Н.С. Королёв, Д.С. Кузьменков
Беларусь, Гомель, ГГУ имени Ф. Скорины

КЛАССИФИКАЦИЯ И ОБРАБОТКА НЕСТРУКТУРИРОВАННЫХ ДАННЫХ

В современном мире производители продукции (услуг) для конечного пользователя (B2C сектор) заинтересованы в совершенствовании качества обслуживания. Одним из самых быстрых, современных и действенных способов получить информацию о недостатках (или преимуществах) является сбор и анализ пользовательских данных. Использование анализа данных позволяет производителю реализовать двусторонний механизм обмена информацией и слышать голос пользователей, что позволят позитивно влиять на качество продукта.

Для имплементации этого процесса необходимо собрать данные из публичных источников (и приватных), путем интегрирования решения по обработке в пользовательское решение, надежно сохранить их (с возможностью масштабирования) и произвести непосредственную классификацию.

Сбор данных с публичных источников осуществляется посредством использования API и реализован на примере Twitter. Для приватных данных добавлен способ загружать данные через Excel и через предоставляемый API.

Для хранения данных и последующей работы с ними был выбран свободный движок для поиска Elasticsearch. Elasticsearch – тиражируемая свободная программная поисковая система, основанная на Lucene. Написана на Java, распространяется по лицензии Apache, в основе использует библиотеку Lucene (как и вторая по популярности поисковая система – Solr), официальные клиенты доступны на Java, .NET (C#), Python, Groovy и ряде других языков. Обеспечивает горизонтально масштабируемый поиск, поддерживает многопоточность.

В качестве классификационных правил можно использовать не только ключевые слова, но и другие типы Lucene запросов. Был разработан механизм препарсинга для возможности внедрения собственных имён. Для удобства пользования добавлена функциональность для показа связанных слов и превью для выбранных классификационных правил.

Пользователю предоставлен интерфейс для построения пользовательских моделей классификации с использованием современных средств и просмотра отчетов по ним с использованием HTML 5 и JavaScript (Angular).