

УДК 681.004.89:164.053

<https://doi.org/10.35546/kntu2078-4481.2021.1.12>

N.V. KORNILOVSKA

Kherson national technical university

ORCID: 0000-0002-8331-8027

S. V. VYSHEMYRSKA

Kherson national technical university

ORCID: 0000-0002-6343-7512

V.O. ZHOVTONOG

Kherson national technical university

PARSING INTERNET RESOURCES USING A CHAT BOT TO CREATE A CONSOLIDATED INFORMATION RESOURCES IN THE FIELD OF EMPLOYMENT IN THE FIELD OF INFORMATION TECHNOLOGY

Modern information and communication technologies allow modelling and predicting the development of complex global processes and systems, contribute to the improvement of these systems and increase their degree of stability.

The purpose of our research is to create a consolidated information resource in the field of information technology employment.

A consolidated information resource is one of the types of modern information technologies. Data consolidation is the initial stage of implementation of any analytical task or project. Consolidation is based on the process of collecting and organizing data storage in a form that is optimal from the point of view of its processing on a specific analytical platform or solving a specific analytical task. Consolidation is a set of methods and procedures aimed at extracting data from various sources, ensuring the required level of their informative value and quality, conversion into a single format in which they can be loaded into a data storage or analytical system.

To achieve the optimal research result, we will combine methods of information consolidation with such modern information technologies as parsing which will operate with the help of the chatbot created by us. The search chatbot must find the necessary information in a short period of time, qualify it, display it on a computer or mobile phone screen in a user-friendly form, and save the results of the previous search. We think that nowadays parsing technologies and chat bots can be considered as modern procedures for creating the consolidated information resource in the research area. The main functional will be searching for vacancies throughout Ukraine on one of the most popular sites for IT professionals – “<https://djinni.co/>”.

Keywords: Consolidated information resource, parsing, chatbot.

Н.В. КОРНІЛОВСЬКА

Херсонський національний технічний університет

ORCID: 0000-0002-8331-8027

С.В. ВИШЕМИРСЬКА

Херсонський національний технічний університет

ORCID: 0000-0002-6343-7512

В.О. ЖОВТОНОГ

Херсонський національний технічний університет

СИНТАКСИЧНИЙ АНАЛІЗ ІНТЕРНЕТ РЕСУРСІВ ЗА ДОПОМОГОЮ ЧАТ-БОТА ДЛЯ СТВОРЕННЯ КОНСОЛІДОВАНОГО ІНФОРМАЦІЙНОГО РЕСУРСУ В СФЕРІ ПРАЦЕВЛАШТУВАННЯ В ГАЛУЗІ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Нові інформаційно-комунікаційні технології все більшою мірою дозволяють моделювати і прогнозувати розвиток складних глобальних процесів і систем, що сприяє раціоналізації цих систем і підвищення ступеня їх стійкості.

Метою нашого дослідження є створення консолідованого інформаційного ресурсу в галузі працевлаштування в сфері інформаційних технологій.

Консолідований інформаційний ресурс є одним із видів сучасних інформаційних технологій. Консолідація даних є початковим етапом реалізації будь-якої аналітичної задачі або проекту. В основі консолідації лежить процес збору та організації зберігання даних у вигляді, оптимальному з точки зору їх обробки на конкретній аналітичній платформі або вирішення конкретної аналітичної задачі. Консолідація - комплекс методів і процедур, спрямованих на вилучення даних з різних джерел,

забезпечення необхідного рівня їх інформативності та якості, перетворення в єдиний формат, в якому вони можуть бути завантажені в сховище даних або аналітичну систему.

Для досягнення оптимального результату досліджень ми поєднаємо методи консолідації інформації із такими сучасними інформаційними технологіями, як синтаксичний аналізатор (парсинг) який буде працювати за допомогою створеного нами чат бота. Пошуковий чат-бот знайде нам потрібну інформацію за короткий проміжок часу, систематизує її, виведе на екран комп'ютера або мобільного телефону в зручній для нас формі, та буде зберігати результати попереднього пошуку. Ми вважаємо що ці технології, парсинг та чат-боти, на сьогоднішній момент треба вважати надсучасними процедурами створення консолідованого інформаційного ресурсу досліджуваної галузі.

Основним функціоналом буде пошук вакансій по всій Україні на одному з найпопулярніших сайтів для IT-спеціалістів – « <https://djinni.co/> ».

Ключові слова: Консолідований інформаційний ресурс, парсинг, чат-бот.

Н.В. КОРНИЛОВСКАЯ

Херсонский национальный технический университет

ORCID: 0000-0002-8331-8027

С.В.ВИШЕМИРСКАЯ

Херсонский национальный технический университет

ORCID: 000-0002-6343-7512

В.О. ЖОВТОНОГ

Херсонский национальный технический университет

СИНТАКСИЧЕСКИЙ АНАЛИЗ ИНТЕРНЕТ РЕСУРСОВ С ПОМОЩЬЮ ЧАТ-БОТА ДЛЯ СОЗДАНИЯ КОНСОЛИДИРОВАННОЙ ИНФОРМАЦИОННОГО РЕСУРСОВ В СФЕРЕ ТРУДОУСТРОЙСТВА В ОБЛАСТИ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Новые информационно-коммуникационные технологии все в большей степени позволяют моделировать и прогнозировать развитие сложных глобальных процессов и систем, способствуют рационализации этих систем и повышения степени их устойчивости.

Целью нашего исследования является создание консолидированного информационного ресурса в области трудоустройства в сфере информационных технологий.

Консолидированный информационный ресурс является одним из видов современных информационных технологий. Консолидация данных является начальным этапом реализации любой аналитической задачи или проекта. В основе консолидации лежит процесс сбора и организации хранения данных в виде, оптимальном с точки зрения их обработки на конкретном аналитической платформе или решения конкретной аналитической задачи. Консолидация - комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразование в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.

Для достижения оптимального результата исследований мы соединим методы консолидации информации с такими современными информационными технологиями, как синтаксический анализатор (парсинг), который будет работать с помощью созданного нами чат бота. Поисковый чат-бот найдет нам нужную информацию за короткий промежуток времени, систематизирует ее, выведет на экран компьютера или мобильного телефона в удобной для нас форме, и будет сохранять результаты предыдущего поиска. Мы считаем, что эти технологии, парсинг и чат-боты, на сегодняшний момент следует считать самыми процедурами создания консолидированного информационного ресурса исследуемой области.

Основным функционалом будет поиск вакансий по всей Украине на одном из самых популярных сайтов для IT-специалистов - «<https://djinni.co/>».

Ключевые слова: Консолидированный информационный ресурс, парсинг, чат бот.

Problem Statement

Information technology (IT) is a set of methods and tools for searching, collecting, processing, storing, transmitting and protecting information and knowledge to solve management problems on the basis of the advanced software, computer and telecommunications equipment. In modern management, automated IT is increasingly used, i.e. management technologies implemented by using hardware and software. Each of these technologies is aimed at implementing a making management decisions mechanism necessary to achieve optimal market characteristics of the management object [1]. Currently there is a global transition to the information

society, which development is inseparably linked with the intensification of information processes, the need to collect, process and transmit huge amounts of information.

The main goals of informatization should flow from the general strategy of economic reforms in Ukraine and be determined by the needs of intensive development of the national information infrastructure, which should cover various social spheres of life and activity. Therefore, the development of informatization should be aimed at achieving the following goals: 1) significant increasing in the level of information completeness, relevance and accessibility for users; 2) qualitative improving information and analytical support of the operation of public administration systems; 3) improving the system of information support for economic entities of all forms of ownership; 4) widespread using potential IT capabilities to solve socio-economic problems, 5) revitalizing in the system of international information exchange in the interests of the political, economic, social and humanitarian ties development. In other words, useful and available information is often purposeless because it is simply not presented with such content and in a form that corresponds to this group (or level) of users.

Consolidation of information is put forward as a solution to problems caused by lack of relevant information. Consolidated information in fact is associated with the increased use of information by different groups of users [2]. However, a consolidated information resource becomes a task that itself requires significant skills, efforts and resources. Efforts and resources will be needed for full implementing the information consolidation unit, establishing the necessary cooperation between applied and IT professionals, teaching and training professionals who possess skills in obtaining information consolidation products and providing relevant services. Consolidation of information is not a panacea for solving information problems and needs, but it is becoming one of the important approaches that should be considered along with a huge variety of other information products and services. Consolidated information products and services play an important role in meeting many critical information needs of modern Ukrainian society [3].

Analysis of the latest researches and publications

The National Informatization Program defines the strategy for solving the problem of meeting information needs and informational support of socio-economic, environmental, scientific and technical, defense, national-cultural and other activities in areas of national importance.

New IT, based on computer technology, require radical changes in organizational management structures, its regulations, HR potential, documentation system, recording and transmitting information. The introduction of information management is of particular importance, it significantly expands the possibilities of using information resources [4]. The development of information management is associated with the organization of knowledge and data processing system, its consistent development to the level of integrated automated control systems that cover vertically and horizontally all levels and links of the organization.

“Bot” (short for “robot”) is a program that imitates human activity. A chatbot accordingly imitates an interlocutor in the chat. A bot is a computer program that tries to give the impression that it is not a program, but a real person who sits somewhere in the Internet and has own opinion and intelligence buds. Simply put, a bot is a robot that has been given a laptop, got an account, and has been put to communicate with people who believe he is a human too. Today, chatbots are capable and are already replacing the support services of various facilities in different areas of life.

The main difference of this program is active human participating at all stages of its development. Starting from its development and ending with its use and simultaneous training using neural networks. The spread and use of chatbots has led to the UX-paradigm of interaction “messaging-as-an-interface” [5].

The mobile applications market is oversaturated: they have been numbered in millions, but the users no longer want to install something new. According to ComScore research, users spend 80% of their time in only three applications. On this background, the messenger segment is growing rapidly. Last year the total audience of the most popular messengers overtook the most popular social networks. Privacy messengers require fewer resources, work on cheaper devices and unlike social networks are not yet clogged with unnecessary information, intrusive advertising and other people’s news.

As a result, it became obvious to service developers that it is easier to get to the user in the program that he has already installed and opens every day than to convince him of the need to work with a particular application. The chatbot does not require traffic for download, installation time, does not take up space in memory and on a smartphone screen. It is easy to work with a chatbot: it is necessary to add it to your contact list and start correspondence. Most often in reply the bot will send information about itself, a list of available commands or display buttons that can turn a dialog box into an intuitive mini-application.

On one hand, using chatbots is useful for customers because they can get the necessary information or take any action in a simpler and more convenient way, and on the other hand, it is an advantage to companies that can use chatbots in order to promote their brand. In addition, the active development of artificial intelligence and speech processing technologies leads to certain innovations: the network can change significantly if chatbots can learn to do something that sites and applications cannot do yet.

Parsing is the process of comparing the linear sequence of language tokens (words, phrases) with its formal grammar. The result of such an analysis is a parsing tree (syntactic tree). It is usually used together with lexical analysis. A parser is a program or part of a program that performs parsing, that is, recognizing incoming information. Herewith the incoming data is converted to a form suitable for further processing. This type usually represents a formal model of incoming information in further information processing language [5]. When parsing, the source text is converted into a data structure, usually into a tree that reflects the syntactic structure of the incoming sequence and is suitable for further processing. As a rule, the result of parsing is the syntactic structure of the sentence presented in the form of a tree of dependencies, or in the form of a tree of components, or in the form of some combination of the first and second presentation ways.

Any information that has “syntax” is automatically analyzed:

- programming languages are parsing the programming languages source code in the process of translation (compilation or interpretation);
- structured data, languages of their description, design, etc., for example, XML, HTML, CSS, ini-files, specialized configuration files.
- SQL-queries (DSL-language), mathematical expressions and regular expressions (which in their turn can be used to make a lexical analysis automatic);
- formal grammars, linguistics are human languages, for example machine translation and other text generators.

Parsing is the process of converting a source code into a structured form. A typical parser is a combination of a lexer and a parser. The lexer groups the source code characters into meaningful sequences called tokens. Then the type of token (identifier, number, string, etc.) is determined. A token is a set of a token meaning and its type. In the example in Figure 1, the tokens are `sp = 100`. The parser builds a coherent tree-like structure from a stream of tokens, which is called a parsing tree. In this case, `assign` is one of tree nodes. Abstract syntax tree or AST is a parsing tree at a higher level, from which significant markers such as parentheses, commas are not removed. However, there are parsers in which the step of lexing and parsing are combined [5].

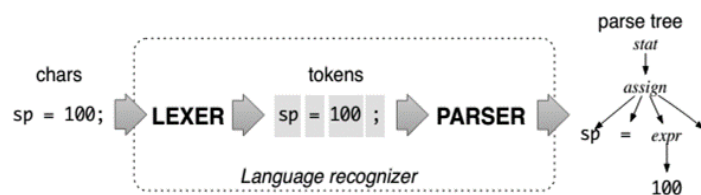


Fig.1. The principle of building a parsing tree by a parser from a stream of the tokens of the connected tree-like structure

The rules are used to describe different AST nodes. The combination of all the rules is called the language grammar. There are tools that generate a code for a certain platform (runtime) for parsing languages on the basis of grammars. They are called parser generators. For example, ANTLR, Bison, Coco / R. However, a parser is often written manually for some reasons, such as Roslyn for example. The advantages of the manual approach are that parsers tend to be more productive and readable [6].

Goal Setting

The aim of our research has become to combine with the maximum efficiency the tools of the consolidated information resource, information analysis (parsing) with the creation of a search chatbot. To implement the set task, the Python language has been selected. It is necessary to create a chatbot that will search and qualify the necessary information in a short period of time, display it on a computer or mobile phone screen in a user-friendly form, and save the results of the previous search. Namely it is required to get the functional that will search for vacancies throughout Ukraine on one of the most popular sites for IT specialists – “https://djinni.co/”.

Presentation of research material

As noted above, a consolidated information resource is a set of methods and procedures aimed at extracting data from various sources, ensuring the required level of its informative value and quality, conversion into a single format in which they can be downloaded into a data storage or analytical system. In this research we will use information parsing technology to extract data from various sources, and we will use search chatbot technology to convert data into a single format. In a simplified form, the chatbot technology looks like this: there is an “engine” recognizing and maintaining a dialogic component. It operates a knowledge base which contains the rules for recognizing questions and creating answers to them. It is possible to connect a messenger, mobile application, website or terminals to this “engine”. If it is needed it is possible to connect technologies for processing incoming voice messages.

We have identified several steps to create a search chatbot:

- 1) preparation;

- 2) creation of a chatbot account and obtaining data for management through the API;
- 3) backend development;
- 4) approval (publication) of the bot.

A more detailed algorithm for creating a chatbot is given in Fig. 2.

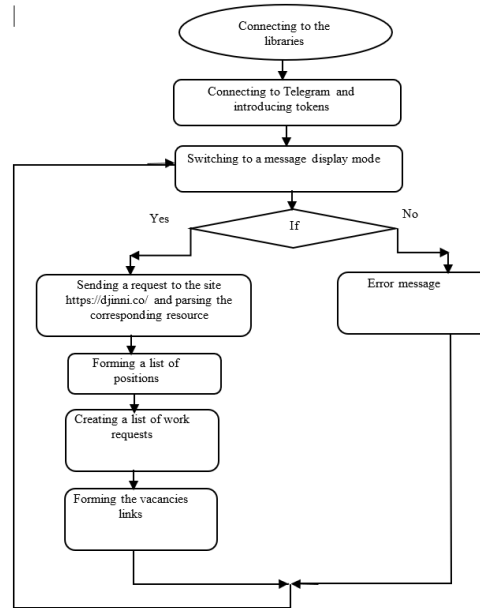


Fig. 2. A detailed algorithm for creating a chatbot.

We will consider these stages in detail.

Preparation. It is required to study the documentation of the platform in order to select a messaging program (platform) in which the bot will interact with users. This is necessary in order to understand what and how the bot will be able to do. It is also necessary to consider in detail the chatbot operation algorithm (Fig. 3).

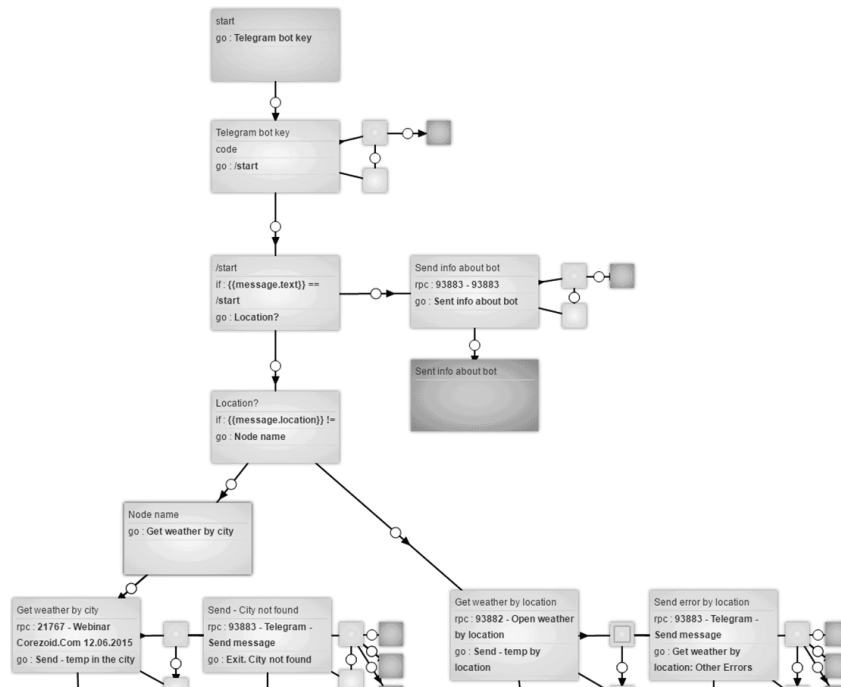


Fig. 3. Example of a chatbot algorithm on the Telegram platform

Creation of a chatbot account and obtaining data for management through the API. In Facebook Messenger a new application is added to the developer’s account on developers.facebook.com, in Viber a bot account is created on partners.viber.com. In Telegram, the chatbot is registered via the @BotFather with the /newbot command, and in Kik with the help of the Botsworth bot. In Skype we create a new bot in the section “My bots” on the site dev.botframework.com. It is necessary to fill in the form and click “Create Microsoft App ID and password”.

We have analysed modern messengers – platforms on which chatbots work in order to choose the working area in which the created chatbot will work. Comparing the Telegram platform with Slack, Skype, Viber, we can highlight the following distinctive features in favour of Telegram:

- 1) cloud-based storage of all correspondence data;
- 2) there is a two-factor user authentication, which makes the use of this platform more secure;
- 3) it has its own encrypted cloud-based storage, which is distributed in different jurisdictions, and it is much better protected than the Google and Apple storages;
- 4) it allows the users to access chats from multiple devices at the same time, thanks to cloud synchronization. Thanks to this, users of Mac, PC, iPad and even of Linux servers get the same experience of interacting with the messenger.

Telegram chatbots have a very available API, the use of which is free and anyone can create their own chatbot. This means that anyone can see how everything works and use the data to develop their own projects. Telegram has neither a publishing procedure nor a chatbot test mode. The bot is immediately available for all users.

Linking a webhook. It is needed to install a webhook on your web server. Webhook is a proper script that is signed to events hosted on the server, and accepts all reports of bot operations (incoming messages, message delivery reports, button clicks, reports about perusal, etc.). In the documentation for all key messengers it is described in details how to link a webhook. This step can be organized in different ways. For example, in Facebook Messenger this is done through the developer's account at developers.facebook.com/apps, and in Viber we use a request to the Viber API, the same in Telegram. A prerequisite for the messenger to work with webhook is to possess an SSL certificate on the server where it resides. None of the key messengers works without a secure certificate [7].

Receiving a token. It is used when requesting the API messenger.

Backend development. In most cases, Node.js or PHP is used to develop bots, but Java or Python also have libraries to accomplish this task.

✓ The project you are creating is a script that will control the chatbot. When creating an account, we link the handler's address (webhook) with the account and subscribe to events in Facebook Messenger, Viber. Other messaging apps automatically subscribe to all events.

✓ Events come to the webhook POST as a request in Json format. This Json stores all data about the current event. For example, a bot has received a text message from a user – Json contains the time of the message receipt, its text, user ID, and so on.

✓ The task of the script is to process this Json and select the answer to the user. When the script has determined the response for the user (picked up the content to be sent in response), a request is sent to the API platform – usually also a POST request. The request specifies the API access key, user ID and the content is sent. The structure of this Json depends on the messenger and the type of message.

Approval (publication) of the bot. Telegram has neither a publishing procedure nor a chatbot test mode. The bot is immediately available for all users.

The main library for creating a bot is Telebot. This library is official for Python from the Telegram developers. The developer's website contains the official documentation on how to use this library. With the help of this library there is a connection between the created program and Telegram.

It is necessary to get a bot key in order to keep in touch. This key can be obtained from an official bot called BotFather. In this bot, it is needed to type the command "/newbot", then type its name and login. After this procedure, we will receive the key that is responsible for the bot operations. Obtaining job information is used by parsing information from official sites, as they are not adapted to sending data on the third-party resources.

Parsing-robots work thanks to the bs4 library (Beautiful Soup 4). The bs4 library is used to store a copy of the site in HTML format in RAM and to structure the information for further search [7]. After saving a copy of the site, we can search for the information we need among HTML tags and classes. In order to make a job search function work, it is required to "extract" all possible professions, job offer titles and links to these offers from the site, each information is contained in the tags "<div class = \"jobs\"> </div>", "<div class = \"vacancy\"> </div>", " " accordingly.

The urllib library is used to process URLs (Uniform Resource Locator). The request is generated and sent using the library requests. This library exactly connects to the desired site, receives a response and generates logs for further decryption.

Decryption is performed using the json library. This library opens a json file and loads the required data into a list that is easily processed by the script.

Adding intelligence to a chatbot. This problem can be solved by connecting natural language processing and machine learning services. Among the most popular are IBM Watson Conversation, Dialogflow (former api.ai), wit.ai (now the service is being transformed, based on Built-In NLP), LUIS. It is possible to create a chatbot from scratch on the basis of one of these services, or to connect the ability to

contact the service for text/speech recognition to an existing bot [8]. Their action is based on the understanding of “INTENT” – the user’s intentions. Depending on the INTENT, the service returns the predetermined answer to the user. When using services, a database of intentions and answers to them is created, on the basis of which the bot will interact with users.

When the program is completed, we can see the job titles and links to the vacancy, when we click on the link, we will get to the job search site (Fig. 4). For the created functional to operate correctly, it is also necessary:

- For the bot to operate, it is required to install python 3 and libraries on the server (write in the power shell): Pip install pytelebotapi; Pip install bs4
- Download the bot executable file for further operation.

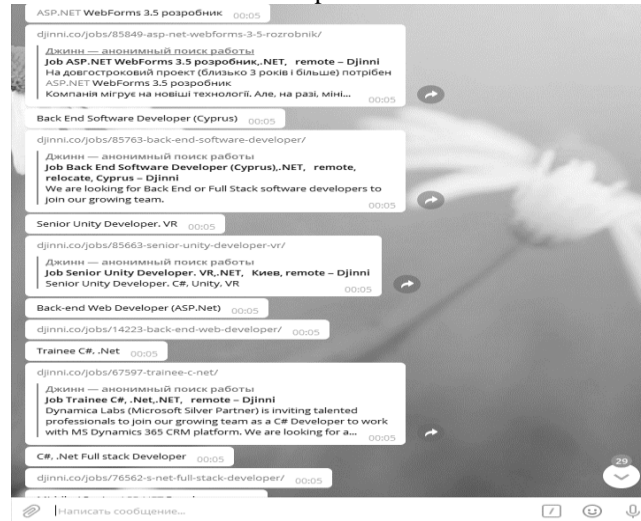


Fig. 4. Completion of the second stage – parsing.

Conclusions

Consolidation of information is not a solution to information problems and needs, but it is becoming one of the important approaches that should be considered along with a huge variety of other information products and services. Consolidated information products and services play an important role in meeting many critical information needs of modern Ukrainian society [3]. Our research has proved that chatbot technology is a modern information technology. And it has also been found that it is easier for service developers to get to the user in the program that he has already installed and opens every day than to convince him of the need to work with a particular application. Therefore, the object of our research has become a search chatbot, which does not require traffic for download, installation time, does not take up space in memory and on the smartphone screen.

The created by us the Python chatbot meets specific needs, namely: job search in the field of IT technologies. The chatbot will find the necessary information in a short period of time, qualify it, display it on a computer or mobile phone screen in a user-friendly form, and save the results of the previous search. As a result, we have obtained the main functional which searches for vacancies throughout Ukraine on one of the most popular sites for IT specialists – <https://djinni.co/>.

References

1. Data consolidation – key concepts [Electronic resource] – Electronic data. – Available at: <http://www.cfin.ru/itm/olap/cons.shtml>
2. Zhezhnych PI Consolidated information resources of databases and knowledge: Textbook / PI Zhezhnych – Lviv: Lviv Polytechnic University, 2010. – 212 p. – ISBN 978-617-607-015-3
3. Kunanets NE Introduction to the specialty "Consolidated Information" / NE Kunanets, VV Pasichnyk. – Lviv: Lviv Polytechnic, 2013. – 196 p. ISBN 978-966-553-975-9
4. Derevyanko AS Technologies and means of information consolidation: Textbook. Manual / A.S Derevyanko, M.N Soloshchuk – Kharkiv: NTU "KhPI", 2008. – 432 p. – ISBN 978-966-593-585-8
5. Makarov V. Parsing html-sites using PHP, Ruby, Python / V. Makarov // Proud member: Web-page. Available at – <http://parsing.valemak.com/>. – Name from the screen.
6. David M. Beasley. Python. Reporting Handbook, 4th Edition. – Translation from English. – SPb.: Symbol-plus, 2010. – 864 p. -ISBN 978-5-93286-157-8.
7. Mark Summerfield. Python Programming 3. Reporting Guide Description. – Translation from English. – SPb.: Symbol-plus, 2009. – 608 p. -ISBN 978-5-93286-161-5.
8. Susie R.A. Python. The most complete guide Description. – SPb.: Petersburg, 2002. – 768 p. -ISBN 5-94157-097-X.