М. П. Шлапак
(*ГрГУ им. Я. Купалы, Гродно*)

СОЗДАНИЕ СИСТЕМЫ КЛАССИФИКАЦИИ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

В настоящее объем информации увеличивается, при этом всё чаще требуется обработка для обнаружения некоторых её характеристик автоматическим образом, в частности классификация, что делает эту задачу актуальной. Использование методов классификации позволяет уменьшить область поиска информации небольшим подмножеством документов, а также имеет прикладное применение в следующих областях обработки информации: фильтрация спама, реклама, системы автоматического перевода текста.

Общая схема задачи выглядит следующим образом:

1. Индексация – преобразование конкретного документа в его логическое представление (например, вектор весов). Процесс индексации можно представить в виде трёх этапов: извлечение термов, взвешивание термов и сокращение размерности.

2. Классификация – этап классификации документа или обучения на множестве преобразованных документов из предыдущего пункта.

Для обучения классификатора были использованы готовые данные из сети, содержащие текстовые документы на русском языке. Для классификации был использован метод MultinominalNB, так как он показал наилучший результат среди других методов классификации на данном наборе данных.

Основной модуль программы реализован с помощью языка Python, т.к. большинство фреймворков для анализа текста были разработаны на этом языке, что расширяет границы различных решений задач классификации. Также написан REST-сервис на Node.js, который агрегирует нужным образом информацию из основного модуля. В качестве базы данных используется MongoDB, клиентская часть написана на TypeScript с использованием фреймворка Angular.

Разработанное приложение позволяет производить классификацию документов, а также REST-сервис и Python-модуль могут быть использованы другими приложениями как вспомогательные сервисы.