УДК 004.89

Тетяна Грищук, В'ячеслав Ковтун, Володимир Вишневський (Вінниця)

ПРОЦЕС ОЦІНЮВАННЯ ВІДПОВІДЕЙ РЕСПОНДЕНТІВ ЯК ЗАДАЧА АНАЛІЗУ ВЕЛИКИХ ДАНИХ

В наш час все більше організаційних процесів на підприємствах автоматизуються за допомогою прикладних програм. Важко уявити бухгалтерський облік, спостереження за виробничими лініями, документообіг, облік кадрів, що ведуться вручну. Але є досі процес, що в більшості випадків відбувається вручну – процес перевірки кандидатів для зайняття посад на підприємстві. Цей процес складається з декількох задач: публікація оголошень, облік кандидатів, організація інтерв'ю, безпосередньо процес інтерв'ю та прийняття рішення.

Найбільш витратним за часом є етап проведення інтерв'ю. Якщо підготовка питань та/або завдань на інтерв'ю може зайняти небагато часу, то безпосереднє інтерв'ювання може відбирати дні і навіть тижні робочого часу. Ця задача частково може бути розв'язана, якщо усі запитання мають вигляд тестів, але для вибору найкращого кандидата все ж таки бажано отримати відповідь у вільній формі, побачити та почути хоча б декілька відповідей. Але такі види відповідей потребують часу для аналізу.

Метою дослідження є підвищення ефективності процесу оцінювання вільних відповідей респондентів на запитання.

Постановка задачі. Подамо процес оцінювання у вигляді графу

$$E = \{O, C, Q, A, Y\},$$

де O – множина посад; C - множина компаній; Q - множина запитань; A - множина відповідей; Y - множина оцінок за відповіді.

Потрібно побудувати модель, що для заданої пари запитання-відповідь обрахує оцінку якості відповіді в певному діапазоні.

Для **розв'язання задачі** пропонується використати машинне навчання, а саме побудувати регресійну модель. Авторами було проаналізовано потенційні проблеми:

1. Запитання охоплюють різні галузі.
2. Відповіді в базі можуть бути різними мовами.
3. Оцінки за відповіді можуть бути суб'єктивними.
4. Великий обсяг даних.

Проблема великої варіативності може бути вирішена за рахунок великого обсягу даних для навчання. Проблема мовонезалежності може бути розв'язана за рахунок використання мовонезалежних ознак. Одним з найбільш ефективних алгоритмів для векторизації є TF-IDF, де TF - відношення числа входжень обраного слова до загальної кількості слів документа, а IDF - інверсія частоти, з якою слово зустрічається в документах колекції [1]. Використання IDF зменшує вагу широковживаних слів, а TF – визначає найбільш вживані елементи. Проблему суб'єктивності можливо усунути лише частково, очистивши навчальну вибірку від порожніх відповідей, за які респонденти отримали більше 0. Четверта проблема усувається за рахунок використання сучасних бібліотек машинного навчання та мов програмування, таких як Python або Apache Spark.

Для досліджень автори обрали мову програмування Phyton, бібліотеки pandas, TfidfVectorizer та sklearn [2]. Середня абсолютна помилка навчання склала 0.6 для діапазону оцінок від 0 до 10.

Висновки. Розроблена модель дозволяє автоматично оцінювати письмові (або транскрибовані) відповіді респондентів за рахунок вивчення попереднього досвіду. Отримана помилка системи є прийнятною, враховуючи складність поставленої задачі.

Список літературних джерел

1. TF-IDF [сайт]. Режим доступу: https://ru.wikipedia.org/wiki/TF-IDF (дата звернення 09.09.2018 року)
2. Мюллер А. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными / Андреас Мюллер, Сара Гвидо. – Вильямс. – 2016. – 480 с.