

АНАЛИЗ ЯЗЫКОВЫХ ГРУПП ДЛЯ ОПТИМИЗАЦИИ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

А. Д. Яблонцева, А. А. Голубничий

Хакасский государственный университет им. Н. Ф. Катанова, пр. Ленина, 90/2, 655017, г. Абакан, Россия, 91919919@mail.ru

В данной статье рассматривается разработка инструмента, предназначенного для анализа языковых групп с целью улучшения алгоритмов машинного обучения. Основная цель исследования заключается в создании программного обеспечения, которое позволит исследователям и разработчикам более эффективно адаптировать алгоритмы под специфические особенности различных языков, что особенно важно в контексте многоязычных моделей обработки естественного языка.

Ключевые слова: машинное обучение, языковые группы, тюркские языки, обработка текста, анализ данных, адаптация алгоритмов.

ANALYSIS OF LANGUAGE GROUPS FOR OPTIMIZATION OF MACHINE LEARNING ALGORITHMS

A. D. Yablonceva, A. A. Golubnichiy

Katanov Khakass State University, ave. Lenin, 90/2, 655017, Abakan, Russia, 91919919@mail.com

This article develops a tool designed to analyze language groups in order to improve machine learning algorithms. The main goal of the research is to create software that will allow researchers and developers to more effectively adapt algorithms to the characteristics of different languages, which is especially important in twenty multilingual natural language processing models.

Keywords: machine learning, language groups, Turkic languages, text processing, data analysis, adaptation of algorithms.

На основе предварительного анализа состояния программных решений и актуальных исследований в области обработки текстов, был предложен метод работы с тюркскими языковыми группами. Одним из ключевых подходов стало использование так называемых «белых списков» для выявления общих черт и различий между языками, что помогает не только лучше понимать особенности языковых групп, но и адаптировать существующие модели машинного обучения к этим специфическим особенностям.

Пример анализа белого списка тюркских языков

Язык	Общие черты с другими	Уникальные особенности	Пример слов
Турецкий	85 %	Грамматика	«ev» (дом)
Казахский	78 %	Фонетика	«үй» (дом)
Узбекский	80 %	Лексика	«чу» (дом)

Для удобства работы с большими объемами данных, разработаны интерфейс, представленный на рисунке, и база данных, обеспечивающие возможность быстрого анализа языковых данных. Важным аспектом разработки является также визуализация полученных результатов, которая позволяет наглядно представлять различия и сходства между языками, что облегчает процесс адаптации алгоритмов машинного обучения.

В конечном итоге разработанное программное обеспечение направлено на улучшение алгоритмов обработки текстов и может быть применимо для работы с различными языковыми семьями, включая тюркские языки. Данный инструмент призван повысить точность и эффективность работы алгоритмов в разнообразных областях применения, от лингвистики до искусственного интеллекта.

Добро пожаловать в оценщик моделей MNIST!

Чтобы использовать эту программу, выполните следующие действия:

1. Нажмите «Загрузить модель», чтобы выбрать файл обученной модели.
2. Нажмите «Загрузить признаки», чтобы выбрать CSV-файл тестовых изображений.
3. Нажмите «Загрузить метки», чтобы выбрать CSV-файл тестовых меток.
4. Нажмите «Оценить модель», чтобы увидеть результаты оценки.
5. Результаты, включая матрицу путаницы и отчет о классификации, будут отображены ниже»

Используемые метрики для оценки алгоритмов машинного обучения

- Точность: доля правильно классифицированных образцов.
- Точность (взвешенная): средневзвешенное значение точности по классу.
- Полнота (взвешенная): средневзвешенное значение полноты по классу.
- Оценка F1 (взвешенная): средневзвешенное значение оценки F1 для каждого класса.
- Оценка F1 (макро): Макро-усредненная F1-мера.
- Оценка F1 (микро): Микро-усредненная F1-мера.

Загрузить модель

Загрузка признаков

Загрузка меток

Оценить модель

Программа оценки эффективности алгоритмов машинного обучения

Библиографический список

1. Алгоритмы распознавания символов. URL: <https://moluch.ru/archive/413/91060/> (дата обращения: 05.05.2024).
2. Алгоритмы синтаксического анализа с использованием кс-грамматик. URL: <https://studfile.net/preview/2817589/page:2/> (дата обращения: 05.05.2024).
3. Как построить контекстно-свободную грамматику по мультиграфу автомата с магазинной памятью. URL: <https://qna.habr.com/q/1023940> (дата обращения: 05.05.2024).
4. Контурный анализ и его применение для распознавания объектов. URL: <https://nauka21veka.ru/articles/tekhnicheskie-nauki/konturnyy-analiz-i-ego-primenenie-dlya-raspoznavaniya-obektov-1480087113/> (дата обращения: 05.05.2024).
5. Расстояние редактирования. URL: <https://stepik.org/lesson/13258/step/2> (дата обращения: 04.06.2024).

© Яблонцева А. Д., Голубничий А. А., 2024