УДК: 004.42

МЕТОДЫ И СРЕДСТВА СБОРА, ОБРАБОТКИ И АНАЛИЗА ДАННЫХ ДЛЯ КЛАССИФИКАЦИИ УЧЕТНЫХ ЗАПИСЕЙ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНОЙ СЕТИ INSTAGRAM

Т.В. Назаров
Научный руководитель – М.В Хлопотов
Университет ИТМО

Данная работа рассматривает проблему классификации учетных записей пользователей социальной сети Instagram в соответствии с методологией CRISP-DM (Cross-Industry Standard Process for Data Mining – межотраслевой стандартный процесс для анализа данных). Более половины трафика в сети Интернет генерируется программами, так называемыми «ботами». Гипотеза работы состоит в том, что по определенным признакам учетной записи можно определить, управляется она живым человеком или же ботом. Также рассматривается возможность разделить учетные записи на большее количество более определенных классов.

Целями данной работы являются:

- исследовать методы и средства сбора, обработки и анализа данных социальной сети Instagram;
- исследовать способы и алгоритмы классификации учетных записей пользователей социальной сети Instagram;

Классификация учетных записей в социальных сетях – востребованная задача, как для маркетологов, так и для социологов. Instagram – одна из самых быстрорастущих социальных сетей на сегодняшний день. В условиях закрытия API Instagram наиболее результативным способом сбора данных является парсинг страниц веб-версии Instagram. Для разметки полученных данных на принадлежность тому или иному классу был использован экспертный метод. На специально разработанном ресурсе опрашиваемые отмечали к какому классу, по их мнению, принадлежит та или иная учетная запись. Затем как итоговый выбирался наиболее популярный ответ. Для размеченных данных проводилась оценка работы различных алгоритмов классификации, с целью выявить наиболее подходящий для дальнейшей реализации в рамках программного продукта.

По итогам работы получены результаты:

- разработан модуль для сбора требуемых данных;
- собран и размечен набор данных, характеризующий пользователей социальной сети Instagram;
- исследованы алгоритмы классификации полученного набора данных.

Автор: _____ Назаров Т.В.

Научный руководитель: _____ Хлопотов М.В.

Руководитель образовательной программы: _____ Горлушкина Н.Н.