УДК 004.42

АВТОМАТИЗАЦІЯ ЗБОРУ ТА АНАЛІЗУ ДАНИХ З ОНЛАЙН-АВТОМАГАЗИНІВ

***А. С. Котляр*[1]**, *Р. В. Пугачов*[2]

[1] *магістрант кафедри інформатики та інтелектуальної власності, НТУ «ХПІ», Харків, Україна*

[2] *доцент кафедри інформатики та інтелектуальної власності, канд. техн. наук, НТУ «ХПІ», Харків, Україна*

artur.kotliar@cs.khpi.edu.ua

Протягом усього життя, пізнаючи навколишній світ, людство постійно має справу зі збором та аналізом інформації. Доречна інформація допомагає правильно оцінити події, що відбуваються, прийняти обдумане рішення, обрати найбільш вдалий варіант своїх дій. З появою мережі Інтернет кількість інформації стала зростати експоненціально, це означає, що потреба в зборі, аналізі та добування корисних даних з цієї інформації також зростає. Через те, що даних стає так багато, обробляти їх в ручному режимі просто не має сенсу – за час такої обробки вони безповоротно застарівають. Тому, зважаючи на неспинне зростання інформації, автоматизація збору та аналізу даних – це актуальна проблема сьогодні і її актуальність у майбутньому буде тільки зростати.

Метою даної роботи є автоматизація збору та аналізу даних з онлайн-автомагазинів. Необхідно реалізувати сервіси збору та очищення даних. Базуючись на очищених даних провести аналіз та візуалізувати окремі результати аналізу.

В цілому, аналіз даних можна розділити на 4 етапи: збір даних, обробка даних, очищення даних, дослідження даних. Збір даних включає в себе ідентифікацію та видобуток неструктурованих даних. Прикладом таких даних є сайт онлайн-магазину. Наступним етапом є обробка зібраних даних. Дані, отримані на попередньому етапі збору, обробляються та організуються для подальшого аналізу. У більш широкому сенсі, на цьому етапі відбувається трансформація неструктурованих даних в структуровані, шляхом розміщення їх у вигляді рядків та стовпців. Після обробки, дані все ще можуть бути неповними, містити дублікати та помилки. Через це виникає потреба в очищенні даних. Очищення даних - це процес запобігання та виправлення помилок в даних. Загальні завдання включають узгодження записів, виявлення неточності даних, загальну якість існуючих даних, дедублікацію. Після очищення дані готові для дослідження. Для дослідження даних може бути створена описова статистика, така як середнє значення або медіана. Візуалізація даних - це також метод, за допомогою якого дані можуть бути досліджені в графічному форматі.

У даній роботі була виконана автоматизація збору та аналізу даних одного з популярних в Україні онлайн-майданчиків для продажів вживаних авто. Наразі на сайті нараховується близько 200 000 активних об’яв. Результати аналізу допомагають зрозуміти вподобання покупців, оцінити ринок та його специфіку, виявити особливості та пропозицію у кожному регіоні.

Збір та обробку даних було виконано за допомогою платформи UiPath технологією веб-скрапінгу (від англ. scraping – «вишкрібання», веб-збирання або витягнення веб-даних) – перетворення у структуровані дані інформації з веб-сторінок, які призначені для перегляду людиною за допомогою браузера. UiPath дозволяє повністю симулювати дії людини, виконуючи такі дії: переміщувати мишу і натискати на кнопки; переміщати файли і директорії; отримувати дані з будь-яких джерел: PDF

файли, картинки, форми; працювати з Word і Excel документами і багато іншого. Такий широкий функціонал дозволив виконати етапи збору та обробку даних в рамках одного UiPath сценарію. Збір даних виконується посторінково, розмір сторінки, кількість сторінок та фільтри вибираються користувачем перед запуском сценарію. Після обробки однієї сторінки структуровані дані зберігаються в CSV (від англ. comma-separated values "значення, розділені комою") файлі.

Для вирішення задачі очищення даних було реалізовано Python скрипт з використанням бібліотеки pandas, яка призначена для маніпулювання даними та їхнього аналізу. Очищення попередньо зібраних даних включає видалення дублікатів, видалення некоректних рядків та нормалізацію даних у кожному стовпці.

У якості інструменту для аналізу та візуалізації зібраних та очищених даних була обрана платформа для бізнес-аналітики Tableau. Tableau дозволяє швидко створювати візуалізації та отримувати корисну інформацію без потреби в написанні складних скриптів для аналізу. Приклад візуалізації 10ти найбільш популярних виробників авто, які продаються, представлений на рис. 1.

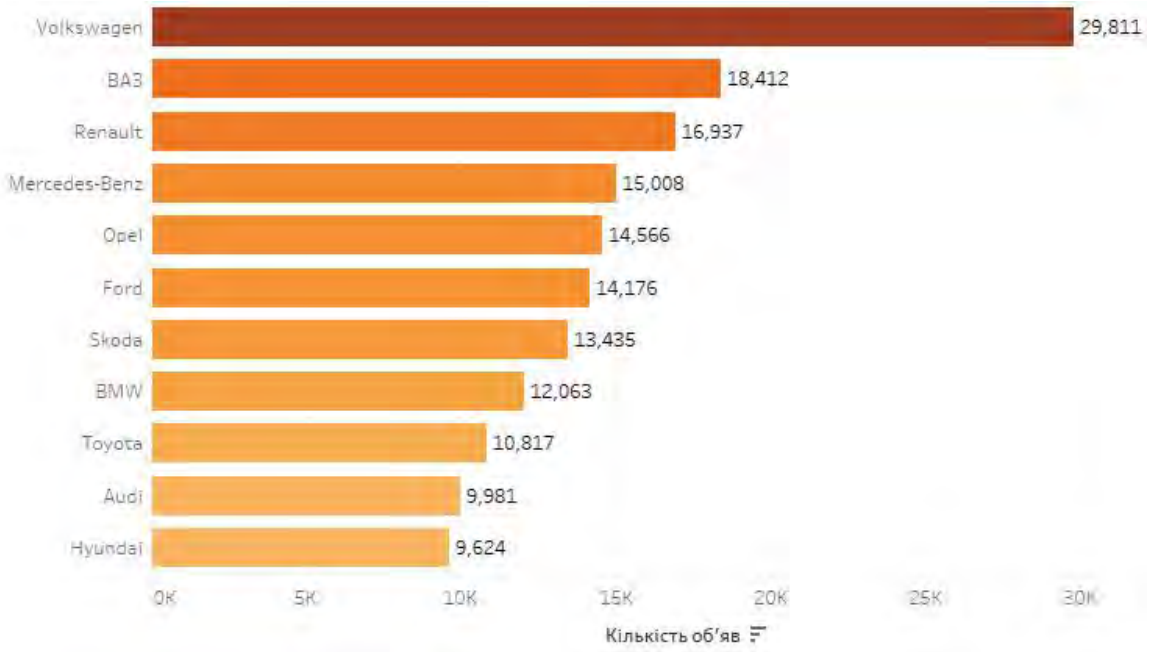

Рис. 1 – Кількість об'яв авто кожного виробника

За допомогою описаного методу було зібрано, оброблено, очищено та проаналізовано близько 260 000 об'яв. Дані зібрані в автоматичному режимі, залучення людини необхідне тільки для налаштування та запуску сервісу. В результаті аналізу були зібрані дані, які можуть бути корисними компаніям виробникам авто, для того, щоб зрозуміти особливості ринку нашої країни та пропонувати більш релевантні позиції та класи автомобілів у майбутньому. Також аналіз може буде корисним і пересічним покупцям, для того, щоб мати максимально повну інформацію про стан ринку та зробити вигідну купівлю.

Список літератури:

1. Data analysis [Електрон. ресурс]. – Режим доступу: https://en.wikipedia.org/wiki/Data_analysis.

2. The UiPath Platform [Електрон. ресурс]. – Режим доступу: https://www.uipath.com/product/platform

3. Tableau Desktop [Електрон. ресурс]. – Режим доступу: https://www.tableau.com/products/desktop