

ДОСЛІДЖЕННЯ ТЕХНОЛОГІЙ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ЗАПИТАНЬ, ЯКІ ПОВТОРЮЮТЬСЯ

д-р техн. наук, проф. С.Ю. Гавриленко, студ. В.Д. Зозуля, НТУ "ХПІ", м. Харків.

Quora (questions and answers) – це платформа для отримання та обміну знаннями на якій можна ставити запитання та спілкуватися з людьми, які надають унікальні ідеї та якісні відповіді. Це дає людям змогу вчитися один у одного та краще розуміти світ.

Понад 100 мільйонів людей відвідують Quora щомісяця, тому не дивно, що багато людей ставлять подібні запитання. Кілька запитань з однаковою метою можуть змусити шукачів витратити більше часу на пошук найкращої відповіді на своє запитання, а автори відчують, що їм потрібно відповісти на кілька версій одного запитання.

Для виявлення повторюваних запитів використовуються різні методи класифікації текстів. Кожна група методів має свої переваги та недоліки, сфери застосування, особливості та обмеження.

В даній роботі досліджено наступні методи машинного навчання для класифікації текстів: Логістична регресія, К-найближчих сусідів, Гаусовський Бассовський класифікатор, Адаптивний бустінг, Дерева рішень, Випадкові ліси (Random Forest).

Програмні моделі методів розроблені у середовищі Google Colab Python.

Виконано попередню обробку даних, що дало можливість виявити низку ключових закономірностей та інсайдів, які сильно впливають на процес класифікації. Наприклад, наявність спільного першого та останнього слова, більше значення співвідношення кількості унікальних слів до загальної кількості слів мають найбільшу кореляцію із однаковими реченнями.

Дослідження показали, що найбільш високу точність мають методи: Випадкові ліси, Адаптивний бустінг, Гаусовський Бассовський класифікатор та Логістична регресія. Точність даних класифікаторів досягає 99%.

Список літератури: 1. Forczmanski P., Markiewicz A. Two-stage approach to extracting visual objects from paperdocuments. *Machine Vision and Applications*, 2016, Vol. 27, p. 1243-1257. 2. Gantuya P., Mungunshagai B., Suvdaa B. Mongolian Traditional Stamp Recognition using Scalable kNN. *International journal of advanced smart convergence*, 2015, p. 170-176. 3. Agichtein E. Scaling information extraction to large document collections. *IEEE Data Eng. Bull.*, 2005, Vol. 28, p. 4. 4. Radhika Madhavan. Natural Language Processing – Current Applications and Future Possibilities. – Режим доступу [www. URL: https://www.techemergence.com/nlp-currentapplications-and-future-possibilities](https://www.techemergence.com/nlp-currentapplications-and-future-possibilities).