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](#_Toc98962731)

[**1.** **Обзор методов тематической классификации текстов** 4](#_Toc98962732)

[**1.1** **Терминология** 4](#_Toc98962733)

[**Заключение** 5](#_Toc98962734)

[**Библиографический список** 7](#_Toc98962735)

[**Приложения** 8](#_Toc98962736)

[**Приложение А. Листинг программы** 8](#_Toc98962737)

# **Введение**

Актуальность – потребность в классификации новостей, статей или блогов.

Цель работы – обучение качественной модели, способной классифицировать небольшие тексты.

Задачи:

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

# **Машинное обучение и его применение**

Машинное обучение – это специализированный способ, позволяющий обучать компьютеры, не прибегая к программированию. 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 методов тематической классификации текстов**

## **Терминология**

# **Подготовка размеченного корпуса**

Для обучения нейронной модели классифицировать текстовые сообщения был взят готовый корпус блогов с сайта хабр, темами которых выступают теги. Данных было очень много (230к), поэтому был написан скрипт, который конвертирует данные из csv в xlsx, параллельно считая количество символов в тексте блога. Медиана = 7438, следовательно количество текстов, длина которых меньше медианы ~115к. Был написан скрипт, который соотносит строки с текстами и их тегами, чтобы образовалась единая таблица, пригодная для использования в обучении. На это было потрачено много сил и времени, поставьте 4.

# **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.

# **Приложения**

## **Приложение А. Листинг программы**