

ХЕРСОНСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ДИЗАЙНУ
КАФЕДРА ІНФОРМАТИКИ І КОМП'ЮТЕРНИХ НАУК

Пояснювальна записка

до дипломної бакалаврської роботи

на тему:

Консолідація даних та агрегування вакансій з веб-джерел.

Виконав: студент 6 курсу, групи 6КІ
спеціальності 122 «Комп'ютерні науки»

Є. В. Молдовану

Керівник: Литвиненко В.І.

Рецензент: Рудакова Г.В.

Хмельницький – 2024 р.

Факультет	Інформаційних технологій та дизайну
Кафедра	Інформатики і комп'ютерних наук
Рівень вищої освіти	<u>другий (магістерський) рівень</u>
Галузь підготовки	<u>12 «Інформаційні технології»</u> (шифр і назва)
Освітньо-професійна програма	<u>Комп'ютерні науки</u> (назва)
Спеціальність	<u>122 «Комп'ютерні науки»</u> (шифр і назва)

ЗАТВЕРДЖУЮ

Завідувач кафедри ІКН,
професор

_____ В.І. Литвиненко

«_____» _____ 2024 року

З А В Д А Н Н Я
НА ДИПЛОМНУ РОБОТУ СТУДЕНТА

Молдовану Євгена Володимировича

(прізвище, ім'я, по батькові)

Тема роботи: Консолідація даних та агрегування вакансій з веб-джерел.

1. Керівник роботи Литвиненко Володимир Іванович, професор, завідувач
кафедри інформатики і комп'ютерних наук

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом ХНТУ від «25» 09 2024 р. № № 499-с

2. Строк подання студентом роботи

3. Вихідні дані до роботи _____

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити): Вступ. Розділ 1. Аналіз функціоналу web сайтів з пошуку роботи. Розділ 2. Методологія програмування для консолідації даних методами NLP. Розділ 3. Практична реалізація прототипу для парсингу і обробки даних вакансій.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень) Малюнків – 58. Таблиць - 2

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання 25.09.2024

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломної роботи	Строк етапів роботи	Прим.
1	Огляд літературних джерел та пошук технологій	10.10.24-12.10.24	
2	Аналіз та тестування основних сайтів пошуку роботи	12.10.24-15.10.24	
3	Огляд потреб до програмного і апаратного забезпечення	15.10.24-16.10.24	
4	Огляд конкурентних сервісів аналогічних тематиці прототипу	16.10.24-17.10.24	
5	Розробка методології і архітектури прототипу	17.10.24-01.11.24	
6	Пошук рішень, бібліотек для вирішення задач семантичного пошуку	01.11.24-8.11.24	
7	Налаштування IDE PyCharm і Python	8.11.24-10.11.24	
8	Реалізація основного функціоналу прототипу	10.11.24-25.11.24	
9	Оформлення роботи, посилань, візуальної частини прототипу	25.11.24-1.12.24	

Студент _____ Є. В. Молдовану
(підпис) (прізвище та ініціали)

Керівник роботи _____ В.І. Литвиненко
(підпис) (прізвище та ініціали)

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

Агрегатор	інформаційна система з агрегації контенту (оголошень або вакансій, резюме)
Портал з пошуку роботи	спеціалізований сайт який публікує резюме або оголошення (вакансії)
Blacklist	можливість додавання оголошень в спеціальний список, після чого вони являються прихованими.
НЛП	Нейролінгвістичне програмування (свідомості)
ПО	програмне забезпечення
API	application programming interface - програмний інтерфейс програми
MVP	minimum viable product – мінімально життєздатний продукт
HRM	human resource management (управління персоналом)
CRM	customer relationship management (Система управління взаємовідносинами з клієнтами)
ATS	Applicant Tracking System (Система відстеження кандидатів)

Реферат

Пояснювальна записка включає в себе малюнків – 58, таблиць – 2, сторінок – 129, 62 літературне посилання на джерела, додатки з вихідним кодом.

NLP, LABSE, PYTHON, VECTOR, STREAMLIT, JS, PYCHARM, VACANCY, PARSER, ANNOY, ФІЛЬТРАЦІЯ.

Актуальність теми.

У зв'язку з війною в Україні дуже зросло безробіття. Мільйони людей перемістилися в більш західні регіони, а частина і зовсім виїхала за кордон. У зв'язку з руйнуванням не тільки осель, але й бізнесу багато підприємств переїхало на захід країни. Саме тому важливо щоб у безробітних і бізнесу були майданчики для пошуку один одного. В Україні вже багато років існують такі сайти, які дозволяють організовувати публікацію оголошень (вакансій) і таким чином сприяють пошуку роботи. В той же час бізнес процеси багатьох таких сайтів не дозволяють і не зацікавлені в максимальній швидкості отримання безробітним роботи. Це пов'язано з методами монетизації цих сайтів, де компанія зацікавлена щоб пошуковець більше часу проводив на їх порталі, більше дивився реклами і таким чином вони заробляли гроші.

Ще однією проблемою є те, що на кожному окремому сайті свої правила і методи монетизації. Як наслідок на ринку багато сайтів і дуже часто так буває, що роботодавець публікує вакансію лише на одному чи на двох порталах. У разі якщо вакансія приваблива, то може пройти не більше 1-2 днів і вона буде закритою.

Тому у безробітних постає велика проблема – потрібно кожного дня аналізувати сотні оголошень, деякі з них вже були відбраковані, деякі неякісні. В результаті людина витрачає час не на підготовку до роботи, а на її пошук.

Тому і виходить, що на пошук нової роботи витрачаються в середньому від 3 місяців, а для дуже рідких вакансій і до 7 місяців.

Тому крім самих порталів з пошуку роботи безробітний часто підписується на телеграм, viber групи з оголошеннями, які завчасно займаються фільтрацією (по місту, ціні, професії). Типовий безробітний підписний на десятки каналів, проглядає 5-10 сайтів пошуку роботи, на що у нього кожен день витрачається по декілька годин на пошуки і взаємодію з менеджерами по персоналу.

Тому для безробітних дуже допомогла б наявність програмного забезпечення, яке допоможе автоматизувати ці процеси. В першу чергу це агрегація вакансій, можливість пошуку і відображення консолідованої інформації.

Мета і задачі дослідження

Метою дослідження є пошук недоліків програмного забезпечення (сайтів пошуку вакансій) на ринку праці України і встановлення основних характеристик для створення прототипу веб-сервісу для аналізу, візуалізації та пошуку вакансій.

Об'єкт дослідження

Процес обробки та консолідації гетерогенних даних.

Предмет дослідження

Процес парсингу та агрегації даних вакансій для формування обробки і візуалізації інформації.

Практичне застосування:

Розроблений прототип дозволяє проводити аналітичні дослідження для відображення реальних значень мінімальної, медіанної і максимальної зарплати. Також візуалізує ці дані у вигляді charts:

- Дані кількості вакансій по містах
- Мінімальна, медіанна та максимальна зарплата

Реалізована можливість обробки даних з фільтрацією по ключовим словам, містам, числовим значенням (наприклад зарплати).

Будь які відфільтровані дані можливо експортувати в csv для наступного аналізу.

Також реалізований пошук по даним з видаленням дублів вакансій. Він може бути використаний для пошуку вакансій безробітними або аналізу конкурентів для компаній.

Наукова новизна:

Використання моделі LaBSE для видалення дублів і ранжування.

Наукова новизна полягає у створенні інформаційної системи з використанням моделі LaBSE для ефективного видалення дублів вакансій та їхнього ранжування за релевантністю. Модель дозволяє порівнювати вакансії за семантичними ознаками, що забезпечує створення бази даних без повторів і підвищує точність пошуку.

База даних вакансій з сайтів Word.ua і Robota.ua в Україні.

Друга новизна полягає у створенні бази даних вакансій ІТ сектору, зібраних з самих великих за відвідуємостю українських сайтів для пошуку роботи - Work.ua і Robota.ua, що дозволяє зібрати актуальні пропозиції в ІТ-сфері під час війни та забезпечує доступ до консолідованої інформації без дублів.

Сформований датасет може бути використаний для аналізу ринку праці в Україні і написання наукових статей по економічним напрямам.

Abstract

Explanatory note includes figures - 58, tables - 2, pages - 129, 62 references, appendices with source code.

NLP, LABSE, PYTHON, VECTOR, STREAMLIT, JS, PYCHARM, VACANCY, PARSER, ANNOY, FILTERING.

Relevance of the topic.

Due to the war in Ukraine, unemployment has increased dramatically. Millions of people have moved to more western regions, and some have gone abroad. Due to the destruction of not only homes but also businesses, many companies moved to the west of the country. That is why it is important for the unemployed and businesses to have platforms to find each other. For many years now, Ukraine has had websites that allow organizing the publication of ads (vacancies) and thus facilitate job search. At the same time, the business processes of many of these sites do not allow and are not interested in maximizing the speed of getting a job for the unemployed. This is due to the methods of monetization of these sites, where the company is interested in the job seeker spending more time on their portal, watching more ads, and thus making money.

Another problem is that each individual site has its own rules and methods of monetization. As a result, there are many sites on the market, and it often happens that an employer posts a vacancy on only one or two portals. If the job is attractive, it may take no more than 1-2 days for it to be closed.

Therefore, the unemployed face a big problem: they need to analyze hundreds of ads every day, some of which have already been rejected, some of which are of poor quality. As a result, a person spends time not preparing for work, but looking for it.

That's why it turns out that it takes an average of 3 months to find a new job, and up to 7 months for very rare vacancies.

Therefore, in addition to the job search portals themselves, the unemployed often subscribe to telegram and viber groups with ads that are filtered in advance (by city, price, profession). A typical unemployed person subscribes to dozens of channels, browses 5-10 job search sites, and spends several hours every day searching and interacting with HR managers.

Therefore, it would be very helpful for the unemployed to have software that would help automate these processes. First of all, it is the aggregation of vacancies, the ability to search and display consolidated information.

The purpose and objectives of the study

The purpose of the study is to find the shortcomings of software (job search sites) in the Ukrainian labor market and to establish the main characteristics for creating a prototype of a web service for analyzing, visualizing and searching for jobs.

Object of research

The process of processing and consolidating heterogeneous data.

Subject of research

The process of parsing and aggregating job data to form information processing and visualization.

Practical application:

The developed prototype allows conducting analytical studies to display the real values of the minimum, median, and maximum salaries. It also visualizes this data in the form of charts:

- Data on the number of vacancies by city
- Minimum, median, and maximum salary

Implemented the ability to process data with filtering by keywords, cities, numeric values (for example, salary).

Any filtered data can be exported to csv for further analysis.

Data search with duplicate job postings removal is also implemented. It can be used to search for vacancies by the unemployed or analyze competitors for companies.

Scientific novelty:

The use of the LaBSE model for duplicate removal and ranking.

The scientific novelty lies in the creation of an information system using the LaBSE model to effectively remove duplicate jobs and rank them by relevance. The model allows comparing jobs by semantic features, which ensures the creation of a database without repetitions and increases the search accuracy.

The database of vacancies from Work.ua and Robota.ua in Ukraine.

The second novelty is the creation of a database of IT sector vacancies collected from the largest Ukrainian job search sites in terms of traffic - Work.ua and Robota.ua, which allows collecting current offers in the IT sector during the war and provides access to consolidated information without duplication.

The generated dataset can be used to analyze the labor market in Ukraine and write scientific articles on economic issues.

ЗМІСТ

ВСТУП	14
РОЗДІЛ 1. АНАЛІЗ ФУНКЦІОНАЛУ WEB САЙТІВ З ПОШУКУ РОБОТИ	16
1.1. Консолідація даних.....	16
1.2. Методи пошуку вакансій	20
1.3. Сайти пошуку вакансій	21
1.5. Основний функціонал сайтів по пошуку роботи.	31
1.6. Основні недоліки порталів пошуку роботи	32
Висновки до розділу 1.....	34
РОЗДІЛ 2. МЕТОДОЛОГІЯ ПРОГРАМУВАННЯ ДЛЯ КОНСОЛІДАЦІЇ ДАНИХ МЕТОДАМИ NLP	35
2.1. Опис програмної реалізації прототипу.....	35
2.2 Аналіз схожих розробок.....	48
2.3. Perfectjobexplorer.com.....	52
2.4 Аналогічне ПЗ поза межами України.....	55
2.5 Основні недоліки аналогічного програмного забезпечення	59
2.6 Етапи і методологія життєвого циклу розробки програмного забезпечення	59
2.7 Архітектура програмного забезпечення.....	62
2.8 NLP. Методи роботи з текстовим описом вакансій.....	68
2.9 Вибір мови програмування та IDE	75
Висновки до розділу 2.....	79
РОЗДІЛ 3. ПРАКТИЧНА РЕАЛІЗАЦІЯ ПРОТОТИПУ ДЛЯ ПАРСИНГУ І ОБРОБКИ ДАНИХ ВАКАНСІЙ	80
3.1 Архітектура парсингу. Юридичні моменти	80
3.2 Бібліотеки і технології для побудови прототипу.....	86
3.3 Структура баз даних SQLITE3 у прототипі	87

3.4 Інтерфейс користувача і візуалізація графіків. Функціонал прототипу..... 92

Висновки до розділу 3..... 104

ВИСНОВКИ ДО РОБОТИ..... 105

ДОДАТКИ 108

ДОДАТОК А. ОСНОВНА ПРОГРАМА...... 108

ПЕРЕЛІК ПОСИЛАНЬ. 123

Вступ.

Наразі в зв'язку з розвитком інформаційних технологій, а в особливості у напрямлені з штучним інтелектом є великий попит на консолідацію інформації. З кожним роком кількість інформації все збільшується і утримувати увагу людині все важче. Це все має вплив на прийняття рішень, коли не оперуючи реальною інформацією можливі похибки.

В ситуації пошуку роботи середньому безробітному важливо отримувати інформацію в агрегованому вигляді і не витрачати час на перегляд дубльованих вакансій. Також агрегація вакансій може дати аналітичну інформацію, яка буде використана для прийняття рішення зміни професії, міста проживання заради кращої роботи. Враховуючи інфляційні процеси і зміну зарплатних очікувань на ринку праці потрібно весь час розуміти, чи достатня оплата на роботі, чи ж є зиск зміни місця проживання.

Відповісти на такі питання може лише сервіс з консолідованими даними, який підтримує відображення унікальних вакансій з підтримкою агрегації інших джерел даних, статистикою по середній зарплаті по професіям, містам або часовим проміжкам.

Апробація результатів магістерської кваліфікаційної роботи.

Результати магістерської кваліфікаційної роботи доповідалися на:

VII Всеукраїнська науково-практична Інтернет-конференція здобувачів вищої освіти та молодих вчених “Сучасні комп'ютерні системи та мережі в управлінні” Україна м. Хмельницький, м. Херсон, 29 листопада. 2024 р.

Публікації:

Молдовану Є., “Проблематика онлайн-пошуку роботи за допомогою сайтів із вакансіями”. “Сучасні стратегії сталого розвитку держави та суспільства: наукові горизонти та перспективи” збірник матеріалів I-ї Науково-практичної

конференції з міжнародною участю / за наук. ред. проф. О. В. Чепелюк, В. Д. Філіппової, В. М. Демченка. Хмельницький: ХНТУ, 2024 с. 405-408.

Молдовану Євген Володимирович, Лур'є Ірина Анатоліївна “Реалізація функціоналу it-сервіса пошуку роботи для незахищених верств населення” "Синергія науки і бізнесу у повоєнному відновленні регіонів України: Матеріали II Міжнародної науково-практичної конференції, 24–26 квітня 2024 року, ТОМ 2, Одеса." Олді+, 2024 с. 53-56.

Молдовану Є. В., Литвиненко В. І. “Виявлення дублікатів вакансій з використанням методів обробки та семантичного аналізу”. Сучасні комп'ютерні системи та мережі в управлінні : матеріали VII Всеукр. наук.-практ. Інтернет-конф. здобувачів вищої освіти та молодих вчених (29 листоп. 2024 р., м. Хмельницький, м. Херсон) / за ред. А. А. Григорової. Херсон : Книжкове видавництво ФОП Вишемирський В. С., 2024. Готується до видання.