УДК 004.942+574.51 doi:10.20998/2413-4295.2017.53.05

РОЗРОБКА ЕКСПЕРТНОЇ СИСТЕМИ АНАЛІЗУ РЕЗУЛЬТАТІВ ВСТУПНИХ КОМПАНІЙ ДО ВНЗ

А. Т. ПАЗИЛОВА*, О. Г. КАПІТОНОВ, Т. М. ДУБОВИК

кафедра спеціалізованих комп'ютерних систем ДВНЗ УДХТУ,м. Дніпро, УКРАЇНА
**anna.empress@gmail.com*

АНОТАЦІЯ *Аналіз мережі Інтернет на наявність великих обсягів даних про студентів спеціальності "комп'ютерна інженерія". Синтаксичний аналіз (парсинг) сайтів https://abit-poisk.org.ua і https://vk.com, з використанням мови програмування Python, за допомогою фреймворка Scrapy і vk API. Статистичний аналіз отриманих даних за методом найменших квадратів. Побудова моделей машинного навчання за методами KNN і Random Forest.*
Ключові слова: *big data; python; scrapy; machine leatning; парсинг; аналіз великих обсягів даних.*

DEVELOPMENT OF EXPERT SYSTEM OF ANALYSIS OF RESULTS OF ENTERED CAMPAIGNS TO HIGH SCHOOLS

A. PAZYLOVA*, O. KAPITONOV, T. DUBOVYK

Department of specialized computer systems of the SHEI USUCT, Dnipro, UKRAINE

ABSTRACT *Analysis of the Internet for the availability of large volumes of data on students specialty "computer engineering". Parsing (parsing) sites https://abit-poisk.org.ua and https://vk.com, using the Python programming language, using the Scrapy framework and the vk API. Statistical and intellectual analysis of collected data to improve the work of the UDCTU receiving commission, using the least squares method. Construction of models of machine learning using KNN and Random Forest methods.*
Obtained results: the largest mean values of the passing ball before joining is shown by the Department of DNU them. O. Gonchar, this can be explained by more budget places; regarding the maximum and minimum average scores, then the approximate picture for all departments of the city is approximately the same; the left bank of the city of Dnipro is an area that provided the minimum number of students who have passed the competitive selection on a specialty; unlike the previous point, the city center produces a large number of potential specialists.
During the study, the means of collecting information on the Internet, methods for analyzing big data and ways of constructing data models with the help of machine learning and neural networks are analyzed and studied.
On the basis of the collected information, an expert system was created that provides information gathering for entrants on the basis of which statistical and intellectual analysis was conducted.
In the future, you can develop an expert system in several directions: a more detailed analysis of each prospective student, with a psychological portrait on behavior in the Internet;tracking the success of students; drawing up a list of persons for targeting advertising in social networks.
Keywords*: big data; python; scrapy; machine leatning; парсинг; parsing; analysis of big data.*

Вступ

Робота полягає в відповіді на питання: як відібрати саме тих абітурієнтів, які дійсно будуть навчатися якісно і в майбутньому стануть вмілими фахівцями? Чи достатньо, для такого важливого вибору, спиратися тільки на результати зовнішнього незалежного тестування і загальний бал атестата середньої освіти? Можливо варто розширити діапазон оціночних критеріїв, спираючись на досвід попередніх років? Завдяки сучасним технологіям, можна спробувати знайти рішення цієї проблеми.

Для більшої наочності, робота конкретизується на дослідженні студентів напряму «комп'ютерна інженерія» всіх вузів міста Дніпро 2012–2016 років прийому.

Мета роботи

Статистичний і інтелектуальний аналіз великого обсягу даних, зібраних за допомогою парсера в мережі Інтернет, для поліпшення роботи приймальної комісії УДХТУ.

Викладення основного матеріалу

Робота починається зі створення парсера [2]. Постановки задачі:

1. Перейти на сторінку з рейтинговими списками абітурієнтів Дніпра, що знаходиться за адресою https: //abit-poisk.org.ua/rate-review /.

2. Переглянути усі сторінки, які мають в адресі текст «univer».

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

3. Переглянути усі сторінки, які мають в адресі текст «direction».

4. Якщо сторінка містить у хедері сторінки текст «комп'ютерна інженерія» спарсити дані про усіх студентів, що пройшли.

5. Зберегти отримані дані в CSV-файлі.

Розробка моделі збору даних [1].

Модель являє собою окремий клас, який містить перелік атрибутивних полів даних, що збираються. Об'єктом парсинга буде перелік абітурієнтів, які пройшли на бюджетне фінансування з напрямку «комп'ютерна інженерія», а набором атрибутів – їх характеристики. Після попереднього аналізу наповненості сторінок абітурієнтів сайту https://abit-poisk.org.ua/ робимо висновок по потрібним характеристикам; це – ПІБ, результат ЗНО та середній бал атестата. Далі необхідно відобразити їх у спеціальному класі. Для цього відкриваємо файл items.py і описуємо клас. Як можна побачити представлений клас містить записи виду: ім'я атрибута = Field (), де в якості ім'я атрибута використовується його англійський варіант.

Отримані дані були доповненні, з використанням соціальної мережі "Вконтакті". Для цього ми скористалися модулем Python'а для роботи з VK API – це інтерфейс, який дозволяє отримувати інформацію з бази даних vk.com за допомогою http-запитів до спеціального серверу.

Для того, щоб знайти лише студентів вищеназваних університетів, потрібно знайти їх ідентифікатори. Для цього достатньо на сайті виконати пошук за приналежністю до ВУЗа, а потім проаналізувати запит, який знаходиться в посиланні, власне потрібна частина такого виду:

university%5D=16620 // для студентів ДНУ

Далі будемо працювати з методами класу Users VK API, а саме з users.search, який повертає список користувачів відповідно до заданого критерію пошуку. Після отримання інформації про всіх студентів, дані будуть записані в файл data.csv.

По завершенні отримані дані були об'єднанні [9]. (рис. 1).

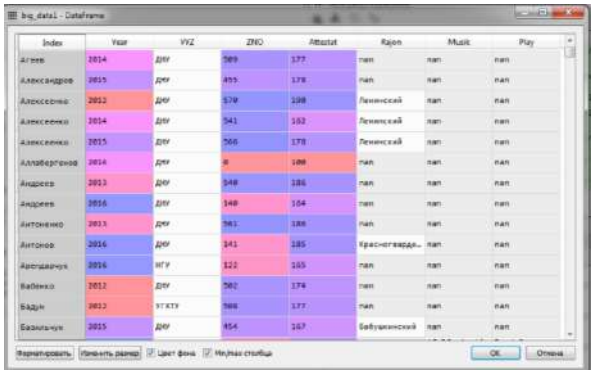

Рис. 1 – Результат об'єднання двох файлів

Як можна побачити, після парсинга залишилося багато не заповнених позицій. Рішення полягає в автоматичному заповненні пропусків [6]. Тому, використовуючи функцію pandas.DataFrame.fillna, з атрибутом method='bfill', пропуски були заповненні значеннями рівними найближчим заповненим даними (рис. 2).

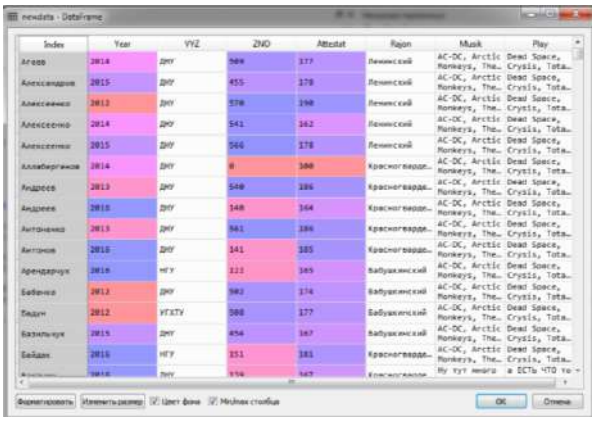

Рис. 2 – Результат після автоматичного заповнення відсутніх даних

Наступний етап передбачав отримання з них загальних характеристик, закономірностей, і їх подальшу інтерпретацію [5].

Розглядалася лінійна модель множинної регресії. Класичний підхід до оцінювання параметрів лінійної моделі множинної регресії заснований на методі найменших квадратів (МНК), який дозволяє отримати такі оцінки параметрів, при яких сума квадратів відхилень фактичних значень результативної ознаки y від розрахункових $\hat{y}$ мінімальна [8].

$$\sum_i (y_i - \hat{y}_{x_i})^2 \rightarrow min$$

Перевірка наявності лінійної кореляції між стовбцями виконувалася за допомогою функції corr(), що розраховує коефіцієнт кореляції Пірсона для усіх пар DataFrame (рис. 3).

Рис. 3 – Кореляційні коефіцієнти Пірсона

Також, оскільки усі дані були приведені до числового вигляду, ми використали функцію scatter_matrix з модуля pandas.tools.plotting, яка дозволила побудувати для кожної кількісної змінної гістограму, а для кожної пари таких змінних – діаграму розсіювання (рис. 4) [7].

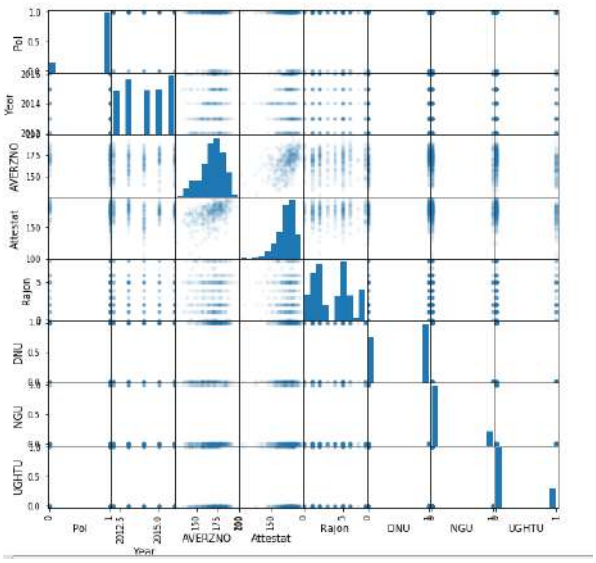

Рис. 4 – Гістограми та діаграми розсіювання

Після проведення аналізу, було вирішено залишити три фактори: середній бал ЗНО, район та стать [3]. Для розрахунку коефіцієнтів по МНК використовували бібліотеку Python – statsmodels, а саме метод OLS (рис. 5).

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Pol	0.4343	0.123	3.541	0.000	0.193	0.675
AVERZNO	0.0075	0.001	9.775	0.000	0.006	0.009
Rajon	0.0068	0.018	0.380	0.704	-0.029	0.042
Omnibus:		260.825	Durbin-Watson:		0.087	
Prob(Omnibus):		0.000	Jarque-Bera (JB):		49.346	
Skew:		0.614	Prob(JB):		1.93e-11	
Kurtosis:		1.800	Cond. No.		468.	

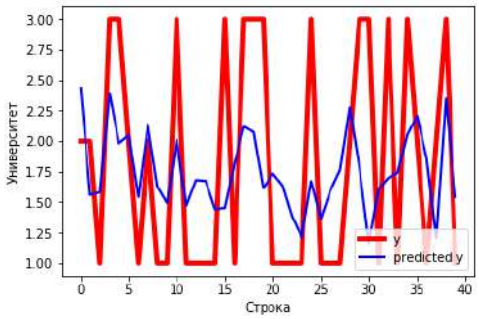

Рис. 5 – Результат праці моделі

Машинне навчання [10].

Для подальшого аналізу даних була використана нейрона мережа, яка мала два схованих нейронних шару, та пройшла навчання за допомогою тестової виборки [4].

Щоб було наочно, моделі було реалізовано за двома алгоритмами: методом K-найближчих сусідів і Random Forest.

Метод найближчих сусідів – метричний класифікатор, заснований на оцінюванні подібності об'єктів. Класифікований об'єкт відноситься до того класу, якому належать найближчі до нього об'єкти навчальної вибірки (рис. 6).

```
Ошибка в методе ближайших соседей:
DNU      0.163701
NGU      0.138790
UGHTU    0.160142
dtype: float64
DNU      0.371901
NGU      0.214876
UGHTU    0.223140
dtype: float64
```

Рис. 6 – Відсоток помилки за методом найближчих сусідів

Random forest, який полягає у використанні комітету (ансамблю) вирішальних дерев, застосовується для задач класифікації, регресії і кластеризації (рис. 7).

```
Ошибка в методе случайный лес:
DNU      0.0
NGU      0.0
UGHTU    0.0
dtype: float64
DNU      0.297521
NGU      0.198347
UGHTU    0.165289
dtype: float64
```

Рис. 7 – Відсоток помилки за методом Random Forest

Обговорення результатів

В ході дослідження були проведені два типи аналізу даних. Отже, статистичний аналіз дає змогу детально розібратися в даних. У випадку з ВНЗ та абітурієнтами, що пройшли за конкурсом по спеціальності «комп'ютерна інженерія», було зроблено наступні висновки:

– найбільші середні значення прохідного балу до вступу показує кафедра ДНУ ім. О. Гончара, це можна пояснити більшою кількістю бюджетних місць;

– щодо максимальних та мінімальних середніх балів, то тут приблизно однакова картина для усіх кафедр міста;

– лівобережжя міста Дніпро є районом, який надав мінімальну кількість студентів, які пройшли конкурсний відбір за фахом;

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

– на відміну від попереднього пункту, центр міста випускає велику кількість потенційних фахівців.

Проте, у теперішній час недостатньо тільки цього. Для прогнозування поведінки математичних моделей використовують штучний інтелект, конкретніше – машинне навчання.

Проаналізувавши результати методу найближчих сусідів, можемо зробити висновок, що для такої задачі метод не дуже адекватний. У другому варіанті – методі Random Forest – відсоток помилок нижчій ніж у попередньому методі, але теж не показує бажаний результат. Із цього можна зробити висновок, що фактори, які навчають модель або занадто щільно розміщені, або їх недостатньо і потрібно їх більш детально вивчити, проте все рівно сучасні програмні засоби дозволяють проводити машинне навчання з досить високим рівнем вірогідності отримання очікуваних результатів.

Висновки

В ході дослідження проаналізовано і вивчено засоби збору інформації в мережі Інтернет, методи аналізу великих обсягів даних і способи побудови моделей даних за допомогою машинного навчання і нейронних мереж.

На основі зібраної інформації створена експертна система, що забезпечує збір інформації про абітурієнтів, на підставі якої проведено статистичний і інтелектуальний аналіз.

У перспективі можна розвивати експертну систему в декількох напрямках:

– більш детальний аналіз кожного перспективного абітурієнта, зі складанням психологічного портрета за поведінкою в мережі Інтеренет;

– відстежування успішності студентів;

– складання списку осіб для таргетинговой реклами в соціальних мережах.

Список літератури

1. Examples of Big Data Projects. URL: http://www.acquia.com/examples-big-data-projects.
2. Feed exports — Scrapy 1.4.0 documentation. URL: https://scrapy.readthedocs.io/en/latest/topics/feed-exports.html/.
3. Habrahabr // Analysis of Big Data URL: http://habrahabr.ru/company/moex/blog/256747/.
4. Habrahabr // Big Data from A to Z. URL: https://habrahabr.ru/company/dca/blog/267361/.
5. **Hulianytskyi, L. F.** Automatic Classification Method Based on a Fuzzy Similarity Relation / **L. F. Hulianytskyi**, **I. I. Riasna** // *Cybernetics and Systems Analysis.* – 2016. – № 1. – P.32-38. – doi:10.1007/sl10559-016-9796-3.
6. **Madaan, A.** Hadoop: Solution to Unstructured Data Handling / **A. Madaan, V. Sharma, P. Pahwa** // *Big Data Analytics. Advances in Intelligent Systems and Computing.* – 2017. – № 1. – P.47-54. – doi: 10.1007/978-981-10-6620-7-6.
7. **Nadich, A.** BigData: problem, technology, market. URL: http://www.compress.ru/artide.aspx?id=22725&iid=1044
8. **Skobelev, V. V.** Analysis of the structure of attributed transition systems without hidden transitions / **V. V. Skobelev** // *Cybernetics and Systems Analysis.* – 2017. – № 2. – P. 165-170. – doi:10.1007/sl10559-017-9916-8.
9. **Артемов, С.** Big Data: новые возможности для растущего бизнеса. // URL: http://www.pcweek.ru/upload/iblock/d05/jet-big-data.pdf
10. **Барсегян, А.** Методы и модели анализа данных: OLAP и Data Mining/, **Барсегян А., М. Куприянов, И. Холод, Е. Степаненко.** - БХВ — Петербург, БХВ — Петербург, 2004. - с. 90.

Bibliography

1. Examples of Big Data Projects. URL: http://www.acquia.com/examples-big-data-projects.
2. Feed exports — Scrapy 1.4.0 documentation. URL: https://scrapy.readthedocs.io/en/latest/topics/feed-exports.html/.
3. Habrahabr // Analysis of Big Data URL: http://habrahabr.ru/company/moex/blog/256747/.
4. Habrahabr // Big Data from A to Z. URL: https://habrahabr.ru/company/dca/blog/267361/.
5. **Hulianytskyi, L. F., Riasna I. I** Automatic Classification Method Based on a Fuzzy Similarity Relation. *Cybernetics and Systems Analysis,* 2016, № **1**, P. 32-38. – doi:10.1007/sl10559-016-9796-3.
6. **Madaan, A., Sharma, V., Pahwa, P.** Hadoop: Solution to Unstructured Data Handling. *Big Data Analytics. Advances in Intelligent Systems and Computing,* 2017, № **1,** P. 47-54. – doi: 10.1007/978-981-10-6620-7-6.
7. **Nadich, A.** BigData: problem, technology, market. URL: http://www.compress.ru/artide.aspx?id=22725&iid=1044
8. **Skobelev, V. V.** Analysis of the structure of attributed transition systems without hidden transitions. *Cybernetics and Systems Analysis,* 2017, № **2,** P. 165-170. – doi:10.1007/sl10559-017-9916-8.
9. **Artemov, C.** Big Data: new updates for business. // URL: http://www.pcweek.ru/upload/iblock/d05/jet-big-data.pdf
10. **Barsegan, A., Kyprinov M., Holod I., Stepanenko E.** Methods and models of data analysis: OLAP and Data Mining**.** - BHV - Petersburg, BHV - Petersburg, 2004. - P. 90.

Відомості про авторів (About authors)

Пазилова Анна Таалайбеківна – студентка магістратури кафедри спеціалізованих комп'ютерних систем ДВНЗ УДХТУ, Дніпро, Україна; e-mail: anna.empress@gmail.com.

Anna Pazylova – Master's degree student in the department of specialized computer systems of the SHEI USUCT, м. Dnipro, Ukraine; e-mail: anna.empress@gmail.com.

Капітонов Олександр Георгійович – доцент кафедри спеціалізованих комп'ютерних систем ДВНЗ УДХТУ, к.х.н., доцент, м. Дніпро, Україна.

Oleksandr Kapitonov – Associate Professor of the Department of Special Computer Systems, SHEI USUCT, Ph.D.(Chemistry), Associate Professor, Dnipro, Ukraine.

Дубовик Тетяна Миколаївна – старший викладач кафедри спеціалізованих комп'ютерних систем ДВНЗ УДХТУ, м. Дніпро, Україна.

Tatyana Dubovyk – senior lecturer in the department of specialized computer systems of the SHEI USUCT, Dnipro, Ukraine.

Будь ласка, посилайтесь на цю статтю наступним чином:

Пазилова, А. Т. Розробка експертної системи аналізу результатів вступних компаній до ВНЗ / **А. Т. Пазилова, О. Г. Капітонов, Т. М. Дубовик** // *Вісник НТУ «ХПІ», Серія: Нові рішення в сучасних технологіях*. – Харків: НТУ «ХПІ». – 2017. – № 53 (1274). – С. 35-39. – doi:10.20998/2413-4295.2017.53.05.

Please cite this article as:

Pazylova, A., Kapitonov, O., Dubovyk, T. Development of expert system of analysis of results of entered campaigns to high schools. *Bulletin of NTU "KhPI". Series: New solutions in modern technologies*. – Kharkiv: NTU "KhPI", 2017, **53** (1274), 35–39 doi:10.20998/2413-4295.2017.53.05.

Пожалуйста, ссылайтесь на эту статью следующим образом:

Пазылова, А. Т. Разработка экспертной системы анализа результатов вступительных компаний в вузы / **А. Т. Пазылова, А. Г. Капитонов, Т. Н. Дубовик** // *Вестник НТУ «ХПИ», Серия: Новые решения в современных технологиях*. – Харьков: НТУ «ХПИ». – 2017. – № 53 (1274). – С. 35-39. – doi:10.20998/2413-4295.2017.53.05.

АННОТАЦИЯ *Анализ сети Интернет на наличие больших объемов данных о студентах специальности "компьютерная инженерия". Синтаксический анализ (парсинг) сайтов https://abit-poisk.org.ua и https://vk.com, с использованием языка программирования Python, с помощью фреймворка Scrapy и vk API. Статистический анализ полученных данных по методу наименьших квадратов. Построение моделей машинного обучения по методам KNN и Random Forest.*

Ключевые слова: *big data; python; scrapy; machine leatning; парсинг; анализ больших объемов данных.*

Поступила (received) 08.12.2017