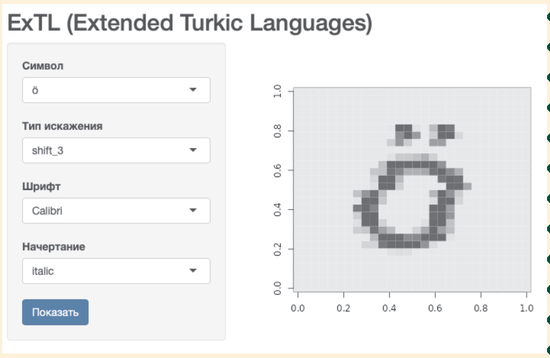
ХАКАССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
им. Н.Ф. Катанова
1939

РАЗРАБОТКА СТАНДАРТИЗИРОВАННОЙ БАЗЫ ДАННЫХ СИМВОЛОВ ДЛЯ ПОСТРОЕНИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ РАСШИРЕННЫХ КИРИЛЛИЧЕСКИХ АЛФАВИТОВ ТЮРКСКИХ ПИСЬМЕННОСТЕЙ

ЦЕЛЬ

Разработка набора датасетов для тренировки моделей машинного обучения по созданию систем оптического распознавания символов тюркских письменностей построенных на основе расширенной кириллицы.

ОПИСАНИЕ РАЗРАБОТКИ



Разработка стандартизированной базы данных для построения моделей машинного обучения на основе кириллических алфавитов тюркских письменностей позволит существенно облегчить процесс их изучения и сохранения, а также способствовать развитию исследований в области тюркологии и лингвистики.

Составлены наборы датасетов EMNIST TCyrillic, содержащие сведения о 74088 изображениях относящихся к 147 символам, из них: 33 основные буквы кириллицы в двух регистрах, 29 букв расширенного кириллического алфавита в двух регистрах, 23 пунктуационных символа. Используется 6 шрифтов в четырех начертаниях без искажений и в 21 варианте искажения изображений. Для проверки результатов было модифицировано 4 нейронные сети для оптического распознавания символов.

Программный продукт уникален, в настоящее время отсутствуют значимые аналогичные информационные системы, позволяющие настраивать модели распознавания символов в создании промышленных систем OCR, что делает программное решение интересным для исследовательских центров и групп разработчиков в области машинного зрения.

ИННОВАЦИОННЫЕ АСПЕКТЫ ПРЕДЛОЖЕНИЯ

Разработана универсальная система распознавания символов языков тюркской семьи с использованием письменности 23 языков, построенных на основе расширенного кириллического алфавита.

ПРАВА ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

- Свидетельство о регистрации базы данных № 2022622859.
- Свидетельство о регистрации программы для ЭВМ № 2022684022.
- Свидетельство о регистрации программы для ЭВМ № 2022685216.

ТЕКУЩАЯ СТАДИЯ

- создана база данных, содержащая различные вариации символов тюркских алфавитов с разными искажениями;
- ведётся разработка программного продукта распознавания символов, тюркских языков.

ПОТЕНЦИАЛЬНЫЕ ПОТРЕБИТЕЛИ

Научно-исследовательские институты, занимающиеся созданием прикладных систем машинного обучения и вопросами развития тюркских языков; туристические компании, люди, интересующиеся культурным наследием.

СРОК РЕАЛИЗАЦИИ: 2021–2024 гг.

ОБЪЕМЫ ВНЕБЮДЖЕТНЫХ СРЕДСТВ ПРИ РЕАЛИЗАЦИИ ПРОЕКТА

300 тыс. руб.

ОБЪЕМ ДОПОЛНИТЕЛЬНЫХ НЕОБХОДИМЫХ ИНВЕСТИЦИЙ

700 тыс. руб.

КОНТАКТЫ

Яблонцева Арина Дмитриевна,
студент
Голубничий Артем Александрович,
ст. преподаватель ПОВТиАС,
ФГБОУ ВО «Хакасский государственный университет им. Н.Ф. Катанова»
тел.:89832580861
e-mail: artem@golubnichij.ru