Н. С. Королёв, Д. С. Кузьменков
(*УО «ГГУ им. Ф. Скорины», Гомель*)

КЛАССИФИКАЦИЯ И ОБРАБОТКА НЕСТРУКТУРИРОВАННЫХ ДАННЫХ

Производители продукции и услуг зачастую заинтересованы в повышении качества. Одним из самых современных и действенных способов получить информацию о недостатках или преимуществах продукта является сбор и анализ пользовательских данных. Он позволяет производителю реализовать двусторонний механизм обмена информацией и слышать голос пользователей, что позволят позитивно влиять на качество продукта.

Для реализации 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 социальной сети Twitter. Для приватных данных в разработанном приложении 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. Система написана на Java, распространяется по лицензии Apache, в основе использует библиотеку Lucene (также 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. Для приложения классификации и обработки неструктурированных данных также был разработан механизм препарсинга для возможности внедрения собственных. Для удобства использования приложения добавлена функциональность для показа связанных слов и превью для вы-

бранных классификационных правил.