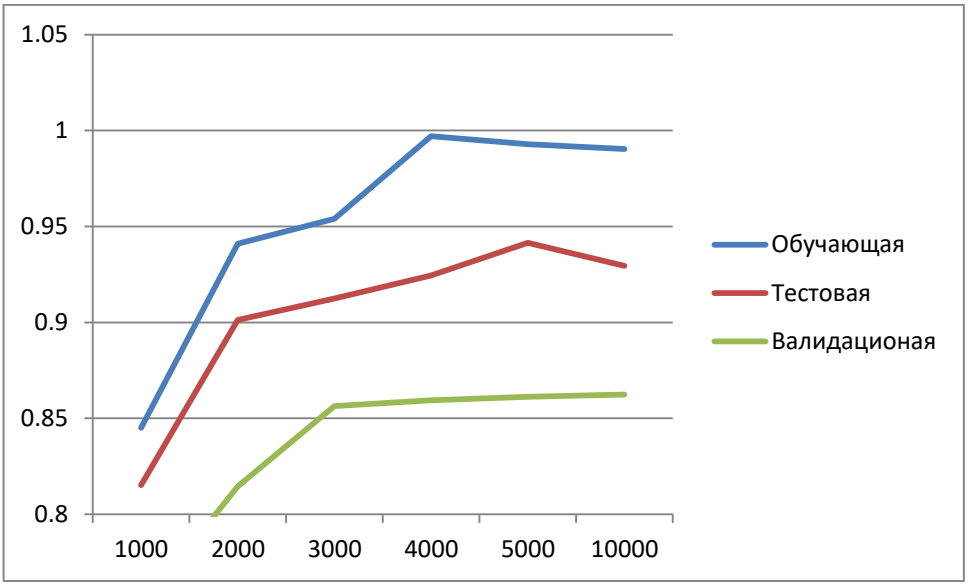

Рисунок 8 – Фрагмент графика точности на различных выборках для метрики TF-IDF

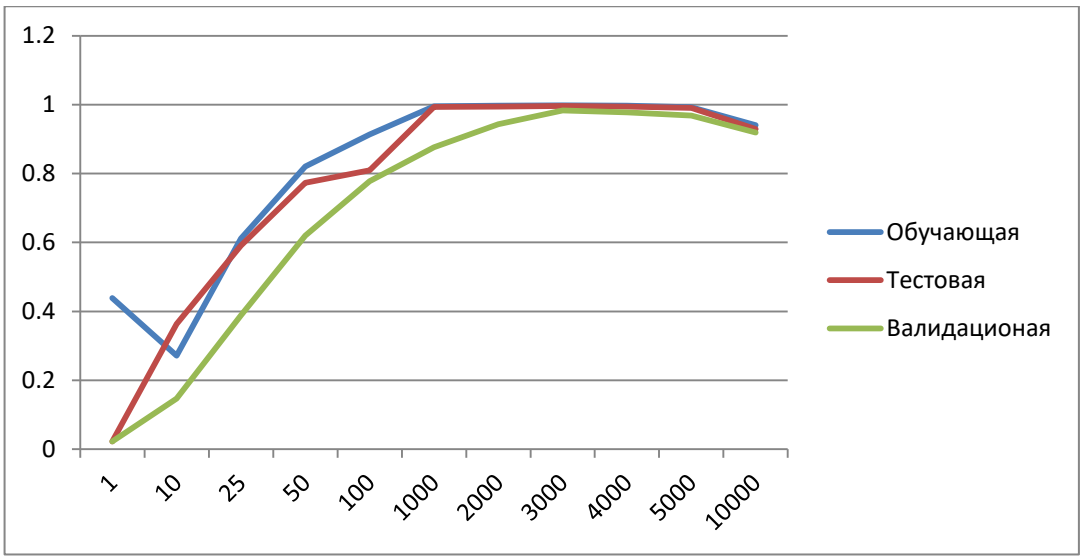

Рисунок 9 – Точность на различных выборках для метрики word2vec

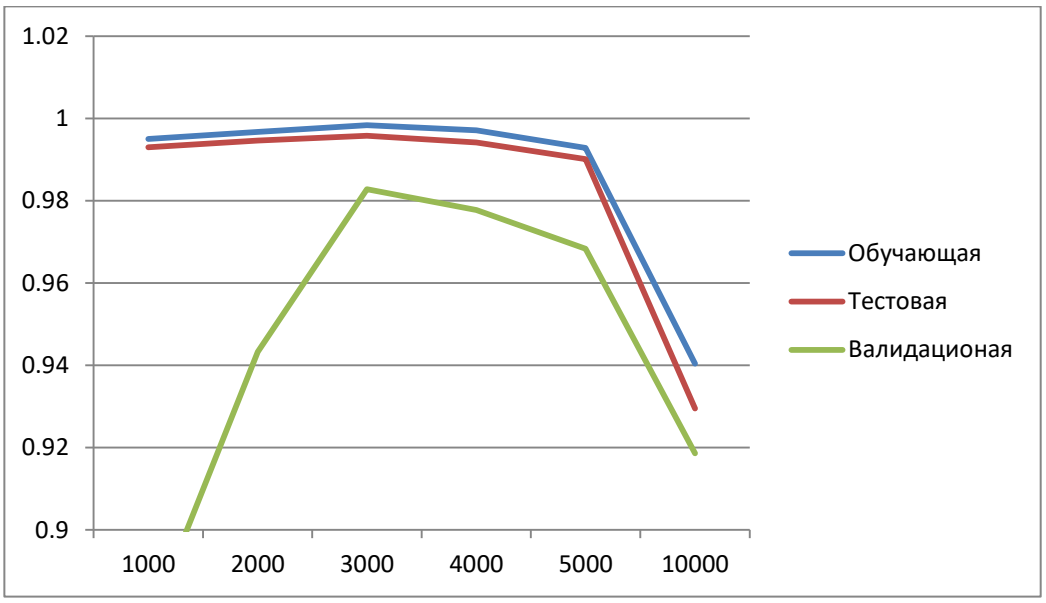

Рисунок 10 – Фрагмент графика точности на различных выборках для метрики word2vec

При числе входных слов, равному единице, точность валидационной выборки в обоих случаях равно 0,022. Эта точность случайного угадывания автора. Далее при увеличении количества слов точность обеих сетей увеличивается. Лучший показатель нейросеть с метрикой word2vec дает при 3000 входных слов (точность валидационной выборки 0.98). При 10000 слов точность выборки немного падает. Возможно, это связано с шумами. Метрика TF-IDF дает максимальную точность валидационной выборки при 10000 слов (0.86), что, однако, меньше, чем метрика word2vec.

4. ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ

4.1. Оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности и ресурсосбережения

4.1.1. Потенциальные потребители результатов исследования

Система предназначена для определения автора текста по его фрагменту. Спектр применения данного продукта достаточно широк: от отыскания автора необходимой вам статьи в интернете или запоминающегося отрывка художественного произведения до достаточно серьёзных военных целей. Потребителями данной системы может стать любой: это может быть литературовед, военный или же обычный пользователь ПК.

4.1.2. Анализ конкурентных технических решений

В настоящее время существует не так много программных продуктов, позволяющих определить автора текста. Однако данные продукты отличаются подходами к выполнению данной задачи.

Например, система «Лингвоанализатор» использует такие методы распознавания, как марковские цепи и информационная энтропия. Анализ текста основан на статистическом анализе частот встречаемости графем. Для успешного распознания текста системе требуется от 40 до 100 тысяч символов, а точность распознавания составляет 84 – 89 %. В базе данной системы находятся 129 авторов, однако, многие авторы представлены 1-2 произведениями, что может не дать достаточной точности определения данных авторов. Также стоит отметить, что данной базе представлены только русские писатели-фантасты. Например, система некорректно определила авторство текста Льва Толстого.

Система «Автороведˮ основана на применении нейронных сетей, метода опорных векторов, а также статистическом анализе наиболее часто встречающихся триграмм и слов русского языка. Требуемый объем текста для данной системы – от 20 до 25 тысяч символов, точность распознавания 95 – 98 %. Однако данной системы нет в открытом доступе.

Система «Атрибуторˮ является онлайн лингвистическим процессором для машинного сравнения текстов и их классификации по параметрам индивидуального авторского стиля. Произведения подбирались так, чтобы тексты разных писателей имели как можно больше различий, а тексты одного писателя имели максимальные сходства. На данный момент система обучена сравнивать только тексты романов. Для атрибуции достаточно примерно 20 – 25 тысяч символов. В базе системы хранятся данные о 103 русских писателях 19 и 20 веков. Согласно отзывам сайта, точность распознавания автора около 60%. По состоянию на апрель 2019 года при попытке определить автора текста открывается страница с ошибкой сервера.

Определим факторы конкурентоспособности: точность распознавания автора, минимальный объем текста для распознавания, удобство использования и дизайн, размер базы и охват авторов, и составим оценочную карту сравнения конкурентных технических решений (Табл. 6).

Таблица 6 – Оценочная карта для сравнения конкурентных технических решений (разработок)

№ п/п	Конкуренты	Факторы конкурентоспособности					Итоговая оценка
		Точность распознавания	Минимальный объем текста	Удобство использования и дизайн	Размер базы	Охват авторов	
1	Лингво-анализатор	8/1,92	8/1,52	5/0,7	8/1,52	5/1,2	6,86
2	Авторовед	10/2,4	10/1,9	1/0,14	8/1,52	8/0,92	6,88
3	Атрибутор	4/0,96	10/1,9	8/1,12	8/1,52	5/1,2	6,7

Продолжение таблицы 6 – Оценочная карта для сравнения конкурентных технических решений (разработок)

4	Данный проект	8/1,92	8/1,52	9/1,26	7/1,33	9/2,16	8,19
	b_j	5	4	3	4	5	21
	w_j	0,24	0,19	0,14	0,19	0,24	-

Как видно из таблицы, разрабатываемый продукт имеет достаточные преимущества перед конкурентами. В основном аналоги проигрывают в удобстве использования и дизайне (у системы «Авторовед» поставлена 1, так как данной системы нет в свободном доступе) и охвате авторов.

По данным таблицы 6 был построен многоугольник конкурентоспособности (рис. 11).

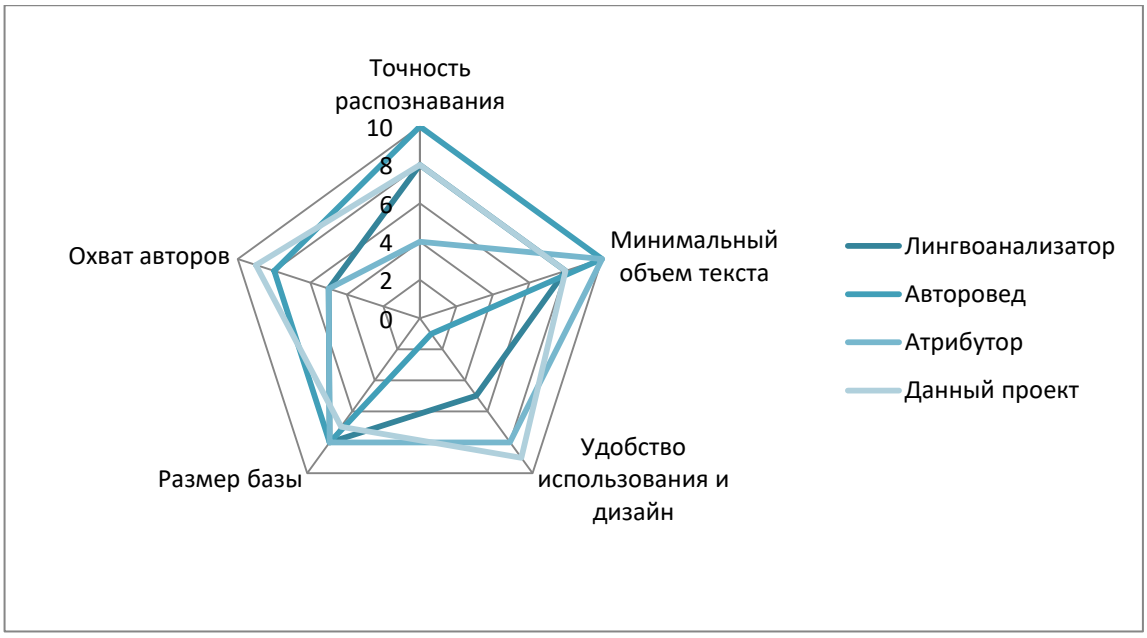

Рисунок 11 – Многоугольник конкурентоспособности

4.1.3. SWOT-анализ

SWOT-анализ – это метод оценки внутренних и внешних факторов, которые влияют на развитие компании или проекта. Эта методика помогает оценить сильные и слабые стороны проекта, найти новые возможности и определить возможные угрозы.

Таблица 7 – SWOT-анализ

	Внутренние факторы		
		Сильные стороны проекта: 1. Высокая конкурентоспособность 2. Небольшая стоимость 3. Удобство использования 4. Использование доступных передовых технологий	Слабые стороны проекта: 1. Сложности в раскрутке бренда 2. Высокие интеллектуальные затраты на реализацию проекта
Внешние факторы	Возможности: 1. Публикация научных статей о продукте 2. Интерес со стороны специалистов ТПУ 3. Невысокая популярность конкурентов	Высокая конкурентоспособность, удобство использования и невысокая популярность конкурентов обеспечивают все возможности для разрабатываемого продукта выбиться в лидеры данной отрасли. Использование передовых технологий и интерес со стороны специалистов ТПУ позволит выпускать различные научные статьи о продукте, что также сделает его популярнее.	Сложности в раскрутке бренда могу быть компенсированы низкой популярностью конкурентов, а также достаточным количеством научных статей о продукте.
	Угрозы: 1. Улучшение качества продукта конкурентов 2. Нехватка времени на реализацию 3. Недостаточная востребованность продукта	Улучшение качества продуктов конкурентов может поставить популярность нашего продукта под угрозу, однако низкая стоимость, высокая конкурентоспособность и удобство использования могут компенсировать это.	Сложности в раскрутке бренда, недостаточная востребованность продукта и улучшение качества конкурентов могут оказаться проблемой для популярности разрабатываемого продукта.

4.2. Планирование научно-исследовательских работ

4.2.1. Структура работ в рамках научного исследования

Проектирование предполагает выполнение определённых стадий и этапов. Оно включает составление в текстовой и (или) графической форме плана работ. Для успешной реализации проекта необходимо устанавливать реальные этапы с чётко обозначенными началом и окончанием. Разработка детального плана работ связана с описанием процессов и их последовательности, выполняемых на каждом этапе, необходимых для этого специалистов, средств и ресурсов. Такой подход в большей степени позволяет избежать упущений и ошибок. Он необходим работникам, реализующим внедрение проекта автоматизации, а также оказывает положительное воздействие на лиц, его финансирующих.

Таблица 8 – Перечень работ и распределение исполнителей

№ работы	Наименование работы	Исполнители работы
1	Выбор научного руководителя бакалаврской работы	Демиденко Л. Р.
2	Составление и утверждение темы бакалаврской работы	Цапко С.Г., Демиденко Л.Р.
3	Составление календарного плана-графика выполнения бакалаврской работы	Цапко С.Г.
4	Подбор и изучение литературы по теме бакалаврской работы	Цапко С.Г., Демиденко Л.Р.
5	Анализ предметной области	Цапко С.Г., Демиденко Л.Р.
6	Подбор корпуса текстов	Цапко С.Г., Демиденко Л.Р.
7	Реализация программного кода для обработки корпуса текстов	Цапко С.Г., Демиденко Л.Р.
8	Обработка корпуса текстов	Цапко С.Г., Демиденко Л.Р.
9	Проектирование нейронной сети	Цапко С.Г., Демиденко Л.Р.

Продолжение таблицы 8 – Перечень работ и распределение исполнителей

10	Разработка нейронной сети	Цапко С.Г., Демиденко Л.Р.
11	Тестирование нейронной сети	Цапко С.Г., Демиденко Л.Р.
12	Согласование выполненной работы с научным руководителем	Цапко С.Г., Демиденко Л.Р.
13	Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	Демиденко Л.Р.
14	Подведение итогов, оформление работы	Демиденко Л.Р.

Из таблицы 8 можно видеть, что вклад студента в исследование составляет около 87%, тогда как вклад научного руководителя – 13%.

4.2.2. Определение трудоемкости выполнения работ

Большой вклад в стоимость внедрения системы дают трудовые затраты, для расчета которых необходимо определить трудоемкость работ.

Трудоемкость оценивается по следующей формуле:

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

где $t_{\text{ож}\,i}$ – ожидаемая трудоемкость i-ой работы, чел.-дни,

$t_{\min i}$ – оптимистическая оценка (минимально возможная трудоемкость выполнения i-ой работы), чел.-дни,

$t_{\max i}$ – пессимистическая оценка (максимально возможная трудоемкость выполнения i-ой работы), чел.-дни.

Далее необходимо определить продолжительности каждой работы в рабочих днях:

$$T_{p_i} = \frac{t_{\text{ож}\,i}}{\text{Ч}_i} \quad (6)$$

где T_{p_i} – продолжительность одной работы, раб. дни,

$t_{\text{ож}\,i}$ – ожидаемая трудоемкость выполнения одной работы, чел.-дни,

$Ч_i$ – это численность исполнителей, выполняющих одновременно одну и ту же работу на этом этапе, чел..

Переведем длительность работ в календарные дни. В дальнейшем это будет необходимо для построения диаграммы Ганта.

$$T_{ki} = T_{pi} \cdot k_{кал} \quad (7)$$

где T_{ki} – продолжительность выполнения i-й работы в календарных днях;

T_{pi} – продолжительность выполнения i-й работы в рабочих днях;

$k_{кал}$ – коэффициент календарности.

4.2.3. Разработка графика проведения научного исследования

Коэффициент календарности рассчитывается по следующей формуле:

$$k_{кал} = \frac{T_{кал}}{T_{кал} - T_{вых} - T_{пр}} \quad (8)$$

где $k_{кал}$ – коэффициент календарности;

$T_{кал}$– количество календарных дней в году;

$T_{вых}$ – количество выходных дней в году;

$T_{пр}$ – количество праздничных дней в году.

Рассчитаем коэффициент календарности. Согласно производственному календарю (для 6-дневной рабочей недели) в 2019 году 365 календарных дней, 299 рабочих дней, 66 выходных/праздничных дней.

$$k_{кал} = \frac{365}{365 - 66} = 1{,}22$$

Таблица 9 – Временные показатели проведения научного исследования

Наименование работы	Исполнители работы	Трудоемкость работ, чел-дни			Длительность работ, дни	
		t_{min}	t_{max}	$t_{ож}$	$T_р$	$T_к$
Выбор научного руководителя бакалаврской работы	Демиденко Л.Р.	1	2	1,4	1	2
Составление и утверждение темы бакалаврской работы	Демиденко Л.Р.	2	3	2,4	2	3
	Цапко С.Г.	1	2	1,4	1	2
Составление календарного плана-графика выполнения бакалаврской работы	Цапко С.Г.	1	2	1,4	1	2
Подбор и изучение литературы по теме бакалаврской работы	Демиденко Л.Р.	14	18	15,6	16	19
	Цапко С.Г.	1	1	1,0	1	1
Анализ предметной области	Демиденко Л.Р.	5	6	5,4	5	7
	Цапко С.Г.	1	1	1,0	1	1
Подбор корпуса текстов	Демиденко Л.Р.	20	25	22,0	22	27
	Цапко С.Г.	1	1	1,0	1	1
Реализация программного кода для обработки корпуса текстов	Демиденко Л.Р.	5	7	5,8	6	7
	Цапко С.Г.	1	1	1,0	1	1
Обработка корпуса текстов	Демиденко Л.Р.	2	3	2,4	2	3
	Цапко С.Г.	1	1	1,0	1	2
Проектирование нейронной сети	Демиденко Л.Р.	3	4	3,4	3	4
	Цапко С.Г.	1	2	1,4	1	2
Разработка нейронной сети	Демиденко Л.Р.	5	7	5,8	6	7
	Цапко С.Г.	1	2	1,4	1	2

Продолжение таблицы 9 – Временные показатели проведения научного исследования

Тестирование нейронной сети	Демиденко Л.Р.	15	20	17,0	17	21
	Цапко С.Г.	1	2	1,4	1	2
Согласование выполненной работы с научным руководителем	Демиденко Л.Р.	2	3	2,4	2	3
	Цапко С.Г.	2	3	2,4	2	3
Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	Демиденко Л.Р.	12	17	14,0	14	17
Подведение итогов, оформление работы	Демиденко Л.Р.	5	8	6,2	6	8
ИТОГО:	Демиденко Л.Р.	91	123	103,8	104	127
	Цапко С.Г.	12	18	14.4	14	18

По данным таблицы 9 в системе GantPro была построена диаграмма Ганта (рис. 12). Данная диаграмма позволяет легко визуализировать сроки по управлению проектами, трансформируя названия задач, даты начала и окончания выполнения, а также длительность в каскадные горизонтальные гистограммы.

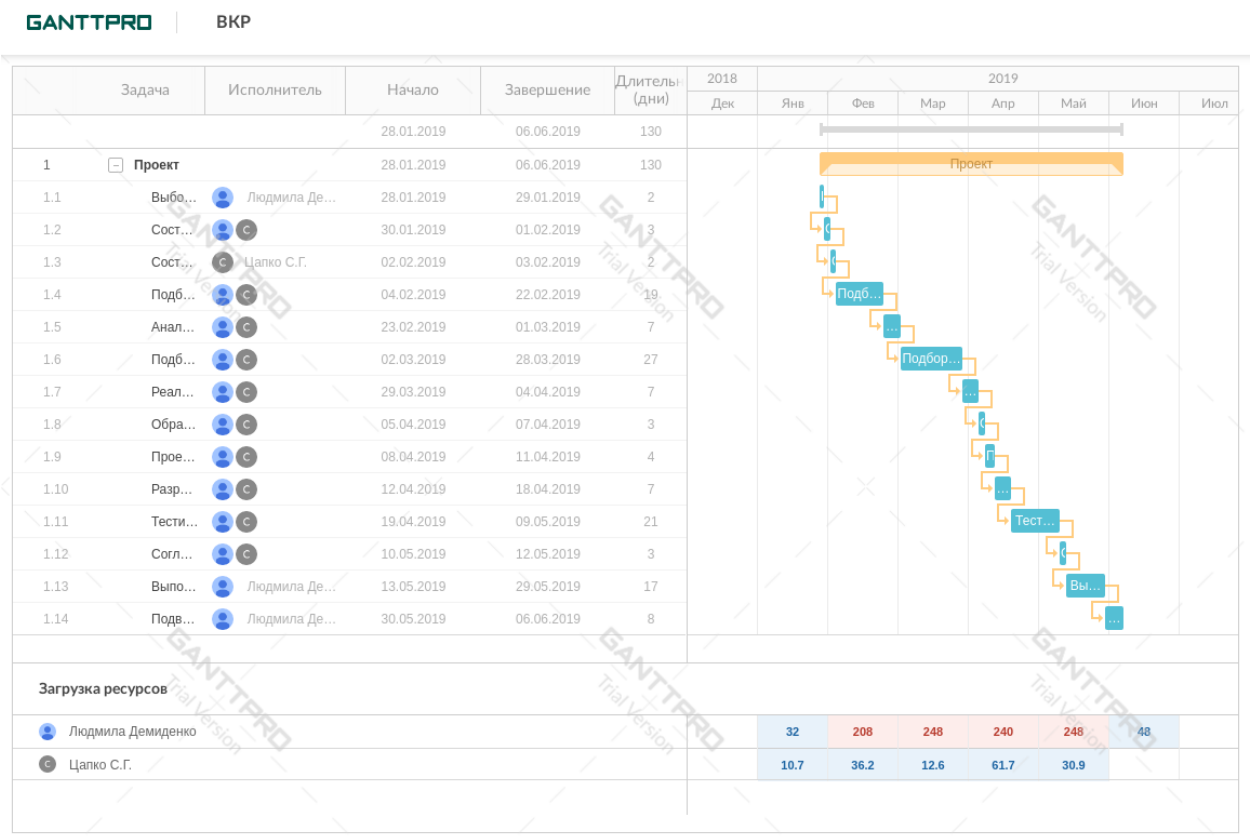

Рисунок 12 – Диаграмма Ганта

4.2.4. Бюджет научно-технического исследования

4.2.4.1. Расчет материальных затрат научно-технического исследования

Материальные расходы в данном исследовании складываются их затрат на канцелярские принадлежности. Данные расходы составляют 2000 рублей.

4.2.4.2. Расчет затрат на специальное оборудование для научных (экспериментальных) целей

Для данной работы не приобреталось никакого специального оборудования. Однако был использован ноутбук и специальное ПО.

Рассчитаем амортизацию ноутбука. Первоначальная стоимость ПК 40000 рублей; срок полезного использования для машин офисных код 330.28.23.23 составляет 2-3 года, берем 3 года. Срок использования ПК для написания ВКР – 5 месяцев.

Норма амортизации рассчитывается по формуле:

$$Ан = \frac{1}{n} * 100\% \quad (9)$$

где Ан – норма амортизации, %,

n – срок полезного использования, г.

Рассчитаем норму амортизации:

$$Ан = \frac{1}{3} \times 100\% = 33{,}33\%$$

Годовые амортизационные отчисления рассчитываются по формуле:

$$Аг = Ст \times Ан \quad (10)$$

где Аг – годовые амортизационные отчисления, руб.,

Ст – стоимость оборудования, руб.

Рассчитаем годовые амортизационные отчисления:

$$Аг = 40000 \times 0{,}33 = 13200 \text{ руб.}$$

Ежемесячные амортизационные отчисления рассчитываются по формуле:

$$Ам = \frac{Аг}{12} \quad (11)$$

Рассчитаем ежемесячные амортизационные отчисления:

$$Ам = \frac{13200}{12} = 1100 \text{ руб.}$$

Итоговая сумма амортизации основных средств рассчитываются по формуле:

$$А = Ам \times t \quad (12)$$

Рассчитаем итоговую сумму амортизации основных средств:

$$А = 1100 \times 5 = 5500 \text{ руб.}$$

Так как в работе использовалось бесплатное ПО, то сумма его амортизации равна 0. Поэтому в затраты на специальное оборудование необходимо включить сумму амортизации основных средств в сумме 5500 руб.

Таблица 10 – Расчет затрат на амортизацию

Наименование	Затраты, руб.
Амортизация ПК	5500

4.2.4.3. Основная заработная плата исполнителей темы

В настоящую статью включается основная заработная плата научных работников, непосредственно участвующих в выполнении работ по данной теме. Величина расходов по заработной плате определяется исходя из трудоемкости выполняемых работ и действующей системы окладов и тарифных ставок.

Для расчета основной заработной платы студента возьмем оклад, равный окладу ассистента без степени, 21760 руб.

Руководителем ВКР является Сергей Геннадьевич Цапко, доцент, кандидат наук. Таким образом, для расчетов берем его оклад равным 33664 руб.

Затраты на заработную плату рассчитываются по следующей формуле:

$$З_п = З_{осн} + З_{доп} \quad (13)$$

где $З_{осн}$ – основная заработная плата, руб.;

$З_{доп}$ – дополнительная заработная плата, руб.

Формула для расчета основной заработной платы:

$$З_{осн} = З_{дн} \times Т_р \times (1 + К_{пр} + К_д) \times К_р \quad (14)$$

где $З_{дн}$– среднедневная заработная плата, руб.;

$К_{пр}$– премиальный коэффициент (0,3);

$К_д$– коэффициент доплат и надбавок (0,2-0,5);

$К_р$– районный коэффициент (для Томска 1,3);

$Т_р$– продолжительность работ, выполняемых работником, раб. дни.

Среднедневная заработная плата:

$$З_{дн} = \frac{З_м \times М}{F_д} \quad (15)$$

где $З_м$– месячный оклад работника, руб.

М – количество месяцев работы без отпуска в течение года (М =11,2 месяца для 5-дневной недели и М=10,4 месяца для 6-дневной);

$F_д$– действительный годовой фонд рабочего времени персонала, раб. дн.

Таблица 11 – Баланс рабочего времени (для 6-дневной недели)

Показатели рабочего времени	Дни
Календарные дни	365
Нерабочие дни (праздники/выходные)	66
Потери рабочего времени (отпуск/невыходы по болезни)	56
Действительный годовой фонд рабочего времени	243

Рассчитаем среднедневную зарплату для студента:

$$З_{дн} = \frac{З_м \times М}{F_д} = \frac{21760 \times 10{,}4}{243} = 931{,}29 \text{ руб.}$$

Рассчитаем среднедневную зарплату для руководителя диплома:

$$З_{дн} = \frac{З_м \times М}{F_д} = \frac{33664 \times 10{,}4}{243} = 1440{,}76 \text{ руб.}$$

Таблица 12 – Расчет основной заработной платы

Исполнители	$З_{дн}$, руб.	$К_{пр}$	$К_д$	$К_р$	$Т_р$	$З_{осн}$
Студент	931,29	0,3	0,2	1,3	104	188865,61
Научный руководитель	1440,76	0,3	0,2	1,3	14	39332,85
Итого:						228198,46

4.2.4.4. Дополнительная заработная плата исполнителей темы

В состав дополнительной заработной платы входят выплаты, положенные работнику за время отсутствия на рабочем месте в рамках трудового законодательства, а также иные выплаты мотивирующего характера, не связанные с исполнением должностных функций.

Дополнительная заработная плата рассчитывается по формуле:

$$З_{доп} = З_{осн} * 0{,}12 \quad (16)$$

Рассчитаем дополнительную заработную плату для студента:

$$З_{доп} = 188865{,}61 * 0{,}12 = 22663{,}87$$

Рассчитаем дополнительную заработную плату для руководителя ВКР:

$$З_{доп} = 39332{,}85 * 0{,}12 = 4719{,}94$$

Таким образом общая дополнительная плата составила 27383,8156 руб.

4.2.4.5. Отчисления во внебюджетные фонды (страховые отчисления)

Страховые взносы – это регулярные обязательные платежи. Уплата взносов дает право на получение больничных и детских пособий, бесплатной медицинской помощи, финансовой поддержки при выходе на пенсию.

Взносы делят на две группы: страховые взносы в фонды и в ИФНС.

К первой группе относят отчисления во внебюджетные фонды из заработной платы работников на страхование от несчастных случаев на производстве и профзаболеваний. Такие отчисления принято называть взносами на травматизм. Делают их в Фонд соцстраха.

Ко второй группе относят взносы на пенсионное, медицинское и социальное страхование на случай временной нетрудоспособности и в связи с материнством.

Согласно Федеральному закону от 24.07.2009 №212-ФЗ размер коэффициента отчислений на уплату во внебюджетные фонды в настоящее время установлен в размере 30%.

Сумма отчислений во внебюджетные фонды рассчитывается по формуле:

$$З_{внеб} = (З_{осн} + З_{доп}) * 0{,}3 \quad (17)$$

Рассчитаем сумму отчислений во внебюджетные фонды для студента:

$$З_{внеб} = (188865{,}61 + 22663{,}87) * 0{,}3 = 63458{,}85$$

Рассчитаем сумму отчислений во внебюджетные фонды для руководителя ВКР:

$$З_{внеб} = (39332{,}85 + 4719{,}94) * 0{,}3 = 13215{,}84$$

Таким образом общая дополнительная плата составила 76674,68 руб.

4.2.4.6. Накладные расходы

Накладные расходы – это затраты на бизнес-процессы, дополняющие и поддерживающие производственную деятельность. К таким бизнес-процессам относятся управление, организация производства, командировки и обучение сотрудников.

Сумма накладных расходов рассчитывается по формуле:

$$З_{нкл} = (З_{мат} + З_{обор} + З_{осн} + З_{доп} + З_{внеб}) * 0{,}16 \quad (18)$$

Рассчитаем сумму накладных расходов:

$$З_{нкл} = (2000 + 5500 + 228198{,}46 + 27383{,}82 + 76674{,}68) * 0{,}16 = 54361{,}11$$

4.2.4.7. Формирование бюджета затрат научно-исследовательского проекта

Выше были посчитаны отдельные статьи расходов на выполнение ВКР. На основе этих данных можно сформировать бюджет проекта (табл. 8).

Таблица 13 – Расчет бюджета затрат НТИ

Наименование	Сумма, руб.	Удельный вес, %
Материальные затраты	2000	0.5
Затраты на специальное оборудование	5500	1,4
Затраты на основную заработную плату	228198,46	57,9
Затраты на дополнительную заработную плату	27383,82	6,9
Страховые взносы	76674,68	19,5
Накладные расходы	54361,11	13,8
Общий бюджет	394118,08	100

Как видно из таблицы, общий бюджет исследования составляет 394118,08 рублей. Самый большой удельный вес имеют затраты на основную заработную плату (57,9%), а самый маленький – материальные затраты (0,5%).

4.3.Определение потенциального эффекта исследования

В ходе анализа было выяснено, что основной целевой группой разрабатываемого продукта являются обычные пользователи ПК. Они получат возможность определить авторство интересующего их текста.

Анализ конкурентоспособности показал, что данная система может смело конкурировать с аналогами. Длительность исследования составит около 5 месяцев. Также был проведен анализ стоимости исследования, в результате которого было рассчитано, что потенциальная стоимость проекта составляет 394118,08.

5. СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ

ВВЕДЕНИЕ

В ходе выполнения данной работы была спроектирована и реализована нейросеть, позволяющая определить автора по его тексту. Пользователями данной системы могут являться как простые пользователи ПК, так и литературоведы и даже военные.

Выпускная квалификационная работа студента выполнялась в десятом корпусе ТПУ (первый этаж) в 108 и 109 аудиториях.

5.1. Правовые и организационные вопросы обеспечения безопасности

Данное исследование выполнялось в основном сидя за компьютером. Поэтому главным аспектом в организационных вопросах обеспечения безопасности стали требования к организации рабочих мест пользователей:

- Рабочее место должно быть организовано с учетом эргономических требований согласно ГОСТ 12.2.032-78 «ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования» и ГОСТ 12.2.061-81 «ССБТ. Оборудование производственное. Общие требования безопасности к рабочим местам». СаНПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы»;
- Конструкция рабочей мебели (рабочий стол, кресло, подставка для ног) должна обеспечивать возможность индивидуальной регулировки соответственно росту пользователя и создавать удобную позу для работы. Вокруг ПК должно быть обеспечено свободное пространство не менее 60-120см;
- На уровне экрана должен быть установлен оригинал-держатель.

На рисунке 13 схематично представлены требования к рабочему месту.

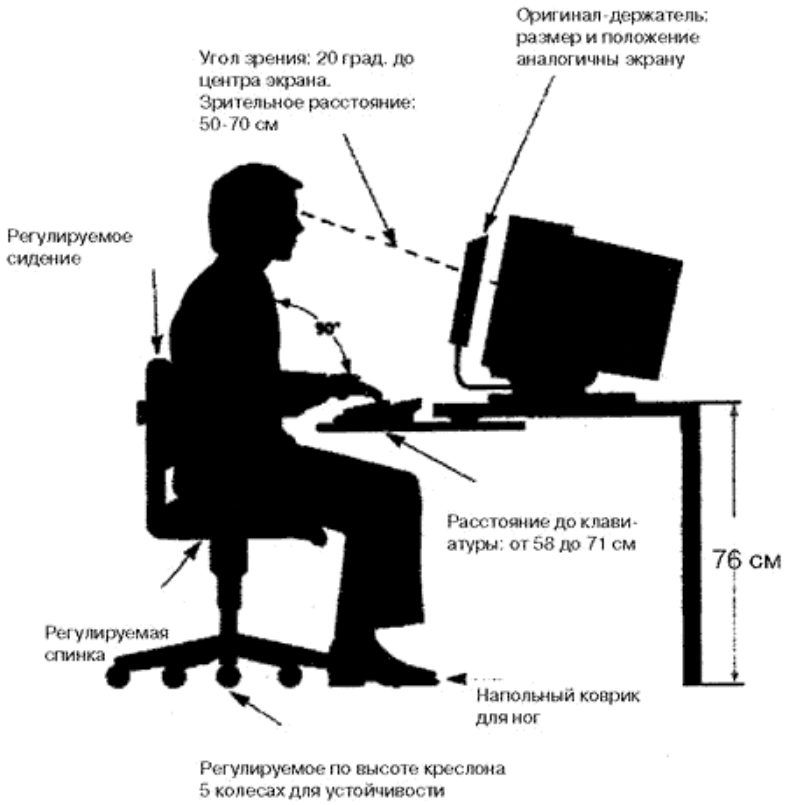

Рисунок 13 – Организация рабочего места

В соответствии с государственными стандартами и правовыми нормами обеспечения безопасности предусмотрена рациональная организация труда в течение смены, которая предусматривает:

- длительность рабочей смены не более 8 часов;
- установление двух регламентируемых перерывов (не менее 20 минут после 1-2 часов работы, не менее 30 минут после 2 часов работы);
- обеденный перерыв не менее 40 минут.

Обязательно предусмотрен предварительный медосмотр при приеме на работу и периодические медосмотры.

Каждый сотрудник должен пройти инструктаж по технике безопасности перед приемом на работу и в дальнейшем, должен быть пройден инструктаж по электробезопасности и охране труда.

5.2. Производственная безопасность

Производственные факторы согласно ГОСТ 12.0.003-2015 подразделяются на опасные и вредные. Опасным производственным

фактором называется фактор, воздействие которого приводит к травме или резкому ухудшению здоровья. Вредным производственным фактором является фактор, воздействие которого приводит к заболеванию или снижению работоспособности.

На оператора ПЭВМ в течение рабочего дня воздействует множество различных производственных факторов, каждый из которых влияет на производительность, работоспособность и физическое состояние.

Возможные опасные и вредные факторы представлены в таблице 14.

Таблица 14 – Возможные опасные и вредные факторы

Факторы (ГОСТ 12.0.003-2015)	Этапы работ			Нормативные документы
	Разработка	Изготовление	Эксплуатация	
1.Отклонение показателей микроклимата	+	+	+	**1.** Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов. ГОСТ 12.1.038-82 ССБТ **2.** Правила устройства электроустановок ПУЭ **3.** Гигиенические требования к микроклимату производственных помещений. СанПиН 2.2.4.548-96. **4.** Естественное и искусственное освещение. СП 51.13330.2011 **5.** Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории застройки СН 2.2.4/2.1.8.562-96 **6.** Электромагнитные поля в производственных условиях. СанПиН 2.2.4.1191-03 **7.** Гигиенические требования к персональным электронно-вычислительным машинам и организации работы СанПиН 2.2.2/2.4.1340-03
2. Превышение уровня шума	-	-	-	
3.Отсутствие или недостаток естественного света	+	+	+	
4.Недостаточная освещенность рабочей зоны	+	+	+	
5.Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека	+	+	+	

5.2.1. Анализ выявленных вредных и опасных факторов рабочего помещения

5.2.1.1. Отклонение показателей микроклимата

Микроклимат производственных (рабочих) помещений – климат внутренней среды этих помещений, который определяется действующими на организм человека сочетаниями температуры, влажности и скорости движения воздуха, а также интенсивности теплового излучения от нагретых поверхностей. Влажность воздуха – содержание в воздухе водяного пара. Абсолютная влажность W – масса водяного пара в 1 м3 воздуха. Максимальная влажность F – масса водяного пара, который может насытить 1 м3 воздуха при данной температуре. Относительная влажность R – это отношение абсолютной влажности к максимальной. Указанные параметры – каждый в отдельности и в совокупности – оказывают значительное влияние на работоспособность человека, его самочувствие и здоровье.

Человек постоянно находится в процессе теплового взаимодействия с окружающей его рабочее место средой. Температура, относительная влажность и скорость движения окружающего воздуха характеризуют процесс теплообмена. Данные параметры оказывают комплексное воздействие на процесс теплообмена на рабочем месте.

В соответствии с СанПиН 2.2.2/2.4.1340-03 в производственных помещениях, в которых работа с использованием ПЭВМ является основной и связана с нервно-эмоциональным напряжением, должны обеспечиваться оптимальные параметры микроклимата в соответствии с действующими санитарно-эпидемиологическими нормативами микроклимата производственных помещений.

Исходя из СанПин 2.2.4.548-96 значения температуры, влажности и скорости движения воздуха устанавливаются для рабочей зоны производственных помещений в зависимости от категории тяжести

выполняемой работы, величины избытков явного тепла, выделяемого в помещении, и периода года.

В таблицах 15 и 16 приведены оптимальные и допустимые величины показателей микроклимата на рабочих местах производственных помещений для оператора ЭВМ. В данном случае работа относится к категории труда легкая.

Таблица 15 – Оптимальные величины показателей микроклимата на рабочих местах производственных помещений

Период года	Температура воздуха, C^0	Температура поверхностей, C^0	Относительная влажность воздуха, %	Скорость движения воздуха, м/с
Холодный	21 – 23	20 – 24	40 – 60	0,1
Теплый	23 – 25	22 – 26	40 – 60	0,1

Таблица 16 – Допустимые величины показателей микроклимата на рабочих местах производственных помещений

Период года	Температура воздуха, °С		Температура поверхностей, °С	Относительная влажность воздуха, %	Скорость движения воздуха, м/с	
	Диапазон ниже оптимальных величин	Диапазон выше оптимальных величин			Для диапазона температур воздуха ниже оптимальных величин	Для диапазона температур воздуха выше оптимальных величин
Холодный	19,0 – 20,9	23,1 – 24,0	18,0 – 25,0	15 – 75	0,1	0,2
Теплый	20,0 – 21,9	24,1 – 28,0	19,0 – 29,0	15 – 75	0,1	0,3

5.2.1.2. Превышение уровня шума

Шум – это совокупность различных шумов, возникающих в процессе производства и неблагоприятно воздействующих на организм. Это понятие

обычно рассматривается с точки зрения экологии и медицины, то есть как угрозу жизнедеятельности, а не как фактор, мешающий работе, потому что постоянное его воздействие может принести непоправимый вред здоровью. Традиционно, рабочий шум был постоянной опасностью для работников, занятых в сфере тяжёлой промышленности и ассоциировался только с ухудшением слуха. Современные понятия охраны труда рассматривают шум как угрозу безопасности и здоровью работников многих профессий по различным причинам.

Шум может привести не только к нарушениям слуха (в случае постоянного нахождения при шуме более 85 децибел(dB), но может быть фактором стресса и повысить систолическое кровяное давление.

Дополнительно, он может способствовать несчастным случаям, маскируя предупреждающие сигналы и мешая сконцентрироваться.

В соответствии с СанПин 2.2.2/2.4.1340-03 допустимые значения уровней звукового давления в октавных полосах частот и уровня звука, создаваемого ПЭВМ приведены в таблице 17.

Таблица 17 – Допустимые значения уровней звукового давления, создаваемого ПЭВМ

Уровни звукового давления в октавных полосах со среднегеометрическими частотами									Уровни звука в дБА
31,5 Гц	63 Гц	125 Гц	250 Гц	500 Гц	1000 Гц	2000 Гц	4000 Гц	8000 Гц	
86 дБ	71 дБ	61 дБ	54 дБ	49 дБ	45 дБ	42 дБ	40 дБ	38 дБ	50

Помещения, в которых для работы используются ПК не должны граничить с помещениями, в которых уровни шума превышают нормируемые значения.

В производственных помещениях, оборудованных ПК, при выполнении основной работы на ПК уровень шума на рабочем месте не должен превышать 50 дБА.

Допустимые уровни звука на рабочих местах нормируются по ГОСТ 12.1.003-83 [10]. Значения допустимых уровней шума приведены в таблице 18.

Таблица 18 – Допустимые уровни шума

Объект	Общий уровень звука, дБ	Уровни звукового давления, дБ в среднегеометрических частотах октавных полос, Гц							
		63	125	250	500	1000	2000	4000	8000
Постоянное рабочее место: 1) при воздействии до 4 ч	80	95	87	82	78	75	73	71	69
2) при воздействии до 8 ч	86	101	93	88	81	79	77	75	

При разработке технологических процессов, проектировании, изготовлении и эксплуатации машин, производственных зданий и сооружений, а также при организации рабочего места следует принимать все необходимые меры по снижению шума, воздействующего на человека на рабочих местах:

- Разработкой шумобезопасной техники;
- Применение средств и методов коллективной защиты по ГОСТ 12.1.029-80;
- Применением средств индивидуальной защиты по ГОСТ 12.4.051-78.
- Зоны с уровнем звука или эквивалентным уровнем звука выше 85 дБ А должны быть обозначены знаками безопасности по ГОСТ 12.4.026-76. Работающих в этих зонах администрация обязана снабжать средствами индивидуальной защиты по ГОСТ 12.4.051-78.
- На предприятиях, в организациях и учреждениях должен быть обеспечен контроль уровней шума на рабочих местах не реже одного раза в год.

5.2.1.3. Недостаточная освещенность рабочей зоны

Освещение – получение, распределение и использование световой энергии для обеспечения благоприятных условий видения предметов и объектов. Оно влияет на настроение и самочувствие, определяет эффективность труда. Рациональное освещение помещений и рабочих мест – одно из важнейших условий создания благоприятных и безопасных условий труда. Около 80 % из общего объема информации человек получает через зрительный аппарат. Качество получаемой информации во многом зависит от освещения: неудовлетворительное в количественном или качественном отношении освещение не только утомляет зрение, но и вызывает утомление организма в целом. Нерационально организованное освещение может, кроме того явиться причиной травматизма: плохо освещенные опасные зоны, слепящие источники света и блики от них, резкие тени и пульсации освещенности ухудшают видимость и могут вызвать неадекватное восприятие наблюдаемого объекта. Поэтому рациональное освещение помещений и рабочих мест – одно из важнейших условий для создания благоприятных и безопасных условий труда.

Искусственное освещение предусматривается в помещениях, в которых испытывается недостаток естественного света, а также для освещения помещения в те часы суток, когда естественная освещенность отсутствует. По принципу организации искусственное освещение можно разделить на два вида: общее и комбинированное [11].

Вопрос освещенности рабочих мест, оборудованных персональными компьютерами (ПЭВМ) изложен в СанПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы» [12].

Таблица 19 – Требования к освещению на рабочих местах, оборудованных ПЭВМ

Освещенность на рабочем столе	300-500 лк
Освещенность на экране ПЭВМ	не выше 300лк
Блики на экране	не выше 40 кд/м²
Прямая блесткость источника света	200 кд/м²
Показатель ослепленности	не более 20
Показатель дискомфорта	не более 15
Отношение яркости - между рабочими поверхностями - между поверхностями стен и оборудования	 3:1-5:1 10:1
Коэффициент пульсации	не более 5%.

Освещенность измеряется в Лк – люксах; 1 люмен/м. кв. (люмен Lm – единица величины светового потока). В Европейских нормах освещенности для ряда помещений введен еще один нормируемый параметр: для рабочих мест, оснащенных мониторами устанавливаются требования к максимальной яркости тех поверхностей светильников, которые могут отражаться в экранах.

5.2.1.4. Электробезопасность

Электробезопасность – система организационных и технических мероприятий, и средств, обеспечивающих защиту людей от вредного и опасного для жизни воздействия электрического тока, электрической дуги, электромагнитного поля и статического электричества [13].

Опасное и вредное воздействия на людей электрического тока, электрической дуги и электромагнитных полей проявляются в виде электротравм и профессиональных заболеваний.

Степень опасного и вредного воздействия на человека электрического тока, электрической дуги и электромагнитных полей зависит от:

- Рода и величины напряжения и тока;
- Частоты электрического тока;
- Пути тока через тело человека;
- Продолжительности воздействия электрического тока или электромагнитного поля на организм человека;
- Условий внешней среды.

Помещение офиса по электробезопасности относится к помещению без повышенной опасности, т.е. сухое, хорошо отапливаемое помещение с ток непроводящими полами, с температурой 18-21^{o}C и влажностью 40-50% согласно ГОСТ Р 12.1.019-2009 ССБТ.

Нормы на допустимые токи и напряжения прикосновения в электроустановках должны устанавливаться в соответствии с предельно допустимыми уровнями воздействия на человека токов и напряжений прикосновения и утверждаться в установленном порядке.

Электробезопасность должна обеспечиваться:

- Конструкцией электроустановок;
- Техническими способами и средствами защиты;
- Организационными и техническими мероприятиями.

Электроустановки и их части должны быть выполнены таким образом, чтобы работающие не подвергались опасным и вредным воздействиям электрического тока и электромагнитных полей, и соответствовать требованиям электробезопасности.

Для обеспечения защиты от случайного прикосновения к токоведущим частям необходимо применять следующие способы и средства [13, 14]:

- Защитные оболочки;
- Защитные ограждения (временные или стационарные);
- Безопасное расположение токоведущих частей;

- Изоляция токоведущих частей (рабочая, дополнительная, усиленная, двойная);
- Изоляция рабочего места;
- Малое напряжение;
- Защитное отключение;
- Предупредительная сигнализация, блокировка, знаки безопасности.

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.

Токи статического электричества, наведенные в процессе работы компьютера на корпусах монитора, системного блока и клавиатуры, могут приводить к разрядам при прикосновении к этим элементам. Такие разряды опасности для человека не представляют, но могут привести к выходу из строя компьютера. Для снижения величин токов статического электричества используются нейтрализаторы, местное и общее увлажнение воздуха, использование покрытия полов с антистатической пропиткой.

5.2.2. Обоснование мероприятий по снижению уровней воздействия опасных и вредных факторов на исследователя (работающего)

Профилактика перегрева организма работника в нагревающем микроклимате включает следующие мероприятия:

- нормирование верхней границы внешней термической нагрузки на допустимом уровне применительно к восьмичасовой рабочей смене;
- регламентация продолжительности воздействия нагревающей среды для поддержания среднесменного теплового состояния на оптимальном или допустимом уровне;

- использование специальных средств коллективной и индивидуальной защиты, уменьшающих поступление тепла извне к поверхности тела человека и обеспечивающих допустимый тепловой режим.

Защита от охлаждения осуществляется посредством:

- одежды, изготовленной в соответствии с требованиями государственных стандартов.
- использования локальных источников тепла, обеспечивающие сохранение должного уровня общего и локального теплообмена организма.
- регламентации продолжительности непрерывного пребывания на холоде и продолжительности пребывания в помещении с комфортными условиями.

Мероприятия по защите от повышенного уровня шума:

- устранение причин возникновения шума или снижение его в источнике;
- применение звукоизоляции, звукопоглощения, демпфирования и глушителей шума (активных, резонансных, комбинированных);
- использование средств индивидуальной защиты;
- введение регламентированных дополнительных перерывов;
- проведение обязательных предварительных и периодических медосмотров.

При недостатке на рабочем месте естественного освещения можно выполнить следующие мероприятия:

- защита временем;
- улучшение условий, создаваемых искусственным освещением;
- анализ степени загрязнения стекол в светопроемах, их чистка;
- в случае наличия в помещении зон с достаточным и недостаточным естественным освещением изменение расположения рабочих мест с их перемещением в зону с достаточным естественным освещением.

Средства защиты от повышенного значения напряжения в электрической цепи, замыкание которой может произойти через тело человека:

- оградительные устройства;
- устройства автоматического контроля и сигнализации;
- изолирующие устройства и покрытия;
- устройства защитного заземления и зануления;
- устройства автоматического отключения;
- устройства выравнивания потенциалов и понижения напряжения;
- устройства дистанционного управления;
- предохранительные устройства;
- молниеотводы и разрядники;
- знаки безопасности.

5.3. Экологическая безопасность

В общем случае под охраной окружающей среды характеризуется различного рода мероприятиями влияющие на следующие природные зоны:

- Селитебная зона;
- Атмосфера;
- Гидросфера;
- Литосфера.

Под селитебной зоной понимается территория занятая спортивными сооружениями, зелеными насаждениями, жилыми зданиями и местами отдыха населения, предприняты меры по облагораживанию близлежащих, к области офиса, территорий, их очистка, озеленение, уборка мусора. Устраиваются субботники, которые направлены на очистку и облагораживание территорий, относящихся непосредственно к офису [15].

При размещении зданий, строений, сооружений и иных объектов должно быть обеспечено выполнение требований в области охраны

окружающей среды, восстановления природной среды, рационального использования и воспроизводства природных ресурсов, обеспечения экологической безопасности с учетом ближайших и отдаленных экологических, экономических, демографических и иных последствий эксплуатации указанных объектов и соблюдением приоритета сохранения благоприятной окружающей среды, биологического разнообразия, рационального использования и воспроизводства природных ресурсов.

Анализ воздействия на литосферу сводится к обычному бытовому мусору и отбросам жизнедеятельности человека. В случае выхода из строя ПК, они списываются и отправляются на специальный склад, который при необходимости принимает меры по утилизации списанной техники и комплектующих. На сегодняшний день одним из самых распространенных источников ртутного загрязнения являются вышедшие из эксплуатации люминесцентные лампы. Каждая такая лампа, кроме стекла и алюминия, содержит около 60 мг ртути. Поэтому отслужившие свой срок люминесцентные лампы, а также другие приборы, содержащие ртуть, представляют собой опасный источник токсичных веществ.

В целом, утилизация ламп предполагает передачу использованных ламп предприятиям – переработчикам, которые с помощью специального оборудования перерабатывают вредные лампы в безвредное сырье – сорбент, которое в последующем используют в качестве материала для производства, например, тротуарной плитки.

Под хранением отходов понимается временное размещение их в специально отведенных для этого местах или объектах до их утилизации, или удаления. Отработанные люминесцентные лампы, согласно Классификатору отходов ДК 005-96, утвержденному приказом Госстандарта № 89 от 29.02.96 г., относятся к отходам, которые сортируются и собираются отдельно, поэтому утилизация люминесцентных ламп и их хранение должны отвечать определенные требованиям.

Хранение и удаление отходов (в данном случае - люминесцентных ламп) осуществляются в соответствии с требованиями экологической безопасности согласно СанПин 2.2.7.029-99 наполнения тару с отходами закрывают герметически стальной крышкой, при необходимости заваривают и передают по договору специализированным предприятиям, имеющим лицензию на их утилизацию.

5.4. Безопасность в чрезвычайных ситуациях

5.4.1. Анализ вероятных ЧС

Чрезвычайная ситуация – это состояние, при котором в результате возникновения источника ЧС на объекте, определенной территории или акватории нарушаются нормальные условия жизни и деятельности людей, возникает угроза их жизни и здоровью, наносится ущерб имуществу населения, народному хозяйству и природной среде.

Наиболее типичной ЧС для помещения операторной является пожар. Он может возникнуть вследствие причин электрического и неэлектрического характеров. К причинам электрического характера можно отнести короткое замыкание, искрение, статическое электричество. К причинам неэлектрического характера относится неосторожное обращение с огнём, курение, оставление без присмотра нагревательных приборов.

5.5. Разработка превентивных мер по предупреждению ЧС

Пожарная безопасность – комплекс организационных и технических мероприятий, направленных на обеспечение безопасности людей, на предотвращение пожара, ограничение его распространения, а также на создание условий для успешного тушения пожара [14].

В данном случае на объекте (офис) могут возникать чрезвычайные ситуации (ЧС) техногенного, экологического и природного характера:

- Техногенные;
- Экологические;

- Природные.

Наиболее типичной ЧС для нашего объекта является пожар. Данная ЧС может произойти в случае замыкания электропроводки оборудования, обрыву проводов, не соблюдению мер пожаробезопасности в офисе и т.д.

Согласно техническому регламенту (НПБ 105-03) о требованиях пожарной безопасности по пожарной и взрывопожарной опасности помещения производственного и складского назначения независимо от их функционального назначения подразделяются 5 категорий.

В операторной присутствуют лишь горючие и трудногорючие вещества и материалы (в том числе пыли и волокна), категория производственного помещения – Г (умеренная пожароопасность).

К противопожарным мероприятиям в помещении относят следующие мероприятия:

1. помещение должно быть оборудовано: средствами тушения пожара (огнетушителями, ящиком с песком, стендом с противопожарным инвентарем); средствами связи; должна быть исправна электрическая проводка осветительных приборов и электрооборудования.

2. каждый сотрудник должен знать место нахождения средств пожаротушения и средств связи; помнить номера телефонов для сообщения о пожаре и уметь пользоваться средствами пожаротушения.

Помещение обеспечено средствами пожаротушения в соответствии с нормами:

1. пенный огнетушитель ОП-10 – 1 шт.;
2. углекислотный огнетушитель ОУ-5 – 1 шт.

Вынужденная эвакуация при пожаре протекает в условиях нарастающего действия опасных факторов пожара. Кратковременность процесса вынужденной эвакуации достигается устройством эвакуационных путей и выходов, число, размеры и конструктивно-планировочные решения которых регламентированы строительными нормами СНиП 2.01.02-85.

Для предотвращения возникновения пожара необходимо проводить следующие профилактические работы, направленные на устранение возможных источников возникновения пожара:

- Периодическая проверка проводки;
- Отключение оборудования при покидании рабочего места;
- Проведение с работниками инструктажа по пожарной безопасности.

Для увеличения устойчивости помещения к ЧС необходимо устанавливать системы противопожарной сигнализации, реагирующие на дым и другие продукты горения. Оборудовать помещение огнетушителями, планами эвакуации, а также назначить ответственных за противопожарную безопасность. Согласно НПБ 166-97 необходимо проводить своевременную проверку огнетушителей. Два раза в год (в летний и зимний период) проводить учебные тревоги для отработки действий при пожаре.

5.5.1. Разработка действий в случае возникновения ЧС

Одними из наиболее вероятных видов чрезвычайных ситуаций являются пожар, а также взрыв на рабочем месте.

Всякий работник при обнаружении пожара должен (ППБ 01-03 [13]):

1. незамедлительно сообщить об это в пожарную охрану;
2. принять меры по эвакуации людей, каких-либо материальных ценностей согласно плану эвакуации;
3. отключить электроэнергию, приступить к тушению пожара первичными средствами пожаротушения.

Учебные аудитории 10 корпуса ТПУ оснащены ручными углекислотными огнетушителями ОУ-2 по одному на аудиторию, а также аптечками первой помощи согласно требованиям ГОСТ Р 51057-01.

При возникновении пожара должна сработать система пожаротушения, передав на пункт пожарной станции сигнал о ЧС. В случае если система не сработала, то необходимо самостоятельно произвести вызов

пожарной службы по телефону 101, сообщить место возникновения ЧС и ожидать приезда специалистов.

ЗАКЛЮЧЕНИЕ

В ходе разработки проекта была получена работающая нейронная сеть, способная определить автора текста с точностью 98%.

Была изучена предметная область, методы перевода текста в вектора чисел, язык python. Большая часть времени разработки была потрачена на разработку и тестирование нейронной сети, подбор ее гиперпараметров.

В рамках работы был собран корпус текстов русской литературы 44 авторов, начиная с 19 века, заканчивая настоящим временем. Все тексты были обработаны, написана специальная программа, позволяющая привести их в нужный для анализа вид.

Было проведено сравнение двух метрик, обе из которых показали высокую точность предсказывания, однако метрика word2vec имеет более высокую точность на валидационной выборке (0,98 против 0,86) при меньшем количестве слов во входном файле (3000). Поэтому для данного корпуса текстов метрика word2vec является лучшей.

Данная система может быть встроена в ресурс для определения автора по тексту. Ее пользователи получат возможность с высокой долей вероятности определить автора интересующего их произведения.

СПИСОК ИСПОЛЬЗОВАНЫХ ИСТОЧНИКОВ

1. TF-IDF с примерами кода [Электронный ресурс] // NLPx // URL: http://nlpx.net/archives/57

2. Немного про word2vec: полезная теория [Электронный ресурс] // NLPx // URL: http://nlpx.net/archives/179

3. Основы ИНС [Электронный ресурс] // NeuralNet // URL: https://neuralnet.info/chapter/основы-инс/

4. Как работают нейронные сети: о сложной системе простыми словами [Электронный ресурс] // Robounter // URL: https://robo-hunter.com/news/kak-rabotayt-neironnie-seti-o-slojnoi-sisteme-prostimi-slovami14200.

5. Краткий обзор языка Python [Электронный ресурс] // HelloWorld // URL: http://www.helloworld.ru/texts/comp/lang/python/python2/index.htm

6. Документация [Электронный ресурс] // Морфологический анализатор pymorphy2 // URL: https://pymorphy2.readthedocs.io/en/latest/user/index.html

7. Т.В. Батура. Формальные методы определения авторства текстов. — SSN 1818-7900. Вестник НГУ. Серия Информационные технологии. Том 10, выпуск 4, 2016.

8. Романов А.С. Программная система для идентификации автора письменной речи «Авторовед» / А.С. Романов // Хроники объединенного фонда электронных ресурсов «Наука и образование». - 2015. - №7. - С. 7.

9. Дроздова И. И., Обухова А. Д. Определение авторства текста по частотным характеристикам [Текст] // Технические науки в России и за рубежом: материалы VII Междунар. науч. конф. (г. Москва, ноябрь 2017 г.). — М.: Буки-Веди, 2017. — С. 18-21. — URL https://moluch.ru/conf/tech/archive/286/13237/

10. ГОСТ 12.1.003-83. ССБТ. Общие требования безопасности. – М.: Издательство стандартов, 2002. – 13 с.

11. СНиП 23-05-95. Естественное и искусственное освещение. – М.: Центр проектной продукции в строительстве, 2011. – 70 с.

12. СанПиН 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы. – М.: Информационно-издательский центр Минздрава России, 2003. – 54 с.

13. ГОСТ 12.1.019-79 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты – М.: Издательство стандартов, 1979. – 10 с.

14. СНиП 21-01-97. «Пожарная безопасность зданий и сооружений» [Электронный ресурс] // Библиотека гостов и нормативов «Охрана труда» // URL: https://www.ohranatruda.ru/ot_biblio/normativ/data_normativ/2/2107/

15. ГОСТ 17.4.3.04-85. «Охрана природы. Почвы. Общие требования к контролю и охране от загрязнения».: Сб. ГОСТов. - М.: Стандартинформ, 2008. – 4с.

ПРИЛОЖЕНИЕ А. ЛИСТИНГ ПРОГРАММЫ MAKEFILES

(справочное)

```
  import io
  import time
  import pymorphy2
  import re
  import os
  import codecs

  # Авторы

authorNames = ["Аксенов", "Акунин", "Астафьев", "Белянин",
                 "Булгаков", "Булычев", "Бурносов", "Бушков",
                 "Волоконский", "Дворецкая", "Демина", "Донцова",
                 "Достоевский", "Звездная", "Ишков", "Казанцев",
                 "Каменистый", "Куприн", "Лесков", "Ливадный",
                 "Лукьяненко", "Макс Фрай", "Мамин-Сибиряк",
                 "Маринина", "Мережковский", "Пелевин",
                 "Перумов", "Платонов", "Прилепин", "Пришвин",
                 "Проханов", "Салтыков-Щедрин", "Семенов",
                 "Силлов", "Сорокин", "Стогов", "Стругацкие",
                 "Толстой", "Трифонов", "Тургенев",
                 "Устинова","Чехов", "Эренбург"]

  # Количество слов в новом файле
  wordsAmount = 10000

  bigPath = 'c:/Диплом'

  class Normalizer:
      morph = pymorphy2.MorphAnalyzer()
      regex = re.compile('[^а-я]')

      def norm(self, x):
          return self.morph.parse(x)[0].normal_form

      def prepareText(self, text):
          return self.regex.sub(" ", text.lower()).split()

      def normText(self, oldPath, newPath):
          oldFile = codecs.open(oldPath, "r", "utf-8").read()
          old = self.prepareText(str(oldFile))
          new = []
          for word in old:
              new.append(self.norm(word))
          newFile = open(newPath, "w")
          for word in new:
              newFile.write(" " + word)
          newFile.close()
          return
```

```
  if __name__ == "__main__":

      n = Normalizer()
      start = time.time()
      for authorName in authorNames:
          print(str(int(time.time() - start)) + " - " + "Start "
+ authorName)
          i = 1
          path1 = bigPath + 'Корпуса/' + authorName +
"/Исходники"
          path2 = bigPath + 'Корпуса/' + authorName +
"/Обработано"

          # нормализация слов, папка Исходники
          for file in os.listdir(path1):
              n.normText(path1 + "/" + file, path2 + "/" +
authorName + " " + str(i) + ".txt")
              i += 1

          words = []
          # папка Обработано
          for file in os.listdir(path2):
              # формируем список всех слов
              # открываем документ
              currentDoc = open(path2 + "/" + file, "r").read()

              # разделяем текст на слова
              words += currentDoc.split(" ")

              # количество уже записаных слов в файл
          iter = 0  # 0

          # счетчик названий файлов
          i = 1

          while iter < len(words):
              newPath = bigPath + "Корпуса/" + authorName + "/По
10000 слов/" + authorName + " " + str(i) + ".txt"
              i += 1
              firstWords = []

              firstWords = words[iter:iter + wordsAmount]
              newFile = open(newPath, "w")
              for word in firstWords:
                  newFile.write(word + " ")
              newFile.close()
              iter += wordsAmount

          print(str(int(time.time()/60 - start/60)) + " - " +
authorName + " is DONE!")
      end = time.time()
      print("TOTAL: " + str(int(end/60 - start/60)) + " sec.")
```

ПРИЛОЖЕНИЕ Б. ЛИСТИНГ НЕЙРОСЕТИ

(справочное)

```
import math
from tensorflow.keras.optimizers import SGD
from numpy.random import seed
seed(1)
from tensorflow import set_random_seed
set_random_seed(2)

import os
import numpy as np

from sklearn.feature_extraction.text import TfidfVectorizer
from keras.utils import np_utils
from sklearn.model_selection import train_test_split

from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense

def create_model(i):
    model = Sequential()
    model.add(Dense(75, input_dim=inp, kernel_initializer=i,
activation='relu'))
    model.add(Dense(55, kernel_initializer=i,
activation='relu'))
    model.add(Dense(authorsCount+1, kernel_initializer=i,
activation='softmax'))
    return model

def batch_generator(X, y, batch_size, shuffle):
    number_of_batches = X.shape[0]/batch_size
    counter = 0
    sample_index = np.arange(X.shape[0])
    if shuffle:
        np.random.shuffle(sample_index)
    while True:
        batch_index =
sample_index[batch_size*counter:batch_size*(counter+1)]
        X_batch = X[batch_index,:].todense()
        y_batch = y[batch_index]
        counter += 1
        yield X_batch, y_batch
        if (counter == number_of_batches):
            if shuffle:
                np.random.shuffle(sample_index)
            counter = 0

# Пути и labels
bigPath = 'c:/Диплом/Корпуса'
```

```
authorNames = ["Аксенов", "Акунин", "Астафьев", "Белянин",
                  "Булгаков", "Булычев", "Бурносов", "Бушков",
                  "Волоконский", "Дворецкая", "Демина", "Донцова",
                  "Достоевский", "Звездная", "Ишков", "Казанцев",
                  "Каменистый", "Куприн", "Лесков", "Ливадный",
                  "Лукьяненко", "Макс Фрай", "Мамин-Сибиряк",
                  "Маринина", "Мережковский", "Пелевин",
                  "Перумов", "Платонов", "Прилепин", "Пришвин",
                  "Проханов", "Салтыков-Щедрин", "Семенов",
                  "Силлов", "Сорокин", "Стогов", "Стругацкие",
                  "Толстой", "Трифонов", "Тургенев",
                  "Устинова","Чехов", "Эренбург"]

authorsAmount = len(authorNames)

paths = []
labels = []
wordAmmount = 10000
epoch = 1 #количество эпох обучения

authorsCount = -1
for authorName in authorNames:  # Для каждого автора

    path = bigPath + authorName + "/По " + str(wordAmmount) + "
слов"
    authorsCount += 1

    for file in os.listdir(path):  # Для каждого файла в папке
        currentPath = path + "/" + file
        paths.append(currentPath)  # Записываем путь
        labels.append(authorsCount)  # Записываем в вектор
                                        labels номер автора

print(authorsAmount)

# Построение словаря
vectorizer = TfidfVectorizer(max_df=1.0, min_df=1, norm=None)  #
объявляем векторайзер
allTexts = []  # массив всех текстов

for path in paths:  # Для каждого файла из папок
    currentDoc = open(path, "r")  # открываем документ
    currentText = currentDoc.read()  # читаем документ
    allTexts.append(currentText)  # записываем текст документа в
массив всех текстов
    currentDoc.close()

X = vectorizer.fit_transform(allTexts)
vocab = vectorizer.get_feature_names()  # получаем словарь со
всеми словами

allTexts.clear()  # очистим ненужный больше список
```

```
paths.clear()
inp = int(len(vocab))
print("Всего уникальных стем: ", str(inp))

print(str(int(time.time() - start)/60) + " - " + "Подготавливаем
данные для обучения")

#Подготовка данных
y = np_utils.to_categorical(labels, authorsCount + 1)
#преобразование меток из номера в массив
labels.clear()
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.1, random_state=42)

print(str(int(time.time() - start)/60) + " - " + "Запускаем
нейросеть")

if 1:
    # Описываем и обучаем нейросеть
    uniform_model = create_model('glorot_normal')
    _sgd = SGD(lr=0.005, decay=1e-6, momentum=0.5,
nesterov=True, clipvalue=0.5)
    uniform_model.compile(
        loss='categorical_crossentropy', metrics=['accuracy'],
optimizer=_sgd)

uniform_model.fit_generator(generator=batch_generator(X_train,
y_train, 25, True), epochs=epoch,
steps_per_epoch=math.ceil(X_train.shape[0] / 25), verbose=2,

validation_data=batch_generator(X_test, y_test, 10, True),
validation_steps=math.ceil(X_test.shape[0] / 10))

    # Создаем модель на основе загруженных данных
    uniform_model =
tf.keras.models.model_from_json(loaded_model_json)
    # Загружаем веса в модель
    uniform_model.load_weights(bigPath + "mnist_model.h5")

    #Компилируем модель
    uniform_model.compile(loss="categorical_crossentropy",
optimizer="SGD", metrics=["accuracy"])


# подгружаем валидационную выборку

val_paths = []
val_labels = []

val_authorsCount = -1
```

```
val_files = len(os.listdir(bigPath + "Донцова/По " +
str(wordAmmount) + " слов"))
ammount_val_files = val_files * 0.1

for authorName in authorNames:  # Для каждого автора

    val_path = bigPath + authorName + "/Осталось по " +
str(wordAmmount) + " слов"

    val_authorsCount += 1

    iter = 0
    for file in os.listdir(val_path):  # Для каждого файла в
папке
        while iter < ammount_val_files:
            val_currentPath = val_path + "/" + file
            val_paths.append(val_currentPath)  # Записываем путь
            val_labels.append(val_authorsCount)  # Записываем в
вектор labels номер автора
            iter += 1

# Построение словаря
val_texts = []  # массив всех текстов

for path in val_paths:  # Для каждого файла из папок
    val_currentDoc = open(path, "r")  # открываем документ
    val_currentText = val_currentDoc.read()  # читаем документ
    val_texts.append(val_currentText)  # записываем текст
документа в массив всех текстов
    val_currentDoc.close()

X_val = vectorizer.transform(val_texts)
Y_val = np_utils.to_categorical(val_labels, authorsCount + 1)
#преобразование меток из номера в массив

score=uniform_model.evaluate_generator(generator=batch_generator
(X_val, Y_val, 5, True),verbose=0,steps=math.ceil(X_val.shape[0]
/ 5))

print(str(int(time.time() - start)) + " - " + 'Test accuracy:',
score[1])
```