СПЕЦИФІКАЦІЯ ВИМОГ ДО ІНФОРМАЦІЙНОЇ СИСТЕМИ ДЛЯ ПРОВЕДЕННЯ АВТОРОЗНАВЧОЇ ЕКСПЕРТИЗИ

Борисова Н.В., Мельник К.В., Слюсарева Ю.В.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Авторознавча експертиза – це один із різновидів лінгвістичної експертизи, що дозволяє провести ідентифікацію особи. Задачі авторознавчої експертизи поділяються на діагностичні, що стосуються визначення особистісних характеристик автора та фактів свідомого спотворення письмової мови, та ідентифікаційні, що стосуються встановлення або перевірки авторства. Задача, що вирішується розроблюваною програмою, належить до діагностичних задач, а саме визначення особистісних характеристик автора, зокрема гендерної приналежності та віку. Для автоматизації визначення зазначених характеристик використовуються такі методи класифікації: наївний Байєсівський, метод опорних векторів, дерева рішень, метод k-найближчих сусідів. При цьому для англійської мови найкращий результат показав метод опорних векторів. Точність визначення гендерної приналежності становила більше 80%, а точність визначення віку – більше 70%. Проте результати аналізу існуючих підходів до вирішення поставленої задачі виявили певні недоліки, основними з яких на наш погляд є: 1) наявність невеликої кількості відкритих програмних реалізацій та веб-сервісів з обмеженим функціоналом; 2) відсутність відкритих програмних реалізацій та/або веб-сервісів для слов’янських мов.

При розробці власної програми для авторознавчої експертизи, перш за все необхідно визначити функціональні та нефункціональні вимоги. Представимо опис функціональних вимог у вигляді user story. Для класифікації текстів було обрано характеристики. Ці характеристики не залежать від контексту та мови, якою написано текст, а також мають лінгвістичну інтерпретацію. Всі використовувані відмінності буде переведено у форму, прийнятну для програмної обробки. Обрані характеристики умовно було поділено на шість груп: 1) дані щодо частоти використання знаків пунктуації та спеціальних символів; 2) дані щодо частоти використання різних частин мови та їх сполучень; 3) дані щодо довжини речень та слів; 4) дані щодо частоти використання мовних зворотів та фразеологізмів; 5) дані щодо частоти використання смайликів; 6) дані щодо словникового запасу. Матеріалом для аналізу слугуватиме корпус записів з блогів. У якості методу аналізу було обрано метод опорних векторів, який показав високу ефективність у вирішенні основного завдання дослідження, відповідно до літературних джерел. Класифікація за гендерною приналежністю здійснюватиметься на два класи: чоловіки та жінки, а класифікація за віком авторів – на 4 класи: не досягли 16, від 16 до 25, від 25 до 44, від 44 і старше. Такий розподіл відповідає Віковій класифікації Всесвітньої організації охорони здоров’я. Стосовно нефункціональних вимог було розроблено вимоги до інтерфейсу, апаратні та програмні вимоги, операційні вимоги.