**Сохранение культуры малочисленных народов РФ с применением IT технологий**

**Егоров Мичил Прокопьевич**

Студент

Университет ИТМО

Факультет прикладной информатики, Санкт-Петербург, Россия

E-mail: [egorov\_michil@mail.ru](mailto:egorov_michil@mail.ru)

В век цифровых технологий, все больше малочисленных народов теряют свой язык, традиции, а самое главное — культурное достояние. Одной из причин, безусловно, является поликультурная среда, в которой проживает молодежь: социальные сети, Интернет. По итогам исследования Германа Греф, пользователи проводят в Интернете от трети до половины дня.

Но пока в Интернете очень мало отсканированных литературных произведений и документов малочисленных народов РФ, не говоря уже об оцифрованных экземплярах — тех, которые могут прочесть поисковые краулеры, тех, которые можно найти по нужным словам. У молодежи доступ к знаниям и творчеству предков ограничен бумажными материалами, большинство из которых уже в наши дни приходит в негодность.

У коренных народов России, которые не имеют статус малочисленных, есть свои особенности: дом книги, где можно прочитать про историю народа, а также, в местных школах преподают собственный язык, а иногда и все предметы на том языке. Всё это благодаря алфавитам, которые бурно образовались в 1920 – 1930 гг.

Проблема уменьшения численности народов заключается не в том, что их репродуктивная способность падает, а в том, что, рождаясь и взрослея, молодое поколение воспитывается на неродном языке. Телевидение, гаджеты преподносят материал ребенку не на родном языке, поэтому ему свой язык уже чужд, странен.

Но тогда возникает вопрос: как у якутов, алфавит которых на основе кириллицы появился только в 1939 г., численность составляет более полумиллиона, то у коренных народов таких как чукчи, эвенки, эвены, алфавит которых появился уже в 1937 г., суммарное количество не составляет и более 100 тысяч?

После проведенных исследований, сделан вывод: якуты начали бурно записывать истории, традиции и мировидение, что позволило в будущем рассказывать об этом потомкам.

\*\*\*

Появление алфавита – некий «бум» в сфере культуры и языка, поэтому логично будет задуматься: когда же произойдет следующий «бум»?

В современном мире всё настолько быстро развивается, что если немного отстать, то догнать прогресс уже будет сложно. Ускорение вычислительных процессов компьютеров сделало возможным то, что считалось фантастикой полвека назад: прогнозы температуры с точностью до 90%, беспилотные автомобили и т.д. Так же и язык не должен отставать.

Сейчас бурно развивается сфера машинного обучения, в частности NLP (Natural Language Processing), где машина обрабатывает естественный язык и понимает смысл человеческой речи или напечатанного им текста. На основе NLP создают чат-ботов, которых трудно отличить от настоящего консультанта, и более продвинутые проекты: Siri, Алиса, а также некоторые рекомендательные системы.

\*\*\*

На основе вышесказанного выдвигается гипотеза: следующий «бум» в сфере языка и культуры – это оцифровка культурного наследия. Все те народы, которые не успеют за этим прогрессом – исчезнут или их численность станет минимальным.

Оцифровка культурного наследия дает возможность потомкам прочитать в Интернете литературу на собственном языке, тем самым содействуя в дальнейшем развитии и распространении их культуры.

Поэтому была поставлена цель - научиться быстро и качественно оцифровывать сначала якутские книги: как рукописные, так и напечатанные, — и в дальнейшем — всех малочисленных народов.

\*\*\*

Для оцифровки будет использоваться технология машинного обучения, поэтому появляются следующие проблемы:

* Недостаток данных. Если нужно научить модель что-то распознавать, то потребуются размеченные данные: картинки цифр, транспорта и т.п. Поэтому нам нужны примеры рукописных и печатных букв. Для решения данной проблемы был запущен сайт, который был написан на языке python3.6 с использованием фреймворка django, где собирались примеры рукописных якутских букв (рис. 1). Было собрано более 500 экземпляров по каждой букве якутского алфавита.

Изображение выглядит как снимок экрана

Автоматически созданное описание

Рисунок 1. Сайт, где можно вписать якутскую букву.

* Если научимся распознавать буквы, то следующим шагом будет распознавание слов, далее - текстов. Для этого используются технологии сегментации текста, где текст сначала разбивается на слова, а потом на буквы.

На данный момент модель обучена на основе библиотеки tensorflow с применением сверточных нейронных сетей, которая определяет какая буква написана, с точностью до 92% (рис. 2). Точность в данном случае вычисляется делением количества правильно распознанных букв на их общее количество.

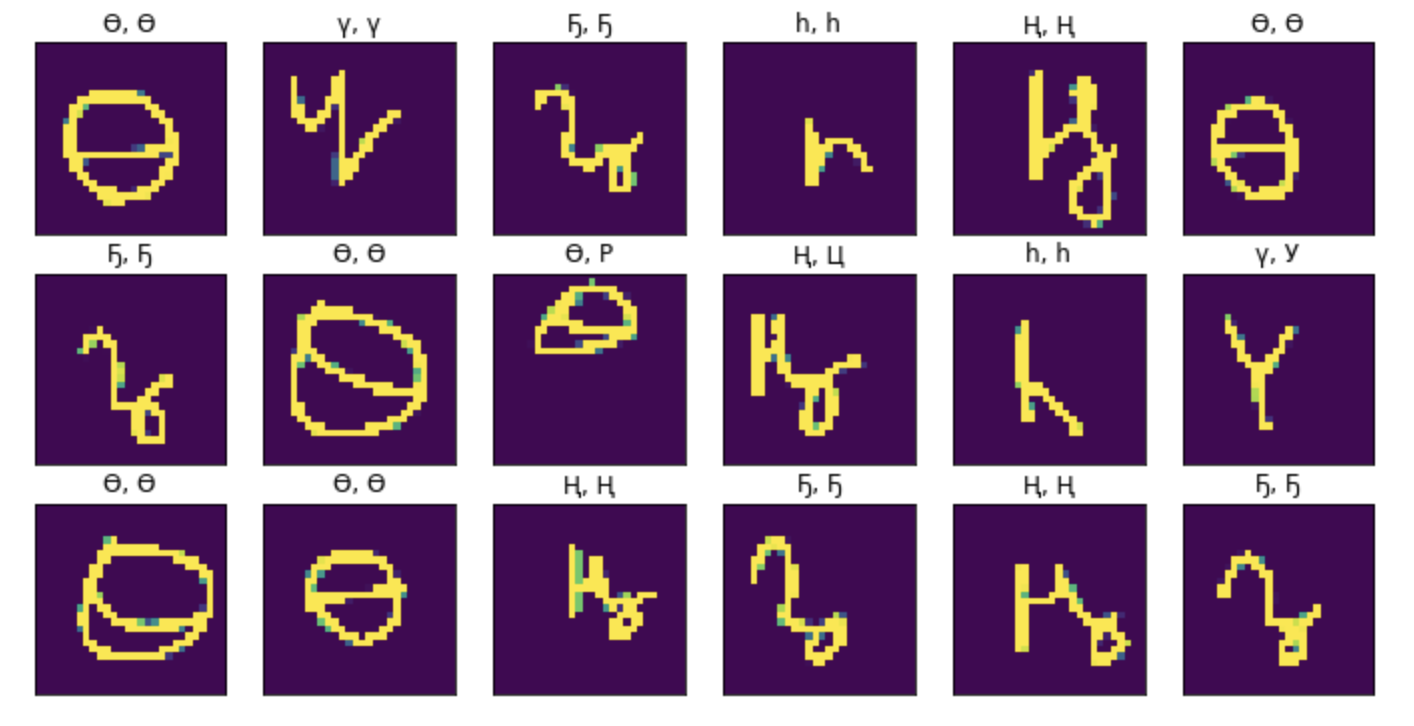


Рисунок 2. Примеры якутских букв из выборки. Первая буква — верный ответ, вторая буква — предсказание модели.

Следующим шагом будет оцифровка якутской сказки – «Таал-Таал эмээхсин», а затем внедрение системы в якутский дом книги – для оцифровки книг, к которым нет доступа из-за того, что они могут развалиться, если обращаться с ними небрежно. А затем в перспективе помочь малочисленным народам выжить в век цифровых технологий.

**Литература**

1. Джесси Рассел. Коренные малочисленные народы Севера, Сибири и Дальнего Востока Российской Федерации. 2012.
2. Nishant Shukla. Machine Learning with TensorFlow. Wis.: Manning Publications. 2018.
3. https://www.[youtu.be/AHCatSM9cow?t=986 (исследование](https://youtu.be/AHCatSM9cow?t=986%20(исследование) Германа Грефа)
4. [https://www.coursera.org/specializations/tensorflow-in-practice (Онлайн](https://www.coursera.org/specializations/tensorflow-in-practice%20(Онлайн)-курс практики по tensorflow)
5. https://www.gks.ru (Федеральная служба государственной статистики России).
6. https://stepik.org/course/1280/syllabus (Онлайн-курс по обработке изображений)