МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ   
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ   
«ВЯТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт математики и информационных систем

ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ И ФИЗИКО-МАТЕМАТИЧЕСКИХ НАУК

КАФЕДРА ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Допущен к защите

Заведующий кафедрой ПМИ

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Е.В. Разова

**Тематическая классификация коротких текстовых сообщений**

Курсовой проект по дисциплине  
«Проектная и научно-исследовательская деятельность»

Выполнил студент группы ФИб-3301-51-00     / В.Р.Кочкин /

Руководитель К. п. н., зав. кафедрой ПМИ     / Е.В.Разова /

Работа защищена с оценкой     \_\_\_\_.\_\_\_\_.2021 г.

Члены комиссии:     /     /

    /     /

Киров 2021

**Содержание**

[Введение 3](#_Toc102311519)

[1. Машинное обучение и его применение 4](#_Toc102311520)

[2. Обзор методов тематической классификации текстов 6](#_Toc102311521)

[2.1 Модель K-ближайших соседей 6](#_Toc102311522)

[2.2 Модель RandomForest 6](#_Toc102311523)

[2.3 Модель LogisticRegression 6](#_Toc102311524)

[2.4 Выводы по разделу 6](#_Toc102311525)

[3. Подготовка размеченного корпуса 6](#_Toc102311526)

[3.1 Поиск и выбор корпуса данных 7](#_Toc102311527)

[3.2 Предварительная обработка корпуса данных 7](#_Toc102311528)

[3.3 Выводы по разделу 7](#_Toc102311529)

[4. Обучение моделей 7](#_Toc102311530)

[4.1 Обучение модели K-ближайших соседей 7](#_Toc102311531)

[4.2 Обучение модели RandomForest 7](#_Toc102311532)

[4.3 Обучение модели LogisticRegression 7](#_Toc102311533)

[4.4 Выводы по разделу 7](#_Toc102311534)

[5. Сравнение моделей 7](#_Toc102311535)

[5.1 Результаты оценщика 7](#_Toc102311536)

[5.2 Выводы по разделу 7](#_Toc102311537)

[Заключение 8](#_Toc102311538)

[Библиографический список 10](#_Toc102311539)

[Приложения 11](#_Toc102311540)

[Приложение А. Листинг программы 11](#_Toc102311541)

# 

# Введение

С каждым днем в интернете появляется огромное количество текстовой информации, и ориентироваться в ней становится практически невозможно. Миллионы новостей, статей, блогов и других данных ежедневно загружается во всемирную паутину. Появляется потребность в их тематической классификации для упрощения поиска нужной информации в интернете. Поэтому задача классификации текстов стала одной из актуальных тем нашего времени.

На помощь в классификации информации приходит популярное на сегодняшний день направление в анализе данных — машинное обучение. Это наука о том, как заставить искусственный интеллект учиться и действовать как человек, а также сделать так, чтобы он сам постоянно улучшал свое обучение и способности на основе предоставленных нами данных о реальном мире.

Целью работы является обучение качественной модели, способной классифицировать небольшие тексты на русском языке по темам.

Для достижения поставленной цели должны быть достигнуты следующие задачи:

* провести анализ литературы по данной теме;
* определить предметную область текстов;
* подобрать готовый набор данных или написать парсер для его формирования;
* сделать обзор методов тематической классификации текстов;
* построить нейронную модель;
* дать оценку качества построенной модели.

# Машинное обучение и его применение

Машинное обучение – это класс методов [искусственного интеллекта](https://ru.wikipedia.org/wiki/Искусственный_интеллект), характерной чертой которых является не прямое решение задачи, а обучение за счёт применения решений множества сходных задач. Для построения таких методов используются средства [математической статистики](https://ru.wikipedia.org/wiki/Математическая_статистика), [численных методов](https://ru.wikipedia.org/wiki/Численные_методы), [математического анализа](https://ru.wikipedia.org/wiki/Математический_анализ), [методов оптимизации](https://ru.wikipedia.org/wiki/Оптимизация_(математика)), [теории вероятностей](https://ru.wikipedia.org/wiki/Теория_вероятностей), [теории графов](https://ru.wikipedia.org/wiki/Теория_графов), различные техники работы с [данными в цифровой форме](https://ru.wikipedia.org/wiki/Данные_(вычислительная_техника)). Отчасти это похоже на процесс обучения младенца, который учится самостоятельно классифицировать объекты и события, определять взаимосвязи между ними.

В основе машинного обучения лежат три одинаково важных компонента:

* **Данные.** Собираются всевозможными способами. Чем больше данных, тем эффективней машинное обучение и точнее будущий результат.
* **Признаки.** Определяют, на каких параметрах строится машинное обучение.
* **Алгоритм.** Выбор метода машинного обучения (при условии наличия хороших данных) будет влиять на точность, скорость работы и размер готовой модели.

Задачи, которые способно решить машинное обучение, напрямую определяют выгоды для бизнеса и возможности решения социальных проблем государствами разных стран. К основным задачам относятся:

* **Регрессия.** Предоставляет прогноз на основе выборки объектов с различными признаками. По итогам анализа данных на выходе получается число или числовой вектор. Например, таким образом работает кредитный скоринг — оценка кредитоспособности потенциального заёмщика.
* **Классификация.** Выявляет категории объектов на основе имеющихся параметров. Продолжает традиции машинного зрения, поэтому часто можно встретить термин «распознавание образов»: например, идентификация разыскиваемых людей по фото или на основании словесного описания внешности.
* **Кластеризация.** Разделяет данные на схожие категории по объединяющему признаку. Например, космические объекты кластеризируют по удаленности, размерам, типам и другим признакам.
* **Идентификация.** Отделяет данные с заданными параметрами от остального массива данных. К примеру, участвует в постановке медицинского диагноза по набору симптомов.
* **Прогнозирование.** Работает с объемами данных за определенный период и предсказывает на основе анализа их значение через заданный период времени. Примером может служить прогноз погоды.
* **Извлечение знаний.** Исследует зависимости между рядом показателей одного и того же явления или события. Например, находит закономерности во взаимодействии биржевых показателей.

В таблице 1.1 перечислены некоторые популярные области применения машинного обучения.

*Таблица 1.1. Популярные области применения машинного обучения*

|  |  |
| --- | --- |
| Автономные машины  Анализ эмоциональной окраски (например, классификация рецензий на фильмы на отрицательные, положительные и нейтральные)  Выявление аномалий  Выявление закономерностей в данных  Выявление попыток мошенничества с кредитными картами  Выявление попыток страхового мошенничества  Глубокий анализ данных в социальных сетях (Facebook, Twitter, LinkedIn)  Диагностическая медицина  Исследование данных  Классификация новостей: спорт, финансы, политика и т. д.  Классификация электронной почты и выделение спама  Маркетинг: деление клиентов на группы  Обнаружение вторжений в компьютерные системы | Обнаружение объектов в сценах  Перевод естественных языков (с английского на испанский, с французского на японский и т. д.)  Прогнозирование временных рядов — например, предсказание будущих котировок акций и прогнозы погоды  Прогнозирование нарушений выплат ипотечных кредитов  Прогнозирование оттока клиентов  Распознавание голоса  Распознавание лиц  Распознавание образов и классификация изображений  Распознавание рукописного текста  Рекомендательные системы («тем, кто купил этот продукт, также понравились…»)  Сжатие данных  Фильтрация спама  Чат-боты |

# Обзор размеченных корпусов для обучения

Для построения модели с высокой точностью, в первую очередь, необходимы хорошо организованные данные. Поэтому было решено использовать один из размеченных корпусов данных, представленных в большом количестве на сайте kaggle.com. Так как ресурс иностранный, то в большей степени подходящие под тематическую классификацию корпуса были на английском языке, что не удовлетворяет цели работы.

После анализа сайта были выделены следующие корпусы данных на русском языке, которые подходят для обучения на определение тематики текстов. У каждого из них были выделены положительные и отрицательные аспекты, исходя из которых был выбран более подходящий для данной работы корпус.

## Articles from russian site habr.com

Корпус данных, содержащий статьи с русского информационного ресурса habr.com.

Характеристики:

* размер 2,7 Гб;
* состоит из трех файлов:
  + habs.csv

Соотносит статьи к хабам.

Атрибуты: post\_id, hab.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 1.1 – Содержание файла habs.csv

* + posts.csv

Статьи с метаданными.

Атрибуты: post\_id, title, text, date, views\_count, comments\_count, bookmarks\_count, rating, author\_nickname.

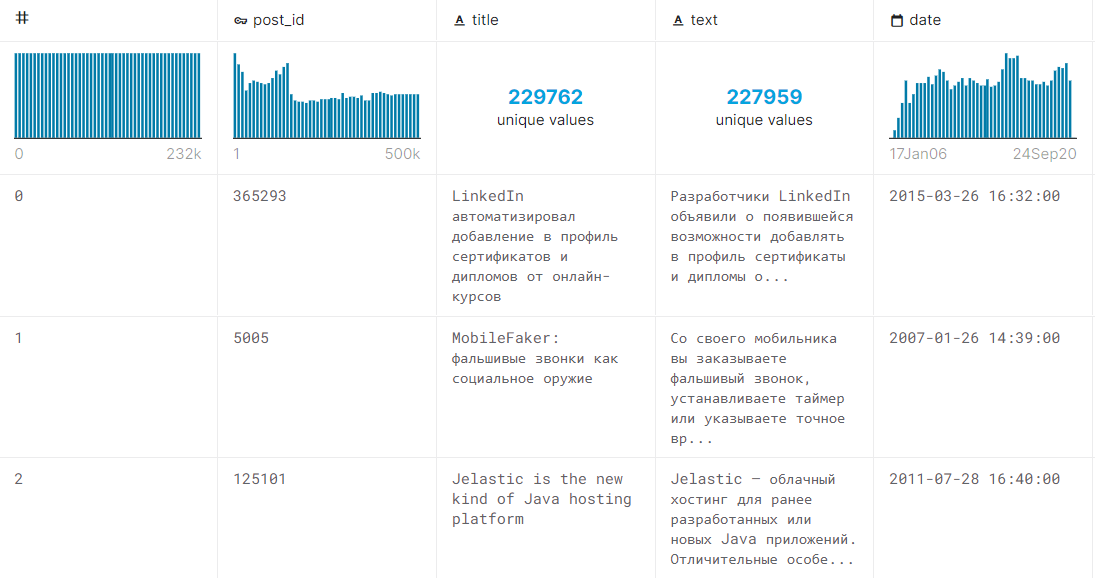


Рисунок 1.2 – Содержание файла posts.csv

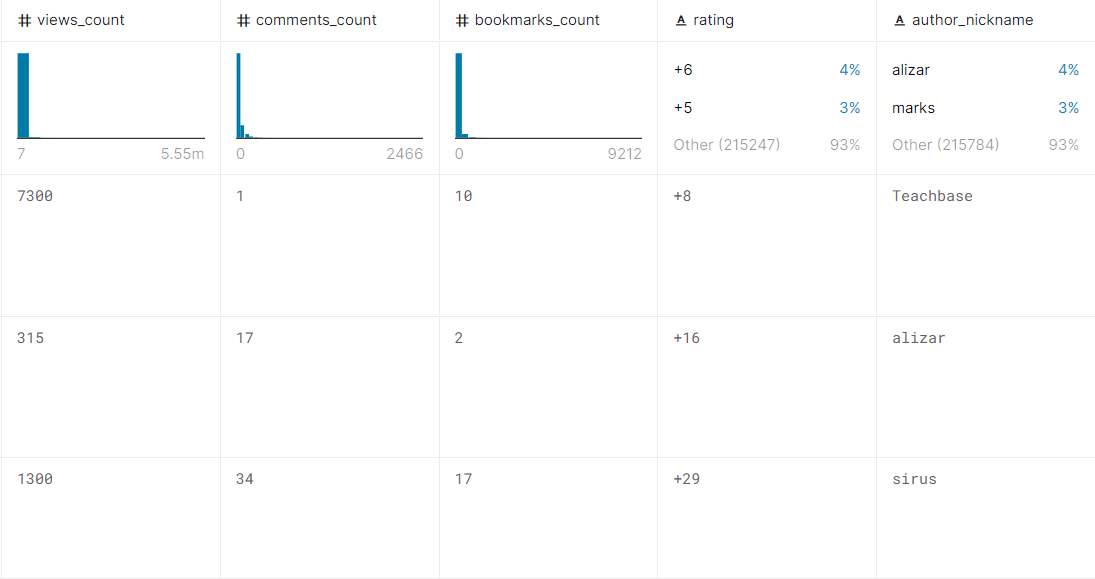


Рисунок 1.3 – Содержание файла posts.csv(продолжение)

* + tags.csv

Соотносит статьи к тегам.

Атрибуты: post\_id, tag.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 1.4 – Содержание файла tags.csv

* количество уникальных текстов 227959.

В качестве атрибута можно взять текст статьи, а в качестве класса – её тег.

Набор данных среднего размера, но с довольно большими текстами (средний размер 7000 символов), что будет сказываться на точности модели. Неудобство использования из-за отделения текстов и тегов.

Решающий недостаток, из-за которого данный набор данных не подходит под задачи проекта, - большинство тегов уникально. То есть сколько текстов, столько и классов. При таком раскладе не получится обучить модель.

## News dataset from Lenta.Ru

Размеченный корпус статей с русского новостного портала Lenta.ru.

Характеристики:

* размер 2,08 Гб;
* состоит из одного файла:
  + lenta-ru-news.csv

Атрибуты: url, title, text, topic, tags, date.

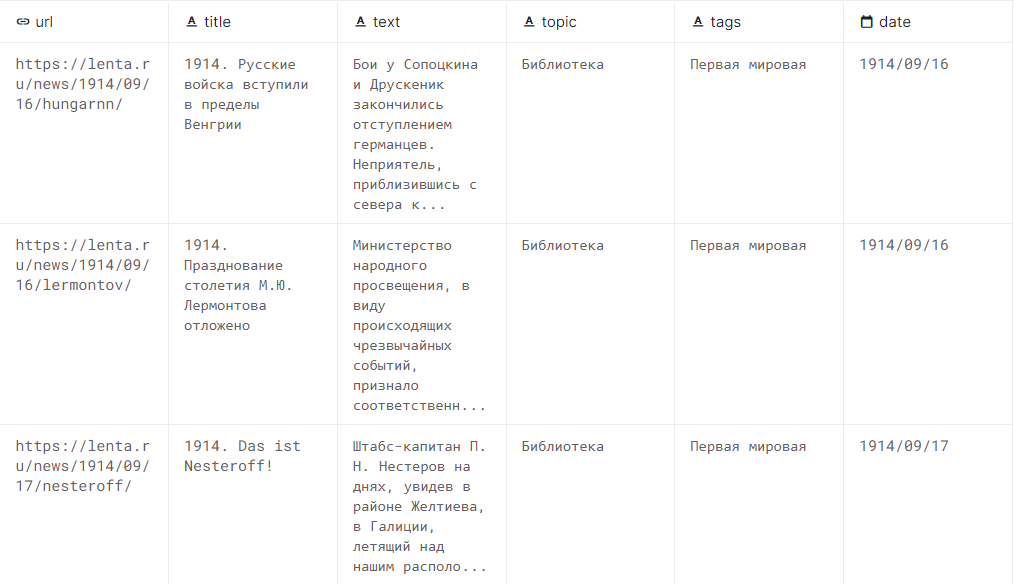


Рисунок 1.5 – Содержание файла lenta-ru-news.csv

* количество уникальных текстов 800038

В качестве атрибута можно использовать текст статьи, а в качестве класса – её топик или тег.

Набор данных довольно большого размера и с небольшими текстами (средний размер N символов). Недостатком является то, что 57% всех тегов являются обобщенными и называются “Все”. Их следует исключить и набора данных. Рассказать, почему следует взять теги, а не топики.

## Выводы по разделу

В этом разделе были рассмотрены готовые размеченные корпуса на русском языке, которые подходят для обучения тематической классификации. В обзор попали два корпуса: статей из habr.com и новостей из lenta.ru. В ходе их анализа было выявлены плюсы и минусы, исходя из которых сделан выбор в пользу корпуса с новостями.

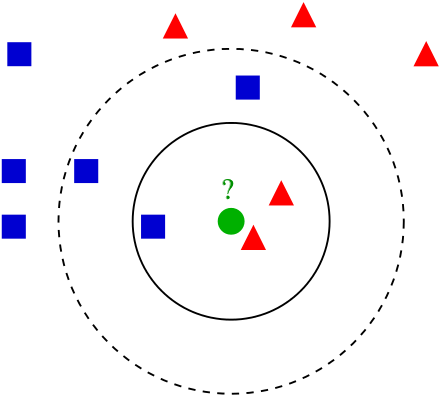
# Обзор алгоритмов тематической классификации текстов

На сегодняшний день существует немалое количество методов (алгоритмов, моделей) машинного обучения. В работе будут рассмотрены и в последующем сравнены 3 одни из самых популярных метода классификации данных.

## Метод K-ближайших соседей

Метод k-ближайших соседей (k Nearest Neighbors, или kNN) – популярный алгоритм классификации, который используется в разных типах задач машинного обучения.

На интуитивном уровне суть метода проста: посмотри на соседей вокруг, какие из них преобладают, таковым ты и являешься. В случае использования метода для классификации объект присваивается тому классу, который является наиболее распространённым среди k соседей данного элемента, классы которых уже известны.

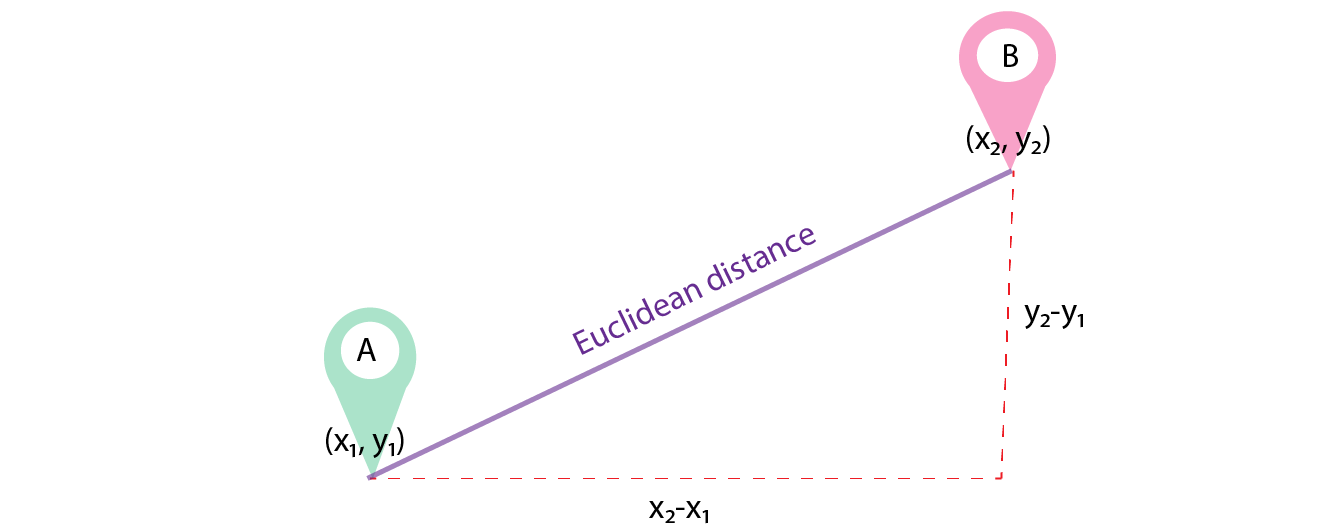


У нас есть тестовый образец в виде зеленого круга. Синие квадраты мы обозначим как класс 1, красные треугольники – класс 2.

Зеленый круг должен быть классифицирован как класс 1 или класс 2. Если рассматриваемая нами область является малым кругом, то объект классифицируется как 2-й класс, потому что внутри данного круга 2 треугольника и только 1 квадрат.

Если мы рассматриваем большой круг (с пунктиром), то круг будет классифицирован как 1-й класс, так как внутри круга 3 квадрата в противовес 2 треугольникам.

Евклидова метрика (евклидово расстояние, или же Euclidean distance) – метрика в евклидовом пространстве, расстояние между двумя точками евклидова пространства, вычисляемое по теореме Пифагора. Проще говоря, это наименьшее возможное расстояние между точками A и B. Хотя евклидово расстояние полезно для малых измерений, оно не работает для больших измерений и для категориальных переменных. Недостатком евклидова расстояния является то, что оно игнорирует сходство между атрибутами. Каждый из них рассматривается как полностью отличный от всех остальных.



Формула вычисления Евклидова расстояния:

Другой важной составляющей метода является нормализация. Разные атрибуты обычно обладают разным диапазоном представленных значений в выборке. К примеру, атрибут А представлен в диапазоне от 0.01 до 0.05, а атрибут Б представлен в диапазоне от 500 до 1000). В таком случае значения дистанции могут сильно зависеть от атрибутов с большими диапазонами. Поэтому данные в большинстве случаев проходят через нормализацию. При кластерном анализе есть два основных способа нормализации данных: MinMax-нормализация и Z-нормализация.

MinMax-нормализация осуществляется следующим образом:

в данном случае все значения будут находиться в диапазоне от 0 до 1; дискретные бинарные значения определяются как 0 и 1.

Z-нормализация:

где σ – среднеквадратичное отклонение. В данном случае большинство значений попадает в диапазон.

Выбор Оптимального значения для k-NN

Не существует конкретного способа определить наилучшее значение для k, поэтому нам нужно попробовать несколько значений, чтобы найти лучшее из них. Но чаще всего наиболее предпочтительным значением для k является 5:

Низкое значение k, например, 1 или 2, может привести к эффекту недообучения модели.

Высокое значение k на первый взгляд выглядит приемлемо, однако возможны трудности с производительностью модели, а также повышается риск переобучения.

Преимущества и Недостатки

Преимущества:

* Алгоритм прост и легко реализуем.
* Не чувствителен к выбросам.
* Нет необходимости строить модель, настраивать несколько параметров или делать дополнительные допущения.
* Алгоритм универсален. Его можно использовать для обоих типов задач: классификации и регрессии.

Недостатки:

* Алгоритм работает значительно медленнее при увеличении объема выборки, предикторов или независимых переменных.
* Из аргумента выше следуют большие вычислительные затраты во время выполнения.
* Всегда нужно определять оптимальное значение k.

<https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19>

## Модель RandomForest

## Модель LogisticRegression

## Выводы по разделу

# Подготовка размеченного корпуса

## Поиск и выбор корпуса данных

## Предварительная обработка корпуса данных

## Выводы по разделу

# Обучение моделей

## Обучение модели K-ближайших соседей

## Обучение модели RandomForest

## Обучение модели LogisticRegression

## Выводы по разделу

# Сравнение моделей

## Результаты оценщика

## Выводы по разделу

# Заключение

В данном курсовом проекте были решены все задачи и достигнута поставленная цель.

Во-первых, был произведен обзор аналогов игры Судоку и выявлен недостающий функционал, а именно отсутствие возможности ввода собственного Судоку и его решения.

Во-вторых, были рассмотрены алгоритм генерации и алгоритмы решения головоломки, а именно Алгоритм перебора с возвратом и его модификация. Они представлены в виде псевдокода и для наглядности продублированы примерами. Данные алгоритмы были протестированы на различных входных данных, благодаря чему сделаны выводы и выбран оптимальный алгоритм для реализации игрового приложения.

В-третьих, было спроектировано и разработано приложение Судоку, которое включает в себя следующие функции:

* Выбор режима игры: генерируемая игра или решение собственного Судоку.
* Выбор размера и сложности игры. Для генерируемой игры доступны размеры 4×4, 9×9, 16×16, 25×25 и сложности легкая, средняя, высокая и эксперт. Для собственной игры доступны размеры 4×4 и 9×9.
* Подсказки для обоих режимов.
* Возможность показать всё решение для обоих режимов.
* Пауза.
* Фоновая музыка, громкостью которой можно управлять из приложения.

В-четвертых, были обработаны все непредвиденные ситуации поведения пользователя при использовании приложения, а также написаны Unit-тесты, которые показывают корректность основных методов программы.

Таким образом, результатом курсового проекта стало игровое приложение Судоку, основанное на рассмотренных алгоритмах, которые были описаны благодаря анализу научной литературы.

# Библиографический список

1. Дейтел Пол, Дейтел Харви. Python: Искусственный интеллект, большие данные и облачные вычисления. — СПб.: Питер, 2020. — 864 с.: ил. — (Серия «Для профессионалов»).
2. Рашид, Тарик. Создаем нейронную сеть. : Пер. с англ. — СПб. : ООО “Альфа-книга”, 2017. — 272 с. : ил. — Парал. тит. англ.
3. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А. А. Слинкина.- М.: ДМАК Пресс, 2015. - 400 с.: ил.
4. Андреас Мюллер, Сара Гвидо. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. Москва 2016-2017.
5. Домингос, Педро. Верховный алгоритм: как машинное обучение изменит наш мир / Педро Домингос ; пер. с англ. В. Горохова ; [науч. ред. А. Сбоев, А. Серенко]. — М. : Манн, Иванов и Фербер, 2016. — 336 с.
6. Шолле Франсуа. Глубокое обучение на Python. — СПб.: Питер, 2018. — 400 с.: ил. — (Серия «Библиотека программиста»).

# Приложения

## Приложение А. Листинг программы