УДК 004.855.5

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ПРИ ГЕНЕРАЦИИ ТЕСТОВЫХ ДАННЫХ ДЛЯ ЗАПОЛНЕНИЯ БД

Рябоус А. Ю. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Научный руководитель – доцент Иванов С.Е.

(федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В докладе приводится процесс рассмотрения и проектирования системы генерации тестовых данных для заполнения баз данных с использованием методов машинного обучения. Предложенная система будет иметь возможность генерировать наборы данных согласно заданной пользователем категории, не ограничивая его только выбором предложенных вариантов.

Введение.

Существующие системы генерации тестовых данных для заполнения баз данных обычно имеют свою, заранее созданную базу данных, из которой псевдослучайным образом производится выборка определённого числа значений, что ограничивает возможности их использования изначально заданными в системе наборами данных. В качестве примера могут быть рассмотрены такие онлайн-сервисы, как http://generatedata.com/#t1 и https://www.mockaroo.com. Они позволяют пользователю задать шаблон генерируемой таблицы и заполнить его данными. Хотя данные сервисы предоставляют возможность выбрать тип и формат генерируемых данных, например, телефонный номер или почтовый адрес, тем не менее, для генерации текстовых данных на выбор предложено лишь ограниченное число системно заданных категорий (имя, фамилия и т.д.).

Основная часть.

В качестве альтернативного решения, предлагается доработать процесс генерации тестовых данных. Часть нетекстовых данных, не требующих осмысленных значений, например, числовые или логические поля, будут генерироваться псевдослучайно. Для полей, подразумевающих текстовые данные, будет реализована система предварительной подготовки данных. Её суть заключается в создании классификатора данных на обширной выборке слов, который, в свою очередь, на основе выделения общих признаков сможет создать группы слов, соответствующих тому или иному признаку. Например, можно использовать для данной цели классификатор дерева решений (или случайного леса для большего объёма данных). Такой подход к генерации текстовых данных позволит значительно увеличить вариативность генерируемых данных.

Выводы.

Применение данного решения обеспечит генерацию данных, наиболее точно соответствующих заданной пользователем категории. Такой подход обеспечит гибкость и большую возможность для генерации различных текстовых данных, что тем самым позволит создавать более реалистичные наборы данных для тестирования.

Рябоус А.Ю. (автор)

Иванов С.Е. (научный руководитель)