

**ПОЛТАВСЬКИЙ ДЕРЖАВНИЙ АГРАРНИЙ УНІВЕРСИТЕТ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ЕКОНОМІКИ, УПРАВЛІННЯ,
ПРАВА ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНФОРМАЦІЙНИХ СИСТЕМ ТА ТЕХНОЛОГІЙ**

Пояснювальна записка

до кваліфікаційної роботи на здобуття ступеня вищої освіти магістр

на тему: **«Аналіз та прогнозування динаміки ринку онлайн-вакансій в
ІТ-галузі України за допомогою інструментів Python»**

Виконала: здобувачка вищої освіти
за освітньо-професійною програмою
Інформаційні управляючі системи та
технології спеціальності
126 Інформаційні системи та технології
ступеня вищої освіти магістр
групи 126ІСТ_мз_2022[1](л.н.)
Дерев'янка К.О.
Керівник: Флегантов Л.О.
Рецензент: Петраш Р.В.

Полтава – 2023 року

ВСТУП

Тема кваліфікаційної роботи є актуальною в сучасному світі, керованому даними. В епоху, коли дані набули великого масштабу, максимальне використання їхнього потенціалу стає ключовим фактором. Інформаційні технології та методи аналізу даних набули статусу невід'ємних інструментів для отримання практичних знань із великих масивів даних. Актуальність цього дослідження полягає також і в можливості поєднати теоретичні знання в галузі інформаційних технологій з їх реальним практичним застосуванням. Дослідження ринку онлайн-вакансій в Україні дозволяє продемонструвати, як ці технології допомагають особам, які приймають рішення, зрозуміти динаміку ринку праці, попит на фахівців та тенденції щодо бажань шукачів роботи.

Зв'язок роботи з науковими програмами, планами, темами. Робота виконана у відповідності до науково-дослідної ініціативної теми «Організаційно-методологічні аспекти впровадження інформаційно-комунікаційних систем і технологій в управлінні діяльністю сучасних організацій та підприємств за умов переходу до цифрової економіки» ДРН 0117U003099.

Метою кваліфікаційної роботи є комплексне вивчення технологій аналізу даних та їх практичного застосування на прикладі дослідження динаміки розвитку ринку праці в Україні, яка відображається в оголошеннях про роботу в Інтернеті.

Основними завданнями роботи є:

- структуроване вивчення технологій аналізу даних та представлення результатів дослідження у вигляді інформаційної панелі (дешборду);
- збір та об'єднання даних із найпопулярніших онлайн-ресурсах для розміщення оголошень та вакансій і платформ, які містять супутні дані, які стосуються України;

– визначити, як ринок праці розвивався у відповідь на зовнішні фактори, такі як економічні зміни, глобальні події та технологічний прогрес.

Об'єкт дослідження – це онлайн-ринок праці України. Він охоплює всі елементи, динаміку та компоненти, пов'язані з ринком онлайн-вакансій в Україні. Сюди відносяться дані про вакансії, шукачів, роботодавців, галузі, регіони та фактори, які впливають на цей ринок.

Предметом дослідження є практичне застосування інформаційних технологій і методів аналізу даних для дослідження, аналізу та отримання висновків з даних онлайн-ринку вакансій в Україні. Сюди також відносяться: збір даних, обробка та підготовка даних, аналіз та візуалізація.

Методами дослідження є: порівняльний аналіз, ситуаційний аналіз, кількісний аналіз, емпіричне дослідження, побудова наочного представлення даних у вигляді інформаційної панелі.

Інформаційна база дослідження: навчальна та науково-технічна література, дані офіційної документації з Python, вітчизняні та зарубіжні стандарти, державна статистика та галузеві публікації, аналітичні джерела мережі інтернет пов'язані з ринком праці в Україні, що включають списки та описи вакансій, профілі кандидатів, історичні тенденції, отримані з провідних українських вебсайтів для пошуку роботи, загальнодоступні звіти про ринок праці.

Елементи наукової новизни. Застосування сучасних методів аналізу даних для дослідження українського ринку праці дозволило дізнатися, які саме зміни відбулися у потребах, кількісних та якісних складових вакансій. Дослідження продемонструвало трансформаційний потенціал рішень, зроблених на основі розгляду даних, для розуміння динаміки ринку праці, що має значення у сучасному дата-центричному світі.

Практичне значення даної роботи полягає отриманні практичної інформації про зміни у вакансіях та характеристиках українського ринку праці завдяки застосуванню сучасних методів аналізу та візуалізації даних. Аналізуючи кількісні та якісні характеристики відповідних даних, це

дослідження дозволить підприємствам і організаціям приймати рішення з оглядом на зміни. Це дозволить краще зрозуміти мінливість ринку праці та ситуації його окремих учасників та галузей. Крім того, розробка та застосування методологій аналізу даних у контексті ринку праці підкреслить їх актуальність у реальному світі та сприятиме розвитку інформаційних технологій і методів аналізу.

Апробація результатів дослідження. За результатами проведеного дослідження опубліковано тези доповідей: «Співвідношення понять ІТ-ринку та ІТ-вакансії у контексті прогнозування ІТ-ринку та ринку ІТ-праці», Матер. VIII Всеукраїнської науково-практичної інтернет-конференції «Управління ресурсним забезпеченням господарської діяльності підприємств реального сектору економіки», 23 листопада 2023 року, м. Полтава; «Моделі прогнозування ринку праці та їх застосування в ІТ-галузі», Матер. VIII Всеукраїнської науково-практичної інтернет-конференції «Управління ресурсним забезпеченням господарської діяльності підприємств реального сектору економіки», 23 листопада 2023 року, м. Полтава. За результатами дослідження здійснено 2 публікації тез доповідей.

Структура та обсяг кваліфікаційної роботи. Кваліфікаційна робота складається зі вступу, трьох розділів та висновків. Викладена на 81 сторінці, містить 28 рисунків, 2 таблиці, 3 додатки. Список використаних джерел налічує 31 найменування.

РОЗДІЛ 1

ТЕОРЕТИЧНІ ОСНОВИ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ РИНКУ ІТ-ВАКАНСІЙ

1.1 Фактори що впливають на зміни на ринку праці

Ринок праці – це динамічна сутність, яка постійно розвивається, і на неї впливає безліч факторів, які формують її контури. Щоб отримати всебічне розуміння змін і тенденцій на ринку праці, вкрай важливо ретельно вивчити ці фактори впливу.

Ринок з точки зору соціології ринку – один із елементів суспільства, який тісно пов'язаний з його соціальною структурою і соціальними процесами [1, с. 162]. Ці соціальні процеси, звичайно, є не лише місцевими, але й включають ситуацію на світовому ринку. Одним із найвпливовіших факторів по всьому світу стало поширення захворювання COVID (початку приблизно у 2020 році). Наслідки вірусу залишили вагомий слід на ринках праці в усьому світі. В Україні пандемія також спровокувала низку трансформацій. Люди змінили спосіб роботи, а віддалена робота стала нормою для багатьох галузей. Порушення колись налагодженого способу роботи економіки призвели до коливань кількості вакантних робочих місць, причому в деяких секторах спостерігалось зростання попиту, а в інших – різкий спад. Поведінка шукачів роботи також змінилася, оскільки неможливо було ігнорувати безпеку праці. Логічним наслідком став занепад для підприємств, робота яких передбачала тісний контакт між людьми, у той час, як ті, кому живе спілкування було не потрібне, стали більш успішними.

Іншим, найвагомим фактором змін для України, а також певною мірою і для інших країн, стали воєнні дії на території України. Зокрема у східних регіонах, така ситуація спричинила появу нових викликів для ринку праці. Низка регіонів лишилася значної кількості робочої сили, підприємства

були змушені або змінити місце ведення діяльності, або були взагалі зруйновані. Велика кількість людей покинула Україну та влаштувала нове життя закордоном. Деяка частина населення перемістилася в більш безпечні регіони.

Розглянемо й інші загальні фактори, які мають вплив на кожну країну. На зміну в потребах у певних кадрах вплинув розвиток технологій. Впровадження автоматизації та штучного інтелекту призвело до зміни робочих ролей у різних галузях. Найочевиднішою є поступова заміна працівників на роботів, тому в певних сферах можливе зменшення бажаної кількості представників деяких професій.

Встановлення розумних кас, які дозволяють користувачу самостійно сплачувати за покупку загрожує професії касира. Це явище спостерігається і в Україні. В деяких країнах (для прикладу Південна Корея) вже певний час існують магазини, де взагалі немає касирів, бо загальний рівень безпеки дозволяє бізнесу повністю довіритися клієнту для проведення розрахунку. Зважаючи на той факт, що мета комерційного підприємства – це отримання прибутку, компанії, які почали впроваджувати апарати самообслуговування, не прагнуть додати нові витрати на їх встановлення та підтримку і при цьому мати незмінний кадровий склад. Із цього слідує, що кількість касирів в українських магазинах буде поступово зменшуватися, аж поки вони не будуть повністю витіснені, де це буде обґрунтовано.

Іншим прикладом є загрозливе становище для дикторів та акторів. Типовим є те, що задовго до популяризації штучного інтелекту актори співпрацювали із компаніями задля створення проєктів, умови яких не передбачали та не описували можливість використання голосів для синтезу із ШІ. Приблизно у 2022-2023 роках актори почали випадково дізнаватися про те, що їх голос тепер вільно продається або навіть доступний безкоштовно. Тому можна стверджувати, що розвиток технологій, з одного боку, полегшує працю бізнесу, але в той же час ставить працівників сфери культури у не вигідне положення і сприяє зменшенню оплати за їх роботу [2]. У

контексті теми цієї роботи розвиток технологій представлений тенденціями у використанні мов програмування.

Отже, такі фактори можна використати для прогнозування динаміки ринку праці:

кількість хворих на covid;

дані про переміщення населення через кордон;

популярність мов програмування.

У таблиці 1.1 наведений структурований опис найбільш потужних глобальних факторів, які впливають на зміни на ринку IT-праці.

Таблиця 1.1 – Глобальні фактори змін на ринку IT-праці

Фактори	Опис	Приклад (IT)
Технологічний прогрес	Розвиток нових технологій та інновацій може змінювати попит на певні навички та професії	Зростання сфери штучного інтелекту збільшує попит на спеціалістів з машинного навчання та аналізу даних.
Глобалізація	Взаємодія і співпраця компаній на міжнародному рівні впливає на ринок праці, роблячи його більш конкурентним	Аутсорсинг та офшоринг програмного забезпечення в країни з нижчими заробітними платами.
Зміни в освіті та навчанні	Розвиток освітніх програм і курсів, що відповідають сучасним вимогам ринку	Поява спеціалізованих курсів та програм, які навчають навичкам роботи з великими даними та хмарними технологіями.
Зміни в законодавстві	Урядові регуляції та закони можуть впливати на ринок праці, наприклад, через зміни в трудовому законодавстві.	Запровадження GDPR (Загального регламенту про захист даних) в ЄС вимагає наявності фахівців з питань конфіденційності даних.
Демографічні зміни	Зміни в віковій структурі населення впливають на попит та пропозицію робочої сили.	Старіння населення може збільшити потребу в IT-рішеннях для охорони здоров'я.
Економічні зміни	Економічні коливання, такі як рецесії або буми, впливають на загальний попит на робочу силу.	В періоди економічного спаду може зменшитися попит на дорогі IT-проекти.
Соціальні та культурні зміни	Зміни у суспільних цінностях і культурі можуть впливати на робоче середовище та корпоративну культуру.	Розвиток культури стартапів і інновацій сприяє зростанню підприємницької активності в IT.

Військова загроза / Загроза війни (зокрема в Україні)**:	Військові конфлікти або загроза війни можуть мати значний вплив на ринок праці через перерозподіл ресурсів, зміну економічних пріоритетів та необхідність адаптації до нових реалій.	В умовах воєнного стану в Україні багато ІТ-компаній були змушені адаптувати свою діяльність, включно з переміщенням офісів та співробітників в безпечніші регіони або за кордон, а також переорієнтацією на проекти, які відповідають зміненим потребам в умовах кризи.
Загроза пандемії (COVID-19)	Пандемія може радикально впливати на ринок праці, змінюючи умови праці, потреби ринку та підхід до організації робочого процесу.	Приклад (ІТ)**: Пандемія COVID-19 спричинила масовий перехід на віддалену роботу в ІТ-галузі, збільшивши попит на інструменти для дистанційної роботи та онлайн комунікації. Також підвищилася потреба в розробці і впровадженні ІТ-рішень для медичної сфери та електронної комерції.

Окремо відмітимо, що всі ці фактори взаємопов'язані та спільно формують динаміку змін на ринку праці, особливо у такій швидкозмінній галузі, як ІТ. Пандемія змусила багато компаній швидко адаптуватися до нових умов роботи, що значно вплинуло на тенденції розвитку та вимоги на ринку праці, особливо в ІТ-секторі. Військова загроза також вносить значні корективи в роботу ІТ-галузі, вимагаючи швидкої адаптації та гнучкості від компаній та працівників.

1.2 Основні поняття та концепції аналізу даних та прогнозування

У цьому розділі ми детально розглядаємо особливості, які визначають методи аналізу даних для дослідження змін на ринку вакансій. Оскільки ринок вакансій за своєю суттю динамічний і на нього впливає безліч факторів, розуміння унікальних характеристик окремих методів аналізу даних у цьому контексті має першочергове значення.

Перший метод полягає в обробці даних у реальному часі. Він найкраще підходить для тих, хто прямо володіє даними для аналізу. Цим методом користуються платформи та вебсайти, на яких розміщуються вакансії, профілі кандидатів та інші дані, які часто оновлюються. Тому необхідна

можливість отримувати, обробляти та візуалізувати дані в режимі реального часу задля представлення найновішої інформації про динаміку ринку праці. Суміжним методом є пакетна обробка даних, яка передбачає попередній збір даних, розподіл на групи даних, а також часові інтервали роботи з даними. У такому випадку кінцевий результат не доступний одразу через групування даних згідно часових проміжків [3]. Обробка в режимі реального часу надає миттєве уявлення про дані, сприяє швидкому прийняттю рішень і вирішенню проблем. Однак цей метод вимагає значних обчислювальних ресурсів і складної інфраструктури. Пакетна обробка ефективна для обробки великих обсягів даних і керування ресурсами, забезпечуючи цілісність даних і масштабованість. Однак це призводить до затримки готової статистики та вимагає складних конвеєрів даних – процесу збору, обробки та передачі даних між елементами пов'язаної аналітичної системи. Вибір між цими підходами залежить від конкретних сценаріїв використання, при цьому обробка в реальному часі ідеально підходить для випадків, що вимагають негайної реакції, а пакетна обробка підходить для масштабної обробки даних за розкладом.

Іншою потужною технікою аналізу даних є багатовимірний аналіз. Цей метод використовується в аналізі даних для одночасного вивчення зв'язків між кількома змінними. Він виходить за рамки традиційного однофакторного аналізу, який розглядає одну змінну за раз, беручи до уваги те, як змінні взаємодіють одна з одною. Цей метод особливо корисний під час роботи зі складними наборами даних. Перевагою мультіваріативного аналізу є удосконалене прогнозне моделювання змінної. Розглядаючи кілька змінних, багатфакторний аналіз може підвищити точність і ефективність прогнозних моделей [4, с. 3]. Щоб отримати надійні результати, багатфакторний аналіз вимагає якісних та чистих даних. Зашумлені або неповні набори даних можуть призвести до оманливих результатів. Іншим викликом є інтерпретація залежностей багатовимірних результатів, оскільки для цього необхідне розуміння зв'язків між декількома змінними, які не завжди можуть

бути інтуїтивно зрозумілими. Прикладом такої ситуації є очікування того, що пропозиції з більшою зарплатнею, вигідне розташування, висока репутація компанії та комплексні пільги позитивно вплинуть на рівень прийняття вакансій шукачами. Крім того, може здатися, що посади вищого рівня будуть прийняті з більшою ймовірністю через їхній статус. Однак багатофакторний аналіз показує й інші аспекти:

- шукачі роботи можуть надавати пріоритет розташуванню компанії або її репутації, зважаючи саме на їх місце постійного проживання;

- IT-фахівці не обов'язково приймуть керівні посади (senior-позиції), навіть задля високої оплати. Саме тому, всупереч очікуванням, посади середнього рівня часто популярніші через збалансованість між викликами на роботі та особистим життям;

- зарплата може бути не головним фактором, оскільки інші переваги, які надає компанія, можуть мати більше значення. Це можуть бути такі переваги, як медичне страхування, пенсійні виплати та гнучкий графік.

Деякі ключові аспекти та методи, що застосовуються в аналізі даних та прогнозуванні, представлені у таблиці 1.2.

Таблиця 1.2 – Ключові аспекти та методи прогнозування

Термін	Опис терміну	Приклади (IT галузь)
Data Collection (Збір даних)	Процес збору релевантних даних з різноманітних джерел.	Збір даних про використання веб-сайту за допомогою Google Analytics.
Data Cleaning (Очищення даних)	Видалення або виправлення неточних, неповних або застарілих даних.	Очищення некоректних або повторюваних записів у базі даних користувачів.
Data Integration (Інтеграція даних)	Об'єднання даних з різних джерел для створення єдиної бази даних.	Інтеграція баз даних клієнтів з різних сервісів CRM для єдиної аналітики.
Data Analysis (Аналіз даних)	Використання статистичних і аналітичних методів для виявлення закономірностей і тенденцій у даних.	Аналіз лог-файлів сервера для виявлення проблем з продуктивністю сайту.
Data Visualization (Візуалізація даних)	Графічне представлення даних для кращого розуміння та інтерпретації.	Використання дашбордів у Tableau для візуалізації трафіку сайту та поведінки користувачів.

Modeling (Моделювання)	Створення математичних моделей для імітації реальних процесів або систем.	Розробка моделі машинного навчання для автоматичної класифікації електронних листів.
Forecasting (Прогнозування)	Використання моделей для передбачення майбутніх подій чи результатів.	Прогнозування навантаження на сервери в певний час доби для оптимізації ресурсів.
Machine Learning (Машинне навчання)	Методики, які дозволяють комп'ютерам вчитися з даних і покращувати свої результати без прямого програмування.	Використання нейронних мереж для розпізнавання образів у мобільному додатку.
Artificial Intelligence (Штучний інтелект)	Створення систем, що можуть виконувати задачі, які традиційно потребують людського інтелекту.	Розробка інтелектуального асистента на базі AI для підтримки користувачів.
Big Data (Великі дані)	Обробка і аналіз величезних масивів даних, які неможливо обробити звичайними методами.	Аналіз великих датасетів з даними про використання програмного забезпечення.
Predictive Analytics (Прогностичний аналіз)	Використання даних, статистичних алгоритмів і машинного навчання для ідентифікації ймовірності майбутніх результатів на основі історичних даних.	Використання алгоритмів прогностичного аналізу для виявлення тенденцій у розвитку програмного забезпечення.
Data Mining (Видобування даних)	Процес аналізу великих наборів даних для виявлення зразків та зв'язків.	Майнінг даних з соціальних мереж для аналізу настроїв користувачів щодо IT продуктів.

1.3 Інструменти Python для аналізу даних та прогнозування

У цій секції детально досліджуються технології та інструменти, необхідні для проведення аналізу та прогнозування змін динамічних характеристик ринку IT-вакансій в Україні. Результативність дослідження обумовлюється, зокрема й розумінням наявних технологій для збору, обробки, аналізу та візуалізації даних.

Одним із засобів всебічного маніпулювання даними є мови програмування, якими користуються фахівці сфери даних. Наразі найпопулярнішим рішенням є Python. Це мова програмування високого рівня

загального призначення, відома інтуїтивно зрозумілим синтаксисом, який імітує природну мову [5]. Можливості Python використовуються для багатьох завдань, але найпопулярнішими є три напрямки:

- наука про дані та аналіз даних;
- розробка вебдодатків;
- автоматизація/ створення сценаріїв.

Потужність Python у будь-яких сценаріях, які стосуються роботи з даними, штучним інтелектом та машинним навчанням зумовлюється не стільки можливостями чистої мови, скільки великою базою готових рішень – фреймворків та спеціальних бібліотек. Хоча поширеність мови програмування Python не обов'язково означає, що вона краща, але це свідчить про її ширше використання. Тому логічно припустити, що вона може мати більш надійну спільноту для постійної підтримки та розвитку. Це і стало головною причиною вибору саме цієї мови замість R [6].

Розглянемо основні бібліотеки для роботи з даними.

По-перше, розглянемо Pandas, одну з найважливіших бібліотек, яка має зручні і доступні методи для операцій з даними, включаючи фільтрацію, сортування, об'єднання та агрегацію. Pandas добре підходить для багатьох типів даних [7]:

- табличні дані з неоднорідно типізованими стовпцями, як у таблиці SQL або електронних Excel- таблицях;
- упорядковані та неупорядковані часові ряди;
- довільні матричні дані (однорідно типізовані або неоднорідні) з мітками рядків і стовпців.

Pandas оперує двома основними структурами: Series (1-вимірна) і DataFrame (2-вимірна). Бібліотека Pandas побудована на основі NumPy і легко інтегрується в обчислювальне середовище з багатьма іншими сторонніми бібліотеками. Об'єкт Series відрізняється від звичайного одновимірного масиву (або списку в Python) наявністю асоційованих міток,

відомих як індекси, для кожного елемента списку. Ця особливість надає йому властивості асоціативного масиву або словника в Python.

Доступ до списку елементів та індексів об'єкта Series можна отримати за допомогою атрибутів `values` та `index` відповідно. Крім того, з об'єктом Series виконувати операції, такі як доступ до декількох індексів, масове присвоєння, фільтрація за допомогою операторів більше/менше та застосування математичних операцій.

DataFrame можна сприймати як звичайну таблицю, оскільки це дійсно таблична структура даних. Кожна таблиця складається з рядків та стовпців. У DataFrame стовпці представлені об'єктами Series, елементи яких формують рядки. DataFrame має два індекси: один для рядків, інший – для стовпців. Якщо індекс стовпця не вказано, pandas автоматично створює цілочисельний індекс RangeIndex від 0 до N-1, де N – це кількість рядків у таблиці.

Доступ до рядків можна отримати різними способами: `.loc` для доступу за міткою рядка, `.iloc` для доступу за числовим індексом (починаючи з 0). Механізм фільтрації працює так само, як і для об'єктів Series. [8]

Вигляд структур даних можна змінювати: додавати та видаляти стовпці, групувати, керувати відсутніми значеннями, об'єднувати стовпці, завантажувати дані із сторонніх ресурсів (бази даних, CSV, Excel тощо) з високою швидкістю.

NumPy (Numerical Python) – бібліотека, спеціально розроблена для ефективних числових операцій і складної обробки даних. Бібліотека має підтримку багатовимірних масивів і матриць разом та набір математичних функцій для роботи з цими масивами. Основною структурою в NumPy є об'єкт `ndarray` (n-вимірний масив). Масив `ndarray` (список) значно швидший та ефективніше використовує пам'ять, ніж вбудовані списки Python. Це досягнуто, окрім іншого, завдяки оптимізованій векторизації, трудомісткі процеси, такі як цикли, індексування тощо відбуваються в оптимізованому, попередньо скомпільованому коді мовою C [9]. NumPy є ключовим компонентом для ефективної обробки числових даних. Він відіграє значну

роль у таких завданнях, як очищення, перетворення та чисельне моделювання даних. Окрім цього, NumPy підтримує широкий спектр апаратного забезпечення та обчислювальних платформ і добре працює з розподіленими бібліотеками, бібліотеками GPU та розрідженими масивами. Як і Pandas, цю бібліотеку легко застосовувати з іншими бібліотеками, наприклад для візуалізації даних за допомогою Matplotlib.

Matplotlib застосовується для візуалізації даних та побудови графіків і діаграм, що сприяє кращому сприйняттю результатів аналізу. Бібліотека має методи для створення статичних, анімованих, а також інтерактивних візуалізацій у Python [10]. На основі цієї бібліотеки будуються більш високорівневі інтерфейси візуалізації та велика кількість пакетів сторонніх розробників: seaborn, HoloViews, ggplot, а також набір інструментів Cartopy для проектування та створення карт. На Рисунок 1.1 зображено приклади візуалізацій значень температури. Рухаючи курсор, ми можемо переглядати, якій точці (x, y) відповідає ця позиція.

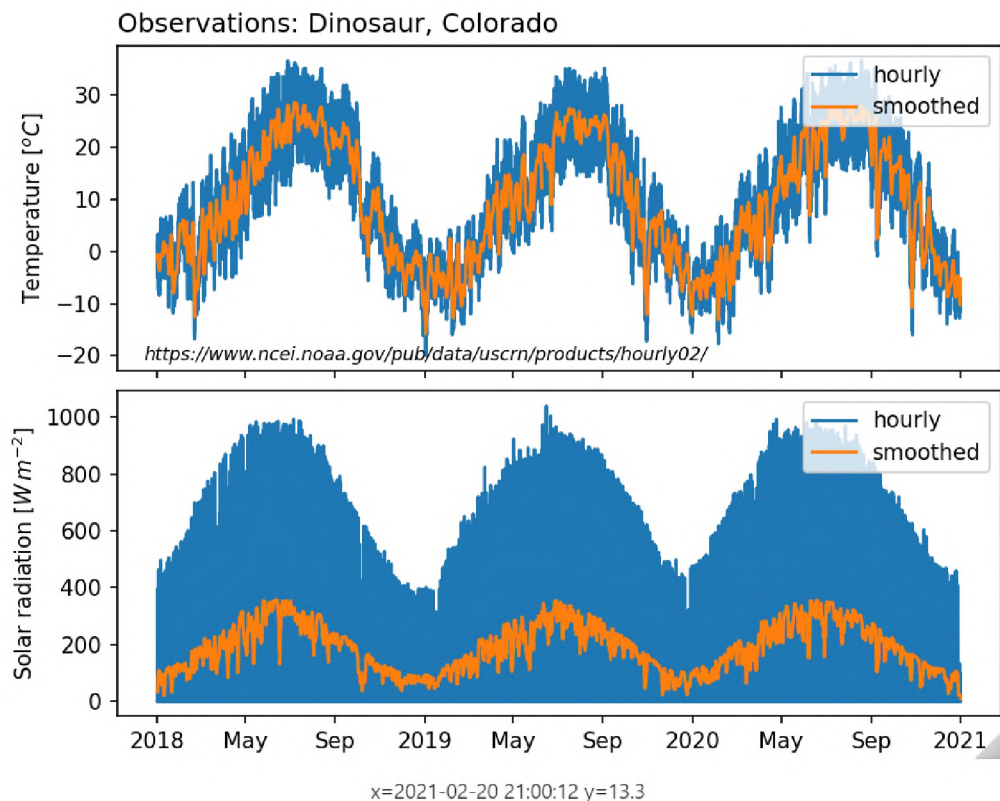


Рисунок 1.1 Приклад візуалізації Matplotlib

Візуалізація є ідеальною формою представлення даних, щоб зрозуміти, що відображають певні дані. Вона дає чітке уявлення про дані різних типів: що демонструють підходи. Але не для всіх даних потрібен однаковий формат представлення. Для цього в Matplotlib існують різні способи генерації візуальних елементів на основі даних. Шість найбільш часто використовуваних типів графіків в Matplotlib є:

- лінійний графік;
- графік у вигляді стовпців;
- діаграма розсіювання;
- кругова діаграма;
- діаграма із зонами;
- гистограма.

Лінійні графіки створюються шляхом з'єднання прямих ліній за точками даних у місцях перетину значень осей X і Y. Лінійні діаграми є найпростішою формою представлення даних. У Matplotlib графік цього типу (рис. 1.2) будується функцією `plot()`.

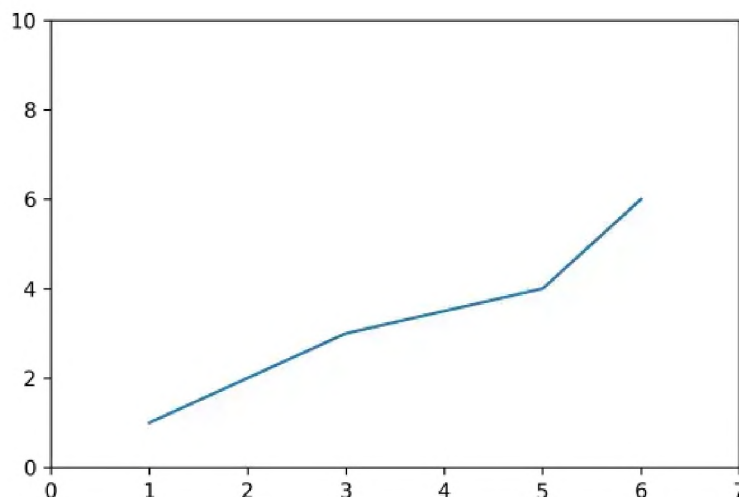


Рисунок 1.2 – Лінійний графік

Діаграми у вигляді стовпців – це вертикальні або горизонтальні прямокутні графіки, які використовуються для порівняння даних. Наприклад, їх зручно застосовувати для оцінки змін за період, представлений на іншій осі (зазвичай на осі X). Кожен стовпчик може зберігати значення одного або кількох наборів даних, поділених згідно з певним співвідношенням. Чим довшим є стовпець, тим більше значення він представляє. У Matplotlib існують функції `bar()` або `barh()` для створення графіків такого типу (рис. 1.3). За допомогою параметрів графіків можна змінювати вигляд діаграм, кольори, розміщення стовпців та інших елементів.

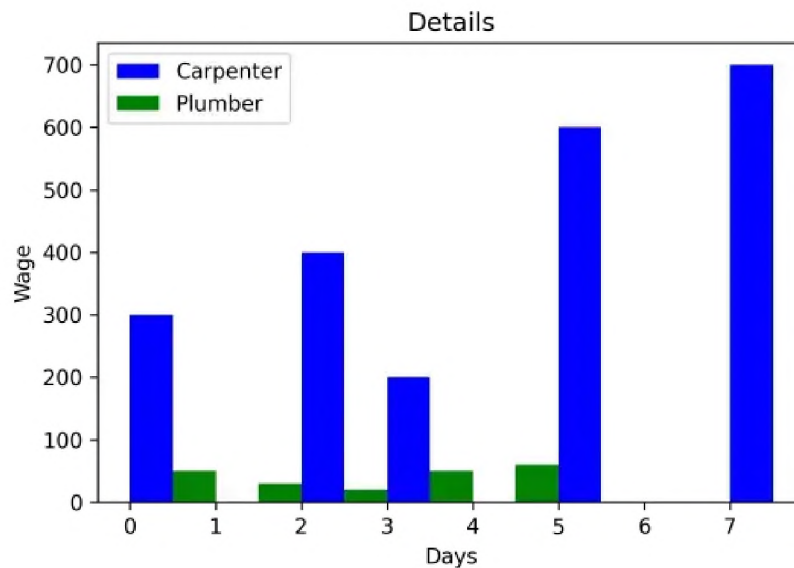


Рисунок 1.3 – Стовпчаста діаграма

Діаграми розсіювання використовуються для порівняння змінних, щоб визначити зв'язок між залежними та незалежними змінними. Дані представляються як набір точок. Метод `scatter()` у бібліотеці Matplotlib використовується для створення діаграми розсіювання даних, представлення зв'язку між змінними та того, як зміна однієї впливає на іншу.

Загалом Matplotlib включає: графіки з можливістю віддаляти або наближати зображення, інтерактивність, налаштування вигляду діаграм,

експорт у різні формати, можливість вбудови у блокноти Jupiter та графічні інтерфейси, інтеграцію сторонніх пакетів.

Іншим напрямком роботи з даними в Python є машинне навчання [11]. Для цього розроблені такі інструменти як scikit-learn і PyTorch. Scikit-learn пропонує різноманітні алгоритми для класифікації, регресії та кластеризації даних, що дозволяє створювати моделі машинного навчання для прогнозування ринкових тенденцій. PyTorch, у свою чергу, надає інструменти для роботи з нейронними мережами і глибоким навчанням, розширюючи можливості дослідження. [12]

Для представлення даних за допомогою Python зазвичай використовують такі методи:

- створення блокноту Jupiter;
- створення інформаційної панелі.

Перший метод передбачає використання Jupiter Notebook. Jupyter Notebook — це вебпрограма з відкритим вихідним кодом, яка дозволяє створювати документи з кодом. Блокноти Jupiter містять код, візуалізації та текстовий опис. Ці можливості вирізняють їх від звичайних файлів з кодом Python або кодом інших мов програмування. Ці файли підтримують лише можливість коментування коду. Блокноти Jupiter мають інструменти для одночасного виконання коду, який можна приховати, вбудови зображень та додавання текстового опису. Jupiter підтримує більше ста мов програмування. Підтримувані мови називаються ядрами. До популярних підтримуваних мов відносяться: Python, R, Matlab, Julia, Java та Scala. Охоплення документування для конкретної мови залежить від розміру її спільноти. Деякі ядра отримують надійну підтримку з великою кількістю активних учасників. Розробники популярних ядер отримують оплату за свій продукт. Підтримка менш популярних мов програмування забезпечується зусиллями невеликої команди або окремих розробників. Обираючи нове ядро, бажано знайти офіційні сторінки спільноти користувачів цього ядра та документацію. Це допоможе визначити, чи відповідає ядро бажаному рівню підтримки та цілям

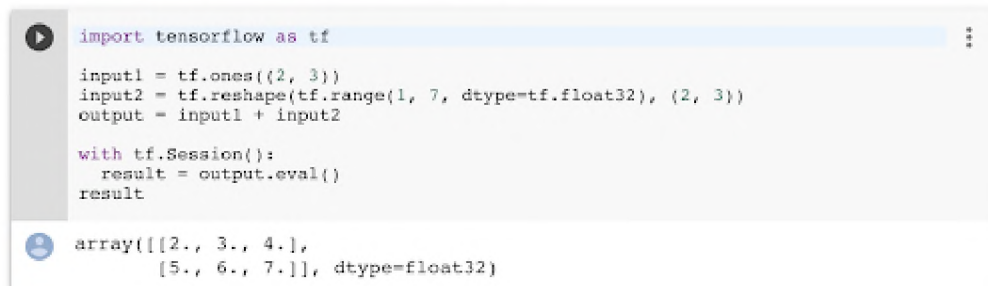
розробки. Для аналізу даних найкраще підтримуваними ядрами є Python і R. Ці мови програмування мають велику спільноту та детальну документацію для роботи в Jupiter Notebook.

Блокноти Jupiter складаються з комірок. Комірки бувають двох типів – текстові та комірки з кодом. Комірки з кодом мають три компоненти: область введення, область відображення та область виведення. Область введення позначається як In []. Число всередині дужок вказує на порядок виконаної комірки. Якщо воно відсутнє, то ця комірка ще не була виконана. Область відображення містить всі елементи, які створені в результаті роботи коду. Тут виводиться текст, створені графіки та зображення. Область виведення призначена для відображень значень, які повертаються в результаті виконання коду. Структуру комірки та її області показано на рис. 1.4.

▼ TensorFlow execution

Colaboratory allows you to execute TensorFlow code in your browser with a single click. The example below adds two matrices:

$$\begin{bmatrix} 1. & 1. & 1. \\ 1. & 1. & 1. \end{bmatrix} + \begin{bmatrix} 1. & 2. & 3. \\ 4. & 5. & 6. \end{bmatrix} = \begin{bmatrix} 2. & 3. & 4. \\ 5. & 6. & 7. \end{bmatrix}$$



```
import tensorflow as tf

input1 = tf.ones((2, 3))
input2 = tf.reshape(tf.range(1, 7, dtype=tf.float32), (2, 3))
output = input1 + input2

with tf.Session():
    result = output.eval()
    result
```

array([[2., 3., 4.],
 [5., 6., 7.]], dtype=float32)

Рисунок 1.4 – Области виведення в Jupiter Notebook

До інших можливостей Jupiter включають:

- адаптивність до різних мов програмування;
- використання кількох мов одночасно;
- стандартний інтерфейс для всіх мов;
- команди редактора – для виконання коду у певній послідовності, виконання всього коду з початку або починаючи з конкретного фрагменту;

- форматування за допомогою шрифтів;
- додавання гіперпосилань;
- LaTeX формули;
- вбудова зображень та зображень які генеруються або виводяться в результаті виконання комірок з кодом;
- виведення документації для команд і бібліотек шляхом виконання комірки з кодом, яка містить назву команди чи бібліотеки та знак питання (в результаті виконання `print?` Буде виведена документація для команди `print` мови Python);
- підтримка віджетів: шкала зміни певної величини, кнопки перемикання параметрів, текстові поля. Віджети доступні для деяких мов програмування та додають інтерактивності при роботі з графіками, картами тощо;
- підтримка особливих команд для зручності виконання комірок з кодом для підтримуваних мов. Такі команди призначені лише для виконання в Jupiter Notebook і стосуються пришвидшення та полегшення роботи з кодом. Особливі команди починаються зі знака `%`;
- виведення таблиць.

Отже, блокноти Jupyter є та універсальним середовищем для аналізу даних і наукових обчислень. Якщо в проєкті передбачено створення інформаційної панелі з коротким описом результатів дослідження, для побудови в Python необхідно виконати такі дії:

1. зібрати дані;
2. створити шаблон вебсторінки;
3. створити файли формату `.ru`, які описують алгоритми обробки та візуалізації даних;
4. передати зображення в файл HTML.

Такий підхід вимагає створення серверу мовою Python, який виконуватиме дії, описані у кроках 3 та 4. Розглянемо інструменти Python, які

необхідні для створення серверу, налагодження зв'язку сервера та HTML-сторінок. Популярним рішенням є Flask, мікروهбфреймворк, який широко використовується через його простоту та гнучкість. Перевагою Flask, зокрема для створення інформаційної панелі є швидкість налаштування. Flask був спеціально розроблений для максимально легкого використання. Встановлення Flask відбувається так само, як для будь-якого Python-пакету. Після імпорту пакету для створення серверу достатньо виконати код:

```
from flask import Flask

app = Flask(__name__)

@app.route("/")
def hello_world():
    return "<p>Hello, World!</p>"
```

На відміну від серверної мови програмування PHP, встановлення повноцінних сторонніх програм не вимагається. Мінімалістичність Flask дозволяє розробникам швидко запустити вебсервер. Незважаючи на невеликий розмір пакету, Flask забезпечує необхідні компоненти для веброзробки. Flask підтримує розширення, які можуть додавати функції за потреби, що робить його адаптованим до різних вимог проєкту. Розширення – це додаткові бібліотеки, які інтегруються в вебсервер для таких цілей:

- валідація форм;
- робота з базами даних. Flask не визначає, яку базу даних необхідно використовувати. Для зв'язку з БД необхідно встановити відповідне розширення;
- обробка завантаження файлів, робота з файлами;
- аутентифікація та інші розширення.

Структура сервера, побудованого з Flask, містить посилання, які представляють сторінки сайту. Маршрутизація у Flask налаштовується

декоратором `.route()`. Декоратор прив'язаний до посилання. Посилання та його частини можуть бути як статичними, так і динамічними. Flask дозволяє створювати URL-посилання для статичних файлів CSS-стилів та файлів Javascript. Flask-додаток може передавати змінні до HTML-структури. Цього можна досягти завдяки генерації Flask-шаблонів. Для візуалізації шаблонів HTML використовують метод `render_template()`. Всередині дужок вказують назву HTML-файлу та змінні у вигляді аргументів, які будуть передані до шаблону. Код нижче показує приклад того, як задати HTML-файл для посилання на корінь вебсайту.

```
app = Flask(__name__)
@app.route("/")
def hello_world():
    return render_template('hello.html', name=name)
```

Для роботи зі Flask HTML-файли змінюють та перетворюють на шаблон Jinja. Jinja — це механізм створення шаблонів, відомий своєю швидкістю та можливостями для розширення. Він дозволяє створювати в HTML-структурі заповнювачі з кодом, синтаксис якого схожий на Python. Шаблон приймає змінні з сервера для візуалізації HTML. Код всередині HTML необхідно розмістити у фігурних дужках. Комбінація інструментів Python для аналізу даних та вебфреймворку Flask дозволяє створити власну інформаційну панель, як альтернативу до представлення даних у Jupiter Notebook і сторонніх сервісів для аналізу даних. Активна спільнота Flask і інших пакетів Python надає розробникам доступну й детальну документацію. Простота та гнучкість Flask робить його чудовим рішенням для створення вебінтерфейсів для демонстрації інформаційної панелі.

1.4 Доступність відкритих джерел даних

Цей розділ присвячений розгляду доступності й надійності відкритих джерел даних, які будуть використані у дослідженні, зокрема для прогнозування кількості вакансій. Важливо враховувати, що існує чимало чинників, які можуть вплинути на надійність та об'єктивність даних [13]. На теперішній момент існують геополітичні обставини та інші чинники, які можуть вплинути на дані, що може викликати суб'єктивізм результатів дослідження або залежність від контексту. На цьому етапі ми детально вивчимо доступні джерела даних та їхню надійність, а також розглянемо можливі фактори, які можуть вплинути на їхню об'єктивність та використання у моделюванні прогнозів.

Прикладом фактору, який ми визначили раніше, є кількість осіб, які перетинають кордон України з моменту введення режиму воєнного стану у 2022 році. Ці дані можуть різнитися в залежності від того, яка сторона їх надає. У цьому дослідженні будемо використовувати дані, зібрані на порталі UNHCR (Управління Верховного комісара ООН у справах біженців) [14]. Міжнародні організації збирають інформацію в урядів одразу кількох країн, які межують з Україною, тому це джерело вважатимемо надійним.

Розглянемо джерела даних щодо кількості вакансій та поширення мов програмування. Важливо звернути увагу на те, чому ці дані мають меншу схильність до геополітичного впливу та неточностей, порівняно з іншими видами інформації.

По-перше, дані щодо кількості вакансій надходять безпосередньо від роботодавців та публікуються на професійних інформаційних ресурсах та дошках оголошень. Оскільки ці ресурси спеціалізуються на комерційних послугах для роботодавців, вони мають менший стимул спотворювати статистику для опублікованих даних. Роботодавцям вигідно надавати точну та достовірну інформацію, щоб залучити найкращих кандидатів.

По-друге, дані про поширення мов програмування, також доступні через спеціалізовані ресурси для програмістів та вебсайтах, де спеціалісти комунікують між собою. Такі ресурси як StackOverflow проводять опитування щодо використання мов програмування [15].

Ці дані також збираються компаніями, які виготовляють програмні продукти, орієнтовані на розробників. Програмісти та розробники в усьому світі використовують різні мови програмування незалежно від політичних подій.

Отже, інформація про поширення мов програмування має більш об'єктивний характер, оскільки вона відображає реальну практику розробки програмного забезпечення.

Збір цих даних із джерел, які спеціалізуються на вакансіях та програмуванні, дозволить знизити вплив геополітичних чинників, оскільки вони підпорядковані потребам ринку праці та технічним вимогам сучасності, а не політичним змінам.

Висновки до розділу 1

Ринок праці – це динамічна сутність, яка постійно розвивається, формується численними факторами, які окреслюють його контури. Повне розуміння змін і тенденцій ринку праці потребує ретельного вивчення цих факторів впливу. З точки зору соціології ринок праці тісно пов'язаний із суспільними структурами та процесами. Воєнні дії створили нові виклики для ринку праці, особливо у східних регіонах. Підприємства або переїхали, або зіткнулися з руйнуванням, що призвело до значного переміщення робочої сили. Через відтік населення деякі регіони відчували надлишок робочої сили, а інші лишилися робочої сили та значної бази підприємств. Це також вплинуло на відкритість та якість даних. Окрім цих конкретних подій,

загальні фактори, що впливають на кожну країну, включають технологічний розвиток.

Існує чималий вибір засобів для вирішення задач цієї роботи. З-поміж інших обрано мову програмування Python та основні бібліотеки, такі як Pandas для роботи з даними та Matplotlib для візуалізації. NumPy забезпечує ефективну обробку числових даних. Окрім цього в рамках суміжних задач використовуються бібліотеки scikit-learn і PyTorch. З їх допомогою можна розробляти моделі машинного навчання. Саме робота із нейронними мережами і машинним навчанням вирізняє Python від інших мов програмування. Цей інструментарій робить можливим глибокий аналіз та прогнозування тенденцій на ринку IT-вакансій.

РОЗДІЛ 2

МЕТОДОЛОГІЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ ЗА ДОПОМОГОЮ ІНСТРУМЕНТІВ PYTHON

2.1 Збір і підготовка даних

У цьому розділі виконаємо збір даних за допомогою кількох методів. Визначимо дані, які необхідно зібрати:

- популярні запити – Jooble;
- дані про перетин кордону;
- дані про кількість IT-вакансій.

Спершу виконаємо збір даних про популярні запити по всій Україні. Для цього використаємо статистику дошки оголошень Jooble для України. За допомогою WayBackMachine (рис. 2.1) отримаємо дані за різні періоди. [16]

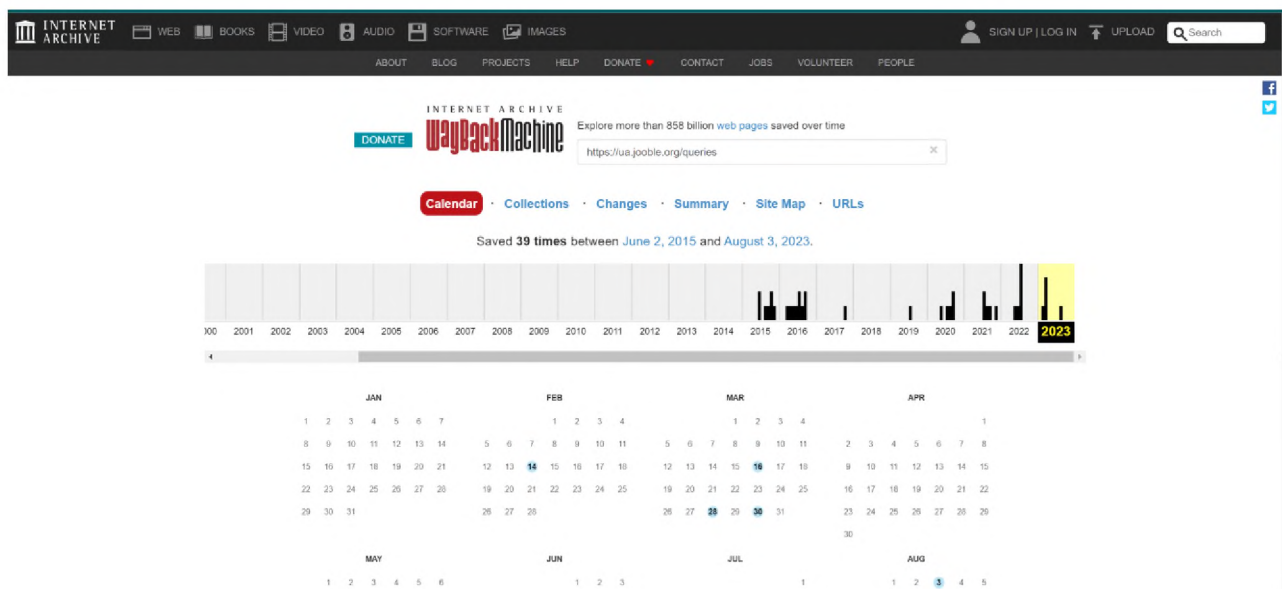


Рисунок 2.1 – Вибір періоду збереженої сторінки

Для збору текстових даних з Jooble використаємо парсинг даних. Парсинг даних є широко використовуваним методом структурування даних.

Процес парсингу включає перетворення одного типу даних в інший. Для збору даних із вебсторінок програма приймає дані – необроблену HTML-розмітку. Отримавши вхідні дані, програма відокремить необхідну частину HTML і перетворить її в потрібний формат даних, який можна легко прочитати та зрозуміти. Програма-парсер містить інструкції у вигляді коду, які визначають дії щодо HTML. Відповідно до вказаних правил парсер обирає необхідну інформацію та зберігає дані в інших форматах, наприклад у JSON, CSV, Excel, у текстовому файлі та інших. Формати вхідних та вихідних даних можуть бути різними, в залежності від потреб і цілей розробників. Парсери можна створювати самостійно або купити доступ до готової програми. Наразі парсинг використовують для таких технологій:

- мови програмування;
- HTML і XML;
- SQL та подібні мови;
- мови моделювання;
- мови сценаріїв;
- HTTP та інші Інтернет-протоколи.

Перевагами створення власного парсеру є:

- відсутність функціональних обмежень. Програма-парсер, створена для потреб конкретного проєкту або організації, міститиме всі потрібні функції;
- розробка власного парсеру зазвичай дешевша, ніж готові комерційні програми;
- свобода рішень щодо майбутніх оновлень. Власний парсер можна змінити на власний розсуд.

Недоліками створення власного парсеру для аналітиків є:

- необхідність в наявності навичок створення подібних програм або спеціаліста, який володіє цими навичками;

- підтримка роботи програми власними зусиллями. Для оновлення парсеру потрібно витратити зусилля учасників проекту, що спричинить збільшення вартості робіт;
- можливі витрати на сервер. В залежності від цілей проекту може знадобитися швидке середовище обробки даних;
- необхідність співпраці з технічною командою, яка розробляє парсер. Для узгодження рішень щодо оновлень, необхідного функціоналу та проведення тестування потрібно виділяти додатковий час.

Якщо використовувати сторонній парсер, то аналітики отримують такі переваги:

- відсутність необхідності витратити кошти на додаткові людські ресурси. Оскільки готова програма підтримується стороннім розробником, то до проектних робіт не буде входити вартість створення власного парсера, а також обслуговування, оновлення парсера та підтримка серверів. Також немає необхідності проводити тестування власного парсеру;
- технічна підтримка розробника. Користувачі готових рішень можуть розраховувати на технічну підтримку розробників парсерів. Вони зможуть швидко вирішувати проблеми, бо розробники комерційних парсерів мають досвід та добре знають свій продукт;
- з доступом до комерційного парсеру ймовірність того, що він не буде працювати належним чином, менша. Відомий комерційний продукт вже пройшов внутрішнє та зовнішнє тестування. Він повністю відповідатиме вимогам ринку.

Отже, вибір інструментів для збору даних за допомогою парсингу залежить від таких факторів: цілі та масштаби проекту, бюджет проекту, а також від даних, які необхідно зібрати. Зважаючи на те, що в даному

дослідженні необхідно збирати дані зі специфічних ресурсів, то найкращим рішенням є створення власних парсерів мовою Python.

Для збору даних, які потребують роботи з HTML, використаємо функції для парсингу даних. Функція `def extract_links_text_and_write_to_file(url, div_id, output_filename)` призначена для сторінок вебсайту Jooble. Щоб дізнатися, який елемент потрібно знайти на сторінці, необхідно скористатися інструментами для веброзробників у браузері. Для цього можна використати команду CTRL + SHIFT + I або перейти до меню «Додаткові інструменти» та «Інструменти розробника». Далі необхідно вибрати будь-який елемент, який відповідає ключовому слову. Для прикладу вибираємо слово «охоронник» (рис. 2.2).



Рисунок 2.2 – Вибраний елемент на сторінці Jooble з пошуковими запитами

Натиснувши на елемент, отримуємо структуру HTML-сторінки. На рис. 2.3 зображено структуру сторінки з популярними ключовими запитами. Усі потрібні елементи розміщені всередині тега гіперпосилання `<a>`. Для парсингу ключових слів необхідно зберегти усі елементи всередині елемента `div` із ідентифікатором `id="divlinks"`. Оскільки пошукові запити – це значення атрибуту `title` гіперпосилання `<a>`, то для кожного гіперпосилання треба виконати операцію збереження значення атрибуту `title`.

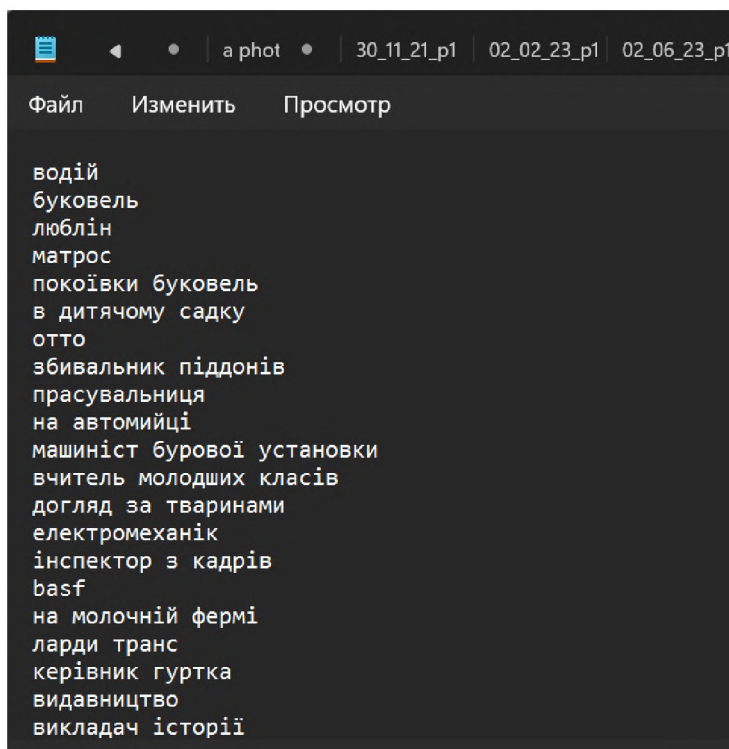


Рисунок 2.4 – Файл з пошуковими запитами

Для збору даних за інші періоди необхідно повернутися на сторінку, показану на рисунок 2.1 і обрати нову дату. Інтернет-архів WayBack не завжди зберігає всі сторінки. В залежності від популярності сайту можуть бути доступні не всі періоди. Хоча для дошки оголошень Jooble деякі дати збережень знімків сторінок відсутні, наявних дат достатньо для проведення дослідження. WayBack Machine відображає кожен сторінку збережених сайтів з посиланням, яка має власний ідентифікатор. Тому сторінка та її посилання є унікальними (рис. 2.5). Це дозволяє звертатися до вмісту вебсторінки за допомогою функції `extract_links_text_and_write_to_file(url, div_id, output_filename)`. Функція дозволяє повторно використовувати код, змінюючи лише параметри – посилання й назву вихідного файлу.

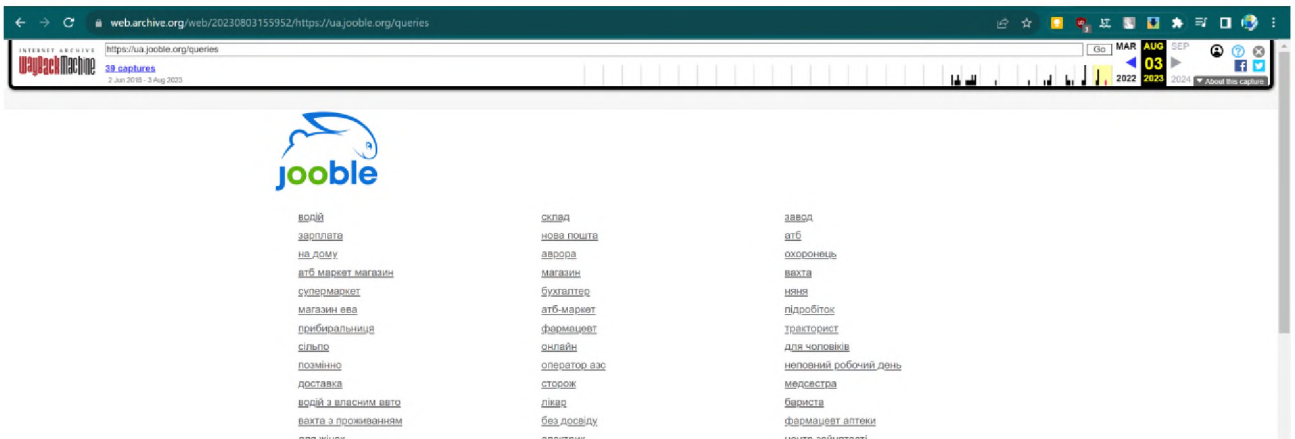


Рисунок 2.5. – Приклад обраної сторінки за датою збереження

В результат виконання функції маємо текстові файли, назви яких відповідають періодам збереження знімку сторінки з ключовими словами:

02_02_23_p1.txt

02_06_23_p1.txt

02_07_22_p1.txt

02_11_23_p1.txt

03_08_23_p1.txt

12_09_21_p1.txt

13_08_21_p1.txt

20_06_22_p1.txt

30_05_22_p1.txt

30_11_21_p1.txt

Отримані файли з ключовими словами будуть оброблені з метою визначення тенденцій пошукових запитів та змін у запитах, які стосуються ІТ-професій.

Наступним етапом є збір даних з порталу UNHCR [17]. Ці дані ілюструють кількість в'їздів та виїздів закордон. Збережемо дані з сайту в форматі xls (рис. 2.6).

	A	B	C	D	E	F
1	Дата	Виїзд, млн	В'їзд, млн	Сальдо		
2	01.08.2022	8.31	4.28	4.03		
3	01.09.2022	9.67	5.46	4.21		
4	01.10.2022	11.27	6.45	4.82		
5	01.11.2022	12	7.36	4.64		
6	01.12.2022	13.05	8.18	4.87		
7	01.01.2023	14.2	9.13	5.07		
8	01.02.2023	15.33	9.98	5.35		
9	01.03.2023	16.24	10.67	5.57		
10	01.04.2023	17.25	11.56	5.69		
11	01.05.2023	18.34	12.53	5.81		
12	01.06.2023	19.46	13.41	6.05		
13	01.07.2023	20.8	14.53	6.27		
14	01.08.2023	22.29	15.73	6.56		
15						
16						
17						

Рисунок 2.6 – Дані про перетин кордону України

Останнім етапом є збереження даних про кількість ІТ-вакансій в Україні. У цьому випадку ми не будемо використовувати файли, оскільки доступ до даних можна отримати через API від UAdata. [18]

Для доступу до словника даних потрібно використати посилання на json-файл: <https://uadata.net/it-industry/programming-languages.json>. Дані знаходяться в об'єкті data:

```
"data": [
  {
    "at": "2022-05-20",
    "val": 15916
  },
  {
    "at": "2022-05-21",
    "val": 15633
  },
  ...
```

Для роботи з даними необхідно використовувати ключі «at» та «val» ключа «data». Так ми отримаємо доступ до дати та значення – кількості вакансій у певний день.

2.2 Вид і структура даних

У цьому розділі розглядається структура даних, а також процес систематизації зібраних даних. Цей процес називається очищенням і підготовкою даних. Підготовка даних — це процес, який включає збір, об'єднання, попередню обробку даних, профілювання, очищення, перевірку, перетворення, структурування та впорядкування даних. Підготовлені дані можна використовувати в програмах для аналітики та візуалізації даних.

Одна з головних цілей підготовки даних полягає в обробці зібраних даних. Необхідно переконатися, що дані точні та послідовні, а їх структура не спричинить недійсних висновків. Часто дані містять відсутні значення, неточності, інші помилки. Окремі набори даних можуть мати несумісні формати, типи даних або потребують об'єднання. Тому ідентифікація помилок і невідповідностей у даних, перевірка якості даних і консолідація наборів даних є невід'ємним етапом будь-якого проєкту.

У контексті цього дослідження ретельна перевірка структури отриманих даних є обов'язковою, не зважаючи на авторитетність джерел. Розуміння структури даних — це фундаментальний крок, який лежить в основі валідності та надійності результатів аналізу даних. Ретельний аналіз структури даних служить основою для отримання значущої та релевантної інформації. За відсутності всебічного розуміння структури даних подальші аналітичні процеси можуть виявитися непереконливими або оманливими.

Для виконання аналізу даних за допомогою інструментів Python розглянемо структуру необхідних файлів. Файли, які містять ключові слова, були зібрані за допомогою функції-парсера даних. Вони представлені в

форматі txt (текстовий файл). Цей набір даних із дошки вакансій Jooble містить набір запитів, пов'язаних із пошуковими запитами щодо різних вакансій по всій Україні. Інформація охоплює різні періоди, забезпечуючи розуміння динаміки пошуків, пов'язаних із працевлаштуванням. Кожен файл відповідає за певний період і містить по одному ключовому слові на кожному рядку. Формат txt найкраще відповідає типу зібраних даних. Оскільки дані зібрані за конкретні місяці, вони повинні бути об'єднані за періодами, до яких відносяться. Зібрані дані містять ключові слова від 03.08.21 до 02_11_23, тому маємо три періоди дослідження – перед воєнними діями, 2022 рік, 2023 рік. Для цього ключові слова об'єднані у трьох файлах, які також мають розширення txt і по одному слові на рядку. Отже, маємо файли (рис. 2.7):

before.txt – до 02.2022;

then.txt – 2022 рік;

after.txt – 2023 рік.

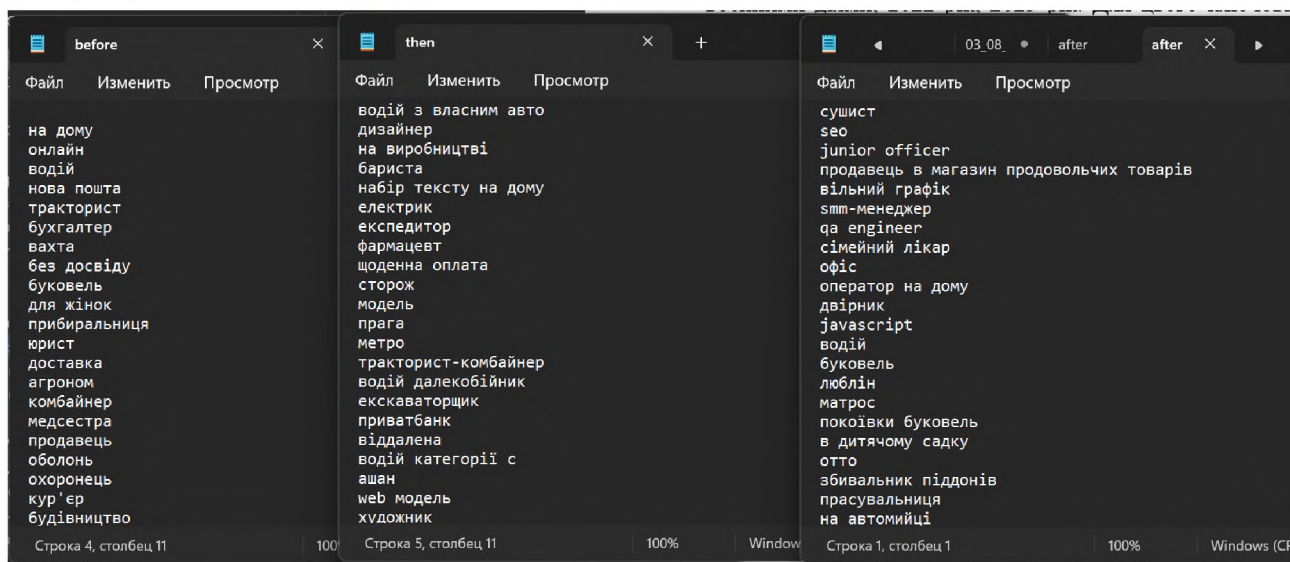


Рисунок 2.7 – Файли з ключовими словами

Наступний файл має назву border.xls. Цей набір даних, отриманий з порталу ООН, описує кількість осіб, які покинули територію України, кількість осіб, які в'їхали, та різницю між цими значеннями. Дані збережені у

структурованому форматі файлу Excel (рис. 2.8), що дозволить систематично дослідити тенденції змін кількості населення після початку воєнних дій. Дані розділені на чотири стовпці. Три останні стовпці мають числовий тип значення. Варто зауважити, що правильність форматування вкрай важлива, оскільки matplotlib та інші бібліотеки дуже чутливі до формату даних і можуть неправильно відображати дані, якщо їх формати не відповідають даним. Перший стовпець містить дані формату дати. Це період спостереження. Числові дані у стовпцях Виїзд, В'їзд та Сальдо показують кількість осіб у мільйонах. Кожне число більше попереднього, бо воно показує не кількість осіб за кожен місяць, а сумарну кількість на заданий момент часу. Дані в таблиці на рис. 2.8 показують переміщення загалом через всі кордони в бік Європейського Союзу, без врахування осіб, які перетнули кордон із Росією та Білоруссю. Тому можемо вважати, що реальна кількість осіб, які з тих чи інших причин покинули Україну, може бути більшою.

	A	B	C	D	E
1	Дата	Виїзд, млн	В'їзд, млн	Сальдо	
2	01.08.2022	8.31	4.28	4.03	
3	01.09.2022	9.67	5.46	4.21	
4	01.10.2022	11.27	6.45	4.82	
5	01.11.2022	12.00	7.36	4.64	
6	01.12.2022	13.05	8.18	4.87	
7	01.01.2023	14.20	9.13	5.07	
8	01.02.2023	15.33	9.98	5.35	
9	01.03.2023	16.24	10.67	5.57	
10	01.04.2023	17.25	11.56	5.69	
11	01.05.2023	18.34	12.53	5.81	
12	01.06.2023	19.46	13.41	6.05	
13	01.07.2023	20.80	14.53	6.27	
14	01.08.2023	22.29	15.73	6.56	
15					
16					

Рисунок 2.8 – Файл з даними про кількість осіб, які перетинали кордон України з сусідніми країнами

Формули, використані для розрахунків, показано на рисунку 2.9. Перший стовпець представлений числами, бо він містить дані формату «Дата», який коректно відображається в звичайному режимі.

	A	B	C	D
1	Дата	Виїзд, млн	В'їзд, млн	Сальдо
2	44774	8.31	4.28	=B2-C2
3	44805	9.67	5.46	=B3-C3
4	44835	11.27	6.45	=B4-C4
5	44866	12	7.36	=B5-C5
6	44896	13.05	8.18	=B6-C6
7	44927	14.2	9.13	=B7-C7
8	44958	15.33	9.98	=B8-C8
9	44986	16.24	10.67	=B9-C9
10	45017	17.25	11.56	=B10-C10
11	45047	18.34	12.53	=B11-C11
12	45078	19.46	13.41	=B12-C12
13	45108	20.8	14.53	=B13-C13
14	45139	22.29	15.73	=B14-C14

Рисунок 2.9 – Формули файлу з даними про кількість осіб, які перетинали кордон України з сусідніми країнами

Останній файл має назву `monthly_median_data.csv` (рис. 2.10). Дані збережені у форматі CSV. Файл CSV (значення, розділені комами) — це особливий тип файлу, який можна створювати та редагувати в Excel. Замість того, щоб зберігати інформацію в стовпцях, дані у CSV-файлах зберігаються у рядках, але в одному рядку вони розділені комами. Це популярний формат файлів, який може зчитатися багатьма програмами. Коли текст і числа зберігаються у файлі CSV, дані можна легко експортувати з однієї програми в іншу. Наприклад, контакти в обліковому записі Google зберігаються у форматі CSV, тому їх можна імпортувати до програмного продукту Outlook від Microsoft.

	A
1	date,value
2	2022-05-31,15867.5
3	2022-06-30,15458.0
4	2022-07-31,14366.0
5	2022-08-31,13588.0
6	2022-09-30,13569.0
7	2022-10-31,13621.0
8	2022-11-30,12519.5
9	2022-12-31,10503.0
10	2023-01-31,9208.0
11	2023-02-28,9982.0
12	2023-03-31,9786.0
13	2023-04-30,9643.0
14	2023-05-31,9853.0
15	2023-06-30,9325.5
16	2023-07-31,9549.0
17	2023-08-31,7600.0
18	2023-09-30,7118.5
19	2023-10-31,7218.0
20	2023-11-30,7172.0
21	

Рисунок 2.10 – Вміст файлу про медіанну кількість вакансій в сфері ІТ

Різноманіття форматів файлів, які досліджуються включає CSV, XLS і txt. Хоча формати даних різні, вони сумісні з різними аналітичними інструментами, які будуть використані для їх обробки та візуалізації. Розглянемо особливості використання цих форматів з інструментами Python. Для зчитування файлів формату CSV достатньо використовувати функцію `pd.read_csv()` із бібліотеки `pandas`, щоб прочитати вміст файлу CSV у `DataFrame`. Для файлів формату `xls` від Microsoft Excel і `txt`-файлів також необхідно впевнитися, що встановлено `pandas`.

Текстові файли містять дані з символами кирилиці, що може спричинити проблеми зчитування символів. У таких мовах програмування, як C, немає різниці між числовими типами даних і текстовим рядком. Рядки – це просто масиви байтів зі значенням у кінці NULL. Єдина відмінність полягає в тому, що якщо намагатися відобразити їх за допомогою

специфікаторів форматування, числа між 65 і 90 будуть виводитися як великі англійські літери. Те саме відбувається і при роботі з текстом у Python — також повертається значення `int`. Для вирішення проблем із виведенням тексту на HTML-сторінці необхідно використовувати шрифт, який підтримує символи кирилиці. Для цього необхідно вказати шлях до шрифту.

Для коректного виведення зображення в HTML дані зображення хмари слів в `img_bytes` потрібно закодувати за допомогою Base64, а потім декодувати в рядок формату UTF-8 (`img_base64`). Цей рядок можна використовувати як джерело для зображення в HTML. Щоб розмістити зображення в HTML, можна використати згенерований рядок Base64 і помістити його у тег HTML ``, наприклад:

```

```

Такий підхід дозволяє включити зображення хмари слів безпосередньо в HTML, гарантуючи, що під час процесу генерації зображення кириличні символи будуть відображаються правильно за допомогою відповідного шрифту.

Оскільки файли містять дані, які представляють формати дат, то для роботи з цим типом даних необхідно імпортувати `matplotlib.dates`. Це модуль у бібліотеці `matplotlib`, який надає можливості для роботи з датами й часом на графіках і візуалізаціях. Це особливо корисно під час роботи з даними часових рядів, де вісь X представляє шкалу часу. Коли файл містить дані про час або дату, то відображення їх у зручному для читання форматі необхідно скористатися модулем `matplotlib.dates` для форматування дати. Так можна змінити вигляд дат на графіку.

2.2 Обґрунтування моделі для прогнозування

В даному дослідженні ми використовуємо методи для прогнозування числових даних, зміна яких залежить як від змінної часу, так і від можливого

впливу інших факторів. Якщо дані не містять категоріальних значень, то найкращими методами можуть стати аналіз часових рядів та регресійні моделі. Обидва підходи надають можливість визначити тенденції та залежності в даних, що дозволяє робити прогнози для майбутніх значень.

Оцінка взаємозв'язків між залежною змінною та однією або кількома незалежними змінними називається регресією. Регресійний аналіз — це сукупність статистичних методів, які використовуються для оцінки зв'язків між залежною змінною та однією або кількома незалежними змінними. Його можна використовувати для оцінки сили зв'язку між змінними та для моделювання майбутнього зв'язку між ними. [19].

До регресійного аналізу включають кілька видів регресії (рис. 2.11): лінійну, багатолінійну і нелінійну. Найпоширенішими моделями є прості лінійні та багатолінійні. Нелінійний регресійний аналіз зазвичай використовується для складніших наборів даних, у яких між залежними та незалежними змінними існує нелінійний зв'язок. [20]

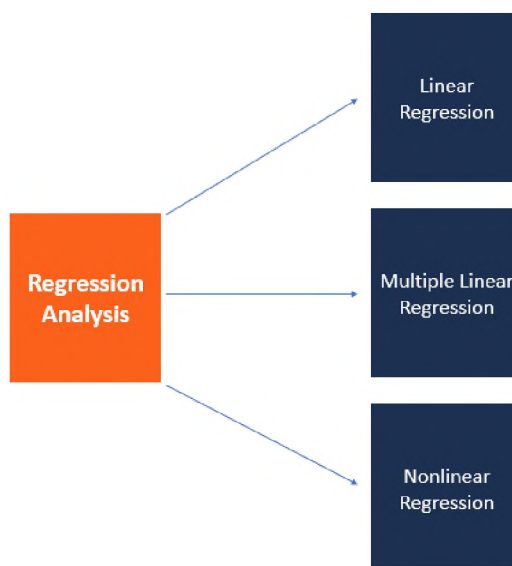


Рисунок 2.11 – Види регресійного аналізу

Проста лінійна регресія – це статистичний метод, який використовується для моделювання зв'язку між двома змінними – залежною

змінною (ту, яку ми хочемо передбачити) і незалежною змінною (ту, яка використовується для прогнозування). Основна ідея простої лінійної регресії полягає в тому, щоб знайти найкращу пряму лінію (лінійне рівняння), яка мінімізує суму квадратів різниць між спостережуваними та прогнозованими значеннями.

У контексті нашого дослідження припустимо, що ми маємо залежну змінну Y , яка представляє кількість пропозицій роботи, і незалежну змінну X , яка може бути значним фактором, що впливає на пропозиції роботи, наприклад, загальна економічна ситуація. Просте рівняння лінійної регресії набуває вигляду $Y = b_0 + b_1X$. Застосувавши лінію регресії до даних, ми отримуємо уявлення про силу та спрямованість зв'язку. Коефіцієнт b_1 вказує на зміну залежної змінної для зміни незалежної змінної на одну одиницю. Статистичні методи, такі як найменші квадрати, зазвичай використовуються для оцінки коефіцієнтів і прогнозування.

Коли зв'язок між залежною змінною та кількома незалежними змінними потребує дослідження, то використовують множинну лінійну регресію. Цей метод розширює принципи простої лінійної регресії для обробки складніших сценаріїв, де на результат впливають кілька факторів. Тоді рівняння регресії має вигляд $Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \dots + b_n \cdot X_n$. Змінна Y позначає залежну змінну, а X – це фактори, на основі яких будується залежність з Y .

Часові ряди відображають зміни значень в часі. Ми використовуємо цей підхід, коли нам потрібно прогнозувати майбутні значення для однієї змінної на основі її минулих значень. Методи аналізу часових рядів включають розглядання трендів, сезонності та випадкових відхилень. Швидко визначити наступні значення можна за допомогою інструментів прогнозування в Excel. За допомогою вбудованих алгоритмів, таких як Exponential Smoothing та ARIMA (Autoregressive Integrated Moving Average), ми можемо отримати прогнози для майбутніх значень. Для побудови такого

прогнозу необхідно визначити зміну часу та числові дані для прогнозування. Лист прогнозів у Excel не приймає додаткових стовпців. [21]

Часовий ряд передбачає, що майбутні значення залежать від попередніх, і не враховує інших факторів, які можуть впливати на змінну.

2.3 Розробка програмних інструментів на Python для обробки та аналізу даних

У цьому розділі виконано побудову інформаційної панелі, який утворений за допомогою діаграм та текстового пояснення до них. Інформаційна панель (дешборд) – це програма, яка показує інформацію та статистику про роботу певного пристрою, явища тощо [22].

У попередньому розділі описано процес збору та підготовки даних про найпопулярніші запити по всій Україні. Усі файли містять найпопулярніші дані з першої сторінки за кожен період. Скористаємося можливостями текстового аналізу штучного інтелекту ChatGPT. Для цього необхідно надіслати вміст файлів та спеціальний запит: «Наступні текстові дані представляють ключові слова для найпопулярніших пошукових запитів на дошках вакансій в Україні. Грунтуючись лише і виключно на цих ключових словах і відповідних датах, виконай аналіз і надай таку інформацію: 1) Як змінилися вподобання шукачів до 02.2022 та після, зважаючи на початок воєнних дій того місяця та року. Які типи професій були найпопулярніші? 2) Розкажи про зміни в пошуках, які стосуються IT та пов'язаних професій». Щоб отримати точні результати, ШІ необхідно описати завдання найточніше, а також вказати, що аналіз необхідно проводити лише на основі пошукових запитів. У результаті маємо відповідь.

До повномасштабних воєнних дій (2021) існував різноманітний спектр пошукових запитів. Популярними запитами є: «бухгалтер», «кур'єр», «бариста», «продавець» і «психолог».

Терміни, пов'язані з «ІТ», не були популярними у цей період, вказуючи на те, що ІТ-поле не було основним фокусом у цих пошуках.

Після початку бойових дій в Україні (2022 і далі) спостерігалися певні зміни в пошукових запитах.

Більш помітними стали ключові слова, пов'язані з безпекою і захистом, такі як «охоронець» і «доглядальниця». Це говорить про зміну ринку праці, оскільки люди могли шукати більше позицій, пов'язаних з темою безпеки.

Спостерігається збільшення пошуків на тему «ІТ», таких як «програміст», «системний адміністратор».

Для візуалізації ключових слів скористаємося методом «хмари слів». Для цього необхідно з'єднати файли за періодами до 01.02.2022, 2022 рік та 2023 рік. Щоб створити хмару слів, визначимо функцію `generate_word_cloud(text)`, яка приймає текст та повертає об'єкт `wordcloud_img`. Виконавши функцію тричі, отримаємо три зображення. Варто зазначити, що більш помітними будуть слова, кількість яких більша. Таким чином можна побачити, які нові запити додалися, порівняно з попередніми періодами [23]. Це також дозволить порівняти висновки, зроблені штучним інтелектом з фактичним представленням ключових слів у хмарі після аналізу. Для побудови візуалізацій і взаємодії Matplotlib і HTML необхідно встановити та імпортувати бібліотеку `mpld3`. Це бібліотека Python, яка забезпечує зв'язок між Matplotlib, для побудови графіків і D3.js, потужною бібліотекою JavaScript для створення інтерактивних візуалізацій даних у веббраузерах. Назва «`mpld3`» розшифровується як «Matplotlib D3», що вказує на трансформацію графіків Matplotlib у D3.js для створення інтерактивних вебвізуалізацій. За допомогою `mpld3` можна генерувати код HTML і JavaScript для візуалізацій Matplotlib. Цей код можна вставляти у вебсторінки, дозволяючи користувачам взаємодіяти з ними. Зображення хмар слів передається в HTML через змінну, визначену в коді Python. Оскільки для створення серверу мовою Python використовується вебфреймворк Flask, то весь код має бути огорнутий у декоратор:

```
@app.route('/')  
def index():  
    return render_template()
```

Визначимо змінні, які мають передаватися до HTML-шаблону:

- `chart_html` – змінна, яка містить графік кількості осіб, які перетинали кордон;
- `wordcloud1_img` – змінна, яка містить зображення графіку хмари слів;
- `wordcloud2_img` – змінна, яка містить зображення графіку хмари слів;
- `wordcloud3_img` – змінна, яка містить зображення графіку хмари слів;
- `monthly_offers_html` – змінна, яка містить зображення графіку змін кількості IT-вакансій в Україні;

Для функціонування додатку необхідно встановити необхідні пакети та бібліотеки:

- `from flask import Flask, render_template` – бібліотека Flask, яка відповідає за роботу сервера, функціонування Jinja шаблонів у HTML;
- `import matplotlib.pyplot as plt` – бібліотека для візуалізації даних;
- `import mpld3` – допоміжна бібліотека для зв'язку з matplotlib;
- `import pandas as pd` – бібліотека для роботи з даними;
- `import matplotlib.dates as mdates` – частина matplotlib для роботи з форматами даних «Дата»;
- `from wordcloud import WordCloud` – бібліотека для створення хмари слів;
- `import io` – базова бібліотека Python. Необхідна для роботи з файлами;

- `import base64` – модуль для кодування бінарних даних. Використовується для конвертації зображень;
- `import requests` – модуль для обробки HTTP-запитів до сервера. Дозволяє створити функціонал для обміну повідомленнями між браузером та сервером.

Кожна хмара слів в інформаційній панелі – це зображення, яке передається Flask до HTML-шаблону через змінну `wordcloud_img` з кодом зображення. Усього маємо три зображення. Кожне з них розміщено в елементі `div` згідно з унікальним ідентифікатором:

```
<div class="col-lg-8 mt-3" id="wordcloud2">
  
</div>
```

Всередині тегу зображення (`img`) маємо дужки з кодом `{{ wordcloud2_img }}`. Цей фрагмент є кодом модуля шаблонів Jinja. Всередині дужок змінна, значення якої було визначене в коді Flask-сервера:

```
return render_template('index.html', chart_html=chart_html,
                       wordcloud1_img=wordcloud1_img,
                       wordcloud2_img=wordcloud2_img,
                       wordcloud3_img=wordcloud3_img,
                       monthly_offers_html=monthly_offers_html,
                       )
```

Розглянемо першу хмару слів. Вона відповідає періоду до 01.02.2022. Його можна описати як період панування COVID та період до розгортання масштабних воєнних дій. Особливо помітними є пошукові запити «на дому», «водій», «онлайн», «бухгалтер», «для жінок», «психолог» «з проживанням», «пошта», «набір тексту», «далекобійник». Ми не спостерігаємо особливого попиту на ключові слова для пошуку вакансій у сфері IT. У цей період були популярні різні професії. Особливого переважання конкретних професій чи груп вакансій немає. Висновок на основі хмари ключових слів відповідає

а отже й кількості товарів. Але варто пам'ятати, що слово «водій» не втратило популярності. У даний період з'явилися запити на професію «охоронець», що може свідчити про прагнення знайти роботу в цій сфері та збільшення попиту на захист майна. Зменшилася позиція слова «тракторист». Причиною може бути погіршення ситуації в аграрному секторі, зменшення площі посівів та їх руйнація через воєнні дії у 2022 році. Бачимо досить помітне слово «онлайн», що показує прагнення до переведення робочих процесів в онлайн-простір. Висновок, зроблений на основі хмари слів, частково відповідає висновку, який зробив ШІ-чатбот. Він не зміг проаналізувати ключові слова так само докладно, як це можна зробити самостійно завдяки хмарам слів. Збережемо опис біля зображення хмари (рис. 2.13).

Особливо помітними є: «на дому» та «водій», як і в попередньому періоді. Помітними є «далекобійник», «пошта», «для жінок і чоловіків», «з власним автомобілем» «з проживанням», «пошта», «набір тексту», хоча набір тексту став менш популярним. Ми не спостерігаємо попиту на ключові слова, пов'язані з IT. У цей період також були популярні різні професії. Помітним стало слово «журналіст». Це може свідчити про попит на цю роботу через воєнні дії та необхідність освітлення подій. Запит на власне авто став більш помітним. Це може бути пов'язане зі змінами в менеджменті ризиків у період воєнних дій. Зменшилася популярність слова «Далекобійник» став менш помітним. Причиною можуть бути: порушення ланцюгів торгівлі з іншими державами, складність виїзду за кордон, зменшення кількості населення в Україні, а отже й кількості товарів. З'явилися

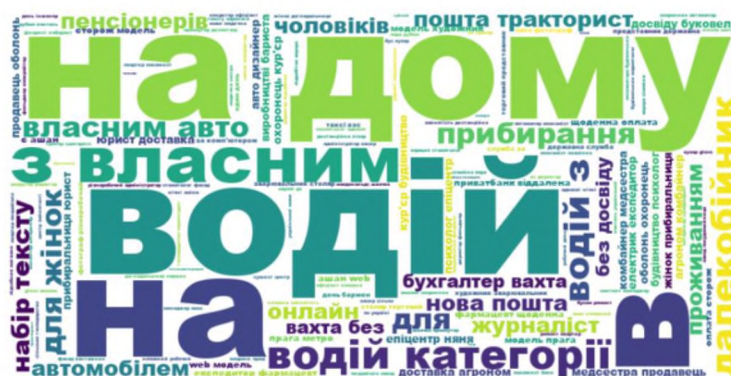


Рисунок 2.13 – Хмара ключових слів за 2022 рік

Розглянемо останній блок. Він містить третю хмару слів. Період дослідження – 2023 рік (рис. 2.14). У цей період одразу бачимо зростання кількості запитів, пов'язаних з IT-професіями. Нові запити, які стосуються сфери IT, включають: QA, engineer, developer, junior. «Водій» так само залишається популярним пошуковим запитом. Іншими популярними запитами є лікар, менеджер, вахта, оператор, санітарка. У цей період зросла частота використання таких слів: лікар, учитель, стоматолог, магазин, Атб,

прибирання, фармацевт, медична сестра. Дещо підвищилася значимість слів «підробіток», «з проживанням». Ми зовсім не бачимо слова «далекобійник», яке раніше було популярним. Збережемо опис зліва від зображення хмари слів (рис. 2.14).

Особливо помітними є: «на дому» та «водій», як і в попередньому періоді. Помітними є «далекобійник», «пошта», «для жінок і чоловіків», «з власним автомобілем» «з проживанням», «пошта», «набір тексту», хоча набір тексту став менш популярним. Ми не спостерігаємо попиту на ключові слова, пов'язані з IT. У цей період також були популярні різні професії. Помітним стало слово «журналіст». Це може свідчити про попит на цю роботу через воєнні дії та необхідність освітлення подій. Запит на власне авто став більш помітним. Це може бути пов'язане зі змінами в менеджменті ризиків у період воєнних дій. Зменшилася популярність слова «Далекобійник» став менш помітним.

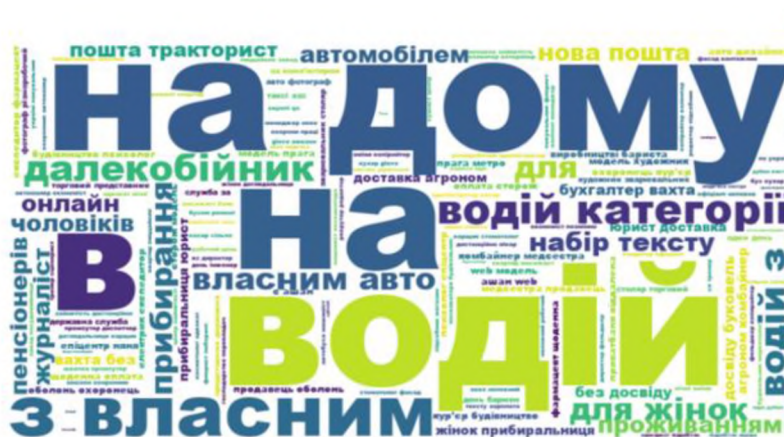


Рисунок 2.14 – Хмара ключових слів за 2023 рік

Наступний блок даних, який потребує додаткового опрацювання, є кількість IT-вакансій. Оскільки дані подаються щоденно, потрібно визначити одне значення для кожного місяця. Для цього найкраще підходить середньомісячна кількість вакансій. Періодами дослідження є масив дат:

['2022-05', '2022-06', '2022-07', '2022-08', '2022-09', '2022-10', '2022-11', '2022-12', '2023-01', '2023-02', '2023-03', '2023-04', '2023-05', '2023-06', '2023-07', '2023-08', '2023-09', '2023-10', '2023-11'].

Визначимо функцію `plot_monthly_median(url)` яка приймає посилання на файл json та створює графік середньомісячної кількості вакансій. Повний код функції наведено в додатку Б. Для передачі зображення в HTML використовується бібліотека `mpld3` і її метод `fig_to_html()`. Так ми конвертуємо графік в формат, який можна відобразити в HTML:

```
monthly_offers_html = mpld3.fig_to_html(fig)
```

Стильове оформлення інформаційної панелі можна виконати двома способами: готовими стилями та власною розміткою зі стилями. Перший спосіб вимагає підключення віддалених файлів `.css` або завантаження цих

файлів та локальне підключення. Комбінація двох методів надає переваги готових стилів та свободу в їх коректуванні. У роботі використано стилі Bootstrap. Bootstrap – це відкритий набір інструментів для веброзробників. До основних засобів та можливостей цього фреймворку відносяться:

- готова система адаптивної сітки. Bootstrap використовує адаптивну сіткову, побудовану на контейнерах із класами. Вона підлаштовується під мобільні пристрої, яка дозволяє створювати гнучкі та адаптивні макети. Поведінка для різних розмірів екрана вже визначена в файлах Bootstrap.
- попередньо стилізовані компоненти. Bootstrap надає набір попередньо розроблених компонентів, таких як панелі навігації, кнопки, форми тощо. Ці компоненти оформлені та готові до використання. Це значно заощаджує час і зусилля розробників на проєктуванні;
- CSS Flexbox і макет сітки. Bootstrap використовує CSS Flexbox і Grid, щоб забезпечити сучасні динамічні параметри макетів. Це полегшує створення складних і адаптивних дизайнів, а також вирівнювання елементів у різних площинах;
- вбудовані стилі для типографіки. Bootstrap надає набір стилів для типографіки, включаючи стилі заголовків, абзаців та списків. Це забезпечує чітку та послідовну типографічну ієрархію в усьому проєкті.
- сумісність з браузерами. Bootstrap сумісний із усіма сучасними браузерами, забезпечуючи стабільну продуктивність і єдиний вигляд на різних платформах.

У цій роботі стилі Bootstrap підключають за допомогою посилання на файл, який знаходиться на сервері розробника фреймворку:

```
<link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/css/bootstrap.min.css">
```

Для підключення стилів за допомогою посилання на Інтернет-ресурс необхідний доступ до мережі Інтернет. Перевагами такого підключення є відсутність необхідності завантажувати та зберігати файли стилів локально. Це зменшує вагу теки проєктом. Основою Bootstrap є адаптивні контейнери. Розмір загального ряду блоків сягає 12 одиниць.

Розмір внутрішніх контейнерів не може перевищувати 12 одиниць. Наприклад, для графіку перетину кордону маємо контейнер `<div class="col-lg-5">` та `<div class="col-lg-7">`. Ці контейнери вміщують опис графіку та зображення графіку. Разом їх розмір становить 12 одиниць. Адаптивність контейнерів забезпечується лише за умови розміщення `<div class="col-lg-5">` та `<div class="col-lg-7">` у батьківському контейнері з Bootstrap-класом `row`.

Для редагування стилів інформаційної панелі також використано Bootstrap.

Після додавання стилів маємо дашборд (рис. 2.15-2.19). Зображення наведені в послідовності візуалізацій. До зображень додано опис. Структура сторінки поділена на секції, які відокремлюються обведенням блоків. Кожен блок також поділений на дві частини. Перша містить опис до зображень, а в іншій знаходиться відповідне зображення.

Залежність ринку IT-вакансій від інших факторів

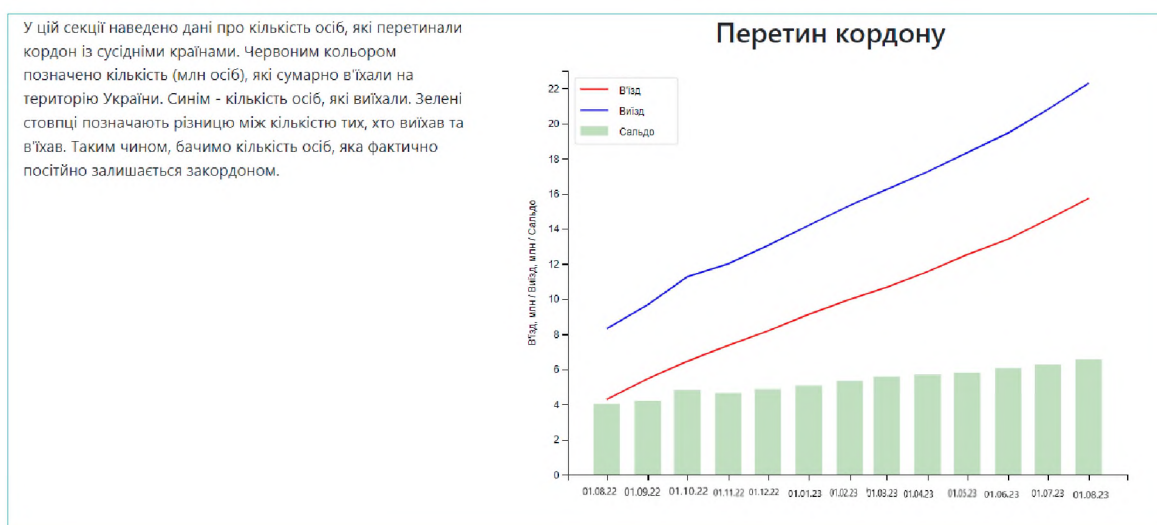


Рисунок 2.15 – Перша частина дашборду

Популярні пошукові запити

До 01.02.2022

Хмара відповідає періоду до 01.02.2022 - панування COVID. Особливо помітними є вирази «на дому», «водій», «онлайн», «бухгалтер», «для жінок», «психолог» «з проживанням», «пошта». Ми не спостерігаємо особливого попиту на ключові слова, пов'язані з IT. У цей період були популярні різні професії.

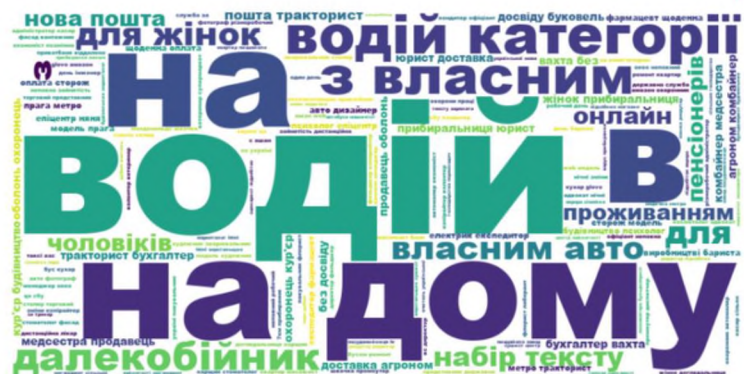


Рисунок 2.16 – Друга частина дашборду

2022 рік

Особливо помітними є: «на дому» та «водій», як і в попередньому періоді. Помітними є «далекобійник», «пошта», «для жінок і чоловіків», «з власним автомобілем» «з проживанням», «пошта», «набір тексту», хоча набір тексту став менш популярним. Ми не спостерігаємо попиту на ключові слова, пов'язані з IT. У цей період також були популярні різні професії. Помітним стало слово «журналіст». Це може свідчити про попит на цю роботу через воєнні дії та необхідність освітлення подій. Запит на власне авто став більш помітним. Це може бути пов'язане зі змінами в менеджменті ризиків у період воєнних дій. Зменшилася популярність слова «далекобійник» став менш помітним. Причиною можуть бути: порушення ланцюгів торгівлі з іншими державами, складність виїзду за кордон, зменшення кількості населення в Україні, а отже й кількості товарів. З'явилися запити «охоронець», що може свідчити про прагнення знайти роботу в цій сфері та збільшення попиту на захист майна. Зменшилася позиція слова «тракторист».

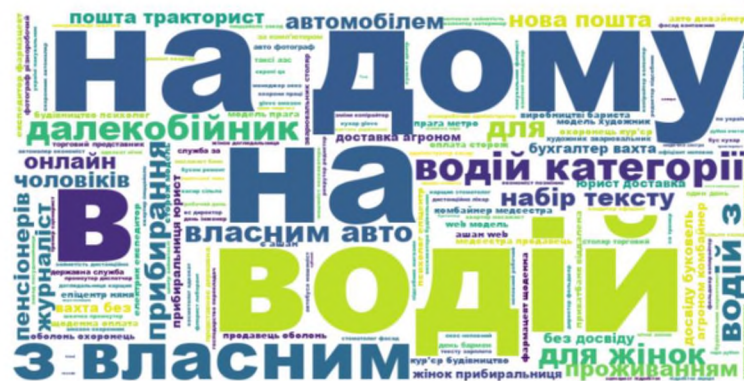


Рисунок 2.17 – Третя частина дашборду

візуалізації в готових рішеннях, наприклад на сторонніх ресурсах, то гнучкість представлення даних зменшується. Натомість додається фактор витрат на використання сторонніх сервісів. Якщо інформаційну панель повністю виконано на Python, ми маємо продукт, виготовлений лише за допомогою відкритих і безкоштовних інструментів. [24].

Висновки до розділу 2

Збір та обробка даних є відповідальним завданням, оскільки деякі дані можуть бути розміщені на ресурсах, доступ до яких вже відсутній. У такому випадку необхідно звертатися до Інтернет-архів вебсторінок. Іншою проблемою є відсутність швидкого доступу до певних даних. Оскільки джерела знаходяться на різних ресурсах, необхідно створювати окремі функції для парсингу даних. Щоб проаналізувати популярні запити в Україні через вебсайт Jooble, необхідно використовувати WayBackMachine. Функція `extract_links_text_and_write_to_file` виконує парсинг текстових даних та створення файлу із запитамі. Дані з порталу UNHCR можна зберегти у форматі xls. Збір даних про кількість IT-вакансій в Україні можна виконати через API від UAdata.

У дослідженні використовуються типові методи аналізу та прогнозування числових даних, зокрема регресійний аналіз та аналіз часових рядів. Зміна даних визначається часом та може бути пов'язана з впливом різних факторів. У схожих дослідженнях застосовуються три методи регресійного аналізу. Проста лінійна регресія використовується для моделювання зв'язку між двома змінними, що дозволяє прогнозувати значення залежної змінної на основі незалежної. Множинна лінійна регресія застосовується, коли зв'язок між залежною та кількома незалежними змінними складніший. Нелінійний регресійний аналіз використовується для складних зв'язків між змінними, де лінійна модель не ефективна. Оскільки в

дослідженні перевіряється залежність змінної від кількох змінних та однієї змінної від іншої, то доречним є використання простої та багатовимірної регресії.

Створення інформаційної панелі без застосування сторонніх ресурсів надає свободу в виборі стильового оформлення та зменшує витрати на проєктування, оскільки мова програмування Python і її бібліотеки відкриті та безкоштовні навіть для комерційних цілей. [25, с. 5]

РОЗДІЛ 3

ПРАКТИЧНІ АСПЕКТИ ПРОГНОЗУВАННЯ РИНКУ ОНЛАЙН- ВАКАНСІЙ В ІТ-ГАЛУЗІ УКРАЇНИ

3.1 Побудова прогнозу на основі історичних даних

Використовуючи історичні дані з 01.08.2022 по 01.08.2023, спочатку виконаємо прогноз для кількості людей, що виїхали. Початком прогнозу встановимо 08.2022. Для прогнозу на один крок вперед необхідно розширити датасет «border», показаний на рисунку 3.1. Цей метод підходить, оскільки маємо змінну часу спостереження та числове значення [26].

Використовуючи інструменти Excel виконаємо прогнозування (рис. 3.1).

	A	B	C
1	Дата	Виїзд, млн	Прогноз(Виїзд, млн)
2	01.08.2022	8.31	
3	01.09.2022	9.67	
4	01.10.2022	11.27	
5	01.11.2022	12.00	
6	01.12.2022	13.05	
7	01.01.2023	14.20	
8	01.02.2023	15.33	
9	01.03.2023	16.24	
10	01.04.2023	17.25	
11	01.05.2023	18.34	
12	01.06.2023	19.46	
13	01.07.2023	20.80	
14	01.08.2023	22.29	22.29
15	01.09.2023		23.29
16	01.10.2023		24.40
17	01.11.2023		25.51
18	01.12.2023		26.62

Рисунок 3.1 – Аркуш прогнозу для змінної «Виїзд»

Знайдемо рівняння регресії у вигляді

$$Y = -2.5716 + 0.8209X$$

де Y – в'їзд, млн. осіб; X – виїзд, млн. осіб.

Дані для визначення параметрів регресії представлені на рисунку 3.2.

The screenshot shows a web-based regression analysis tool. At the top, there are four tabs: "Simple Linear regression" (selected), "Multiple Linear regression", "Logistic regression", and "Multinomial regression". Below the tabs is a data input table with two columns: "X" and "Y". The table contains 14 rows of data. Below the table, there is a section for "X values for prediction: (You may leave empty)" with a text input field containing the values "23.29 24.40 25.51 26.62". At the bottom, there is a note: "You may change the X and Y labels. Separate data by **Enter** or comma, **,** or space after each value."

X	Y
8.31	4.28
9.67	5.46
11.27	6.45
12.00	7.36
13.05	8.18
14.20	9.13
15.33	9.98
16.24	10.67
17.25	11.56
18.34	12.53
19.46	13.41
20.80	14.53
22.29	15.73

X values for prediction: (You may leave empty)

23.29 24.40 25.51 26.62

You may change the X and Y labels. Separate data by **Enter** or comma, **,** or space after each value.

Рисунок 3.2 – Дані для визначення параметрів регресії

Тепер визначимо параметри рівняння регресії для кількості вакансій та перевіримо результати прогнозування. Отримане рівняння регресії має вид:

$$Y = 10729.72582 + 1555.753551 \cdot X_1 - 2391.160726 \cdot X_2.$$

На рисунку 3.3 бачимо, що отримане рівняння визначає кількість вакансій за 9-й місяць 2023 року на рівні 7398 проти реального 7118,5. Наступні значення менші за реальні (рис. 3.3).

	A	B	C	D	E	F	G	H
1	Дата	Виїзд, млн	В'їзд, млн	Сальдо	Кількість вакансій	Прогноз вакансій		
2	01.08.2022	8.31	4.28	4.03	13588	13423.84		
3	01.09.2022	9.67	5.46	4.21	13569	12718.09		
4	01.10.2022	11.27	6.45	4.82	13621	12840.05		
5	01.11.2022	12.00	7.36	4.64	12519.5	11799.79		
6	01.12.2022	13.05	8.18	4.87	10503	11472.57		
7	01.01.2023	14.20	9.13	5.07	9208	10990.08		
8	01.02.2023	15.33	9.98	5.35	9982	10715.59		
9	01.03.2023	16.24	10.67	5.57	9786	10481.43		
10	01.04.2023	17.25	11.56	5.69	9643	9924.6		
11	01.05.2023	18.34	12.53	5.81	9853	9300.941		
12	01.06.2023	19.46	13.41	6.05	9325.5	8939.16		
13	01.07.2023	20.80	14.53	6.27	9549	8345.765		
14	01.08.2023	22.29	15.73	6.56	7600	7794.439	Реальне значення	
15	01.09.2023	23.29	16.546	6.74	7398.383688	7398.305	7118.5	
16	01.10.2023	24.40	17.4571	6.94	6945.545889	6945.463	7218	
17	01.11.2023	25.51	18.3683	7.14	6492.468975	6492.383	7172	
18	01.12.2023	26.62	19.2794	7.34	6039.631176	6039.541		
19								
20								

Рисунок 3.3 – Лист Excel із прогнозованими значеннями

Можемо зробити висновок, що наявні дані не можуть пояснити комплексний склад явища зміни кількості вакансій. З великою ймовірністю такі дані, як захворюваність на Covid-19 також не додадуть точності. Обмеження дослідження розглянемо у наступному розділі.

3.2 Обмеження дослідження та можливості подальших досліджень

Як було визначено раніше, наявні дані не зможуть пояснити поведінку змінної «Кількість вакансій». Це явище явно залежить від великої кількості інших факторів.

Однією з помітних проблем є ненадійність і невизначеність, пов'язана із зібраними даними. Джерела даних можуть містити невідповідності, неточності або упередження у звітах. Крім того, економічна діяльність, на яку впливають геополітичні події та бойові дії, може спричинити непередбачувані коливання, що ускладнює отримання значущої інформації. У дослідженні не розглядається зміна на ринку вакансій сусідніх країн, хоча відомо, що кількість оголошень для українців збільшилася ще на початку відкритого конфлікту. Також не розглядається фактор виїзду компаній закордон. Для прикладу у статті на сайті спільноти програмістів ще за 28 липня 2022 року вказано, що згідно з даними досліджень та опитувань 20% працівників сфери виїхали закордон на той момент [27]. Тому можна припустити, що зменшення кількості працівників в Україні також вплинуло на кількість вакансій в Україні.

Відсутність даних за конкретні періоди також заважає проведенню комплексного аналізу. Наприклад, вебсайт DOU, який надає інформацію про зарплати, надає дані лише за обмежені періоди, переважно в грудні та червні. Це обмежує здатність фіксувати сезонні коливання або виявляти тенденції, які можуть мати місце в різні пори року. Інший фактор, який ми не в змозі точно оцінити, це зміна чисельності та доля чоловіків, які були змушені брати участь у воєнному протистоянні. Воєнні дії, що тривають у багатьох регіонах, суттєво впливають на точність і достовірність зібраних даних. На економічну діяльність, у тому числі на ринки праці, безсумнівно, впливає геополітична невизначеність.

Тіньова економіка, поширена в різних секторах, додає ще один рівень складності. IT-індустрія, зокрема, відома своєю значною неформальною чи незареєстрованою економічною діяльністю. Тіньова економіка додає обмеження на збір точних даних, оскільки значна частина пропозицій про роботу, відомостей про зарплати та іншої відповідної інформації може бути прихованою або неточною.

Незважаючи на ці проблеми, існують можливості для подальшого дослідження. У подальших дослідженнях можуть бути використані додаткові джерела даних, які надають більш повне уявлення про економічну діяльність сфери ІТ. Співпраця з інсайдерами галузі, опитування та цільові ініціативи зі збору даних можуть допомогти зібрати детальну інформацію про ринок праці, особливо в періоди невизначеності.

3.3 Визначення можливих стратегій для роботодавців і працівників

Динамічна природа сучасного ринку праці вимагає стратегічного підходу як для роботодавців, так і для працівників. Розглянувши проблеми, з якими стикнувся ринок праці, стає очевидним, що адаптивність, постійне навчання та спільні зусилля є важливими елементами успіху. Ми встановили, що популярність ІТ-професій серед шукачів роботи зростає, але роботодавці розміщують все менше й менше вакансій. Причинами цього є релокація бізнесу, воєнні дії, недовіра до спроможності робочого середовища зробити якісний продукт в умовах невизначеності. Вакансії для українців почали розміщуватися зокрема й на зарубіжних ресурсах, [28, 29, 30].

Тому очевидною проблемою є переманювання спеціалістів у сусідні країни, що ставить під загрозу сферу ІТ в Україні. Щоб зменшити негативний вплив ситуації уряду варто вести постійні спостереження стану цієї та інших галузей.

Працівники, з іншого боку, мають можливість знайти віддалену вакансію в іншій країні. З одного боку, це збереження спеціаліста в країні з перспективою отримання податків. З іншого боку, в уряді немає механізмів впливу на закордонні компанії. Для повного вирішення проблеми необхідна стабільна ситуація в країні, яка поверне бізнесу впевненість у своїй діяльності саме в Україні.

3.4 Економічна оцінка проєкту

При проведенні комплексного дослідження динаміки ринку праці необхідно забезпечити техніко-економічне обґрунтування необхідних ресурсів. Ефективність дослідницької діяльності залежить від різних факторів, включаючи обсяг дослідження, глибину аналізу та рівень співпраці з різними зацікавленими сторонами. Оцінка витрат, пов'язаних з дослідженням, передбачає розгляд кількох ключових компонентів.

Збір і обробка даних. Вартість отримання надійних і вичерпних наборів даних з різних джерел, включаючи вебсайти, галузеві звіти та економічні показники.

Персонал. Людські ресурси є критично важливим ресурсом. Наймання кваліфікованих професіоналів із досвідом аналізу даних, розумінням економіки та технологій має важливе значення для успіху дослідження.

Технологічна інфраструктура. Інвестиції в технологічну інфраструктуру, включаючи інструменти аналітики, обчислювальні ресурси та ліцензії на програмне забезпечення є важливими для обробки й аналізу великих наборів даних.

Більш комплексне дослідження, яке охоплює ширший спектр факторів, що впливають на ринок праці, потребувало б додаткових ресурсів.

Розширений набір даних дозволить провести більш широке дослідження, але потребує збору даних протягом більш тривалого періоду, що вимагає постійних інвестицій для їх збору і зберігання.

Консультації експертів галузі та взаємодія з експертами з різних галузей і секторів також потребує фінансових ресурсів і часу.

Співпраця з урядом та залучення державних дозволить отримати доступ до закритих даних, але для цього доведеться стикнутися з бюрократичними процесами, що може вимагати додаткових ресурсів для ефективною співпраці.

На рисунку 3.4 наведено розрахунок вартості інформаційної системи. Дані про оплату праці взяті з відкритих джерел та представляють реально можливу вартість роботи спеціалістів.

	A	B
1		Оплата за годину
2	Аналітик	200
3	Frontend-спеціаліст	180
4	Програміст	200
5	Експерт галузі/Написання статті	150
6		
7	Етапи робіт	
8	Оцінка проєкту	2250
9	Збір даних, обробка даних	9000
10	Побудова дашборду	540
11	Створення опису або статті	1200
12	Підсумок	12990
13		

Рисунок 3.4 – Перелік необхідних спеціалістів та етапів робіт

Наведена вартість на Рисунок 3.5 є орієнтовною і не включає додаткових витрат, які можливі при залученні додаткових експертів, платних сервісів, витрат на налагодження контакту з представниками влади, які мають доступ до закритих даних, та інших витрат у залежності від масштабів дослідження. Варто зазначити, що якщо планується створення окремого вебсайту, на якому буде розміщуватися інфографіка, то необхідно врахувати такі витрати:

- хостинг – 300 грн;
- домен com.ua – 480 грн на рік.

Мінімальна вартість підтримки сайту на місяць становить $300 + 40 = 340$ грн. Сюди входить оплата за хостинг і домен на місяць. Вартість роботи працівників розрахована за формулою

Час виконання задачі * Середня заробітна плата за годину,
де усі значення базуються на відкритих даних про середню оплату за годину для стороннього спеціаліста [31].

	A	B
1		Оплата за годину
2	Аналітик	200
3	Frontend-спеціаліст	180
4	Програміст	200
5	Експерт галузі/Написання статті	150
6		
7	Етапи робіт	
8	Оцінка проєкту	=15*150
9	Збір даних, обробка даних	=30*B2+ 15*B4
10	Побудова дашборду	=3*B3
11	Створення опису або статті	=8*B5
12	Підсумок	=СУММ(B8:B11)
13		
14		

Рисунок 3.5 Розрахунок вартості робіт

Висновки до розділу 3

Обмеження дослідження суттєві. На даний момент можливості збору актуальних і точних даних обмежені, що робить результати прогнозування неточними. Відсутність доступу до даних про фактори, які можуть краще пояснити стрімке зменшення кількості вакансій, але збільшення зацікавленості шукачів роботи в сфері ІТ, може бути вирішена шляхом співпраці з урядом і платформами, які мають необхідні дані. Для покращення точності прогнозування необхідно розглянути додаткові змінні. Можливості для майбутніх досліджень включають розширення датасетів, співпрацю з фахівцями в різних сферах економіки для інтерпретації змін і їх причин. Визначено, що окрім дослідження українських джерел необхідно розширити географію збору даних на дошки оголошень і пошуку роботи в сусідніх країнах. Для цього необхідно співпрацювати зі спеціалістами, які знають мови сусідніх країн. Оскільки масштаби дослідження значно впливають на вартість робіт, розрахункова вартість проєкту може змінюватися. Якщо проводити більш комплексне дослідження, то сума може значно збільшитися.

ВИСНОВКИ

Ринок праці є соціальною структурою, на яку впливають внутрішні та зовнішні фактори. У роботі особливу увагу приділено аналізу поточних змін на ринку ІТ-вакансій України, можливостям та обмеженням прогнозного моделювання в контексті сфери ІТ. Використовуючи історичні дані для прогнозування кількості ІТ-вакансій, створено прогноз можливих змін, які піддаються часовим і зовнішнім впливам.

Основною тематикою дослідження є використання відкритих засобів програмування для аналізу та візуалізації даних. Також розглянуто застосування методів регресійного аналізу та методології прогнозування часових рядів. Запропоновано, що комбінування різних методів прогнозування дозволяє ефективно визначити можливі значення змінних. У контексті змін кількості вакансій незалежна змінна, яка відповідає за кількість осіб, що перетнули кордон в обидва боки, може бути прогнозована за допомогою метода часового ряду. Таким чином отримано вхідні дані для прогнозування іншої змінної – кількості вакансій. Цього можна досягти шляхом побудови рівняння регресії між залежною змінною та незалежними факторами, які впливають на значення залежної змінної. У роботі описано практичний приклад застосування простої лінійної регресії. Цей метод відображає фундаментальний зв'язок між двома змінними. Множинна лінійна регресія враховує багатofакторний зв'язок між залежною змінною та факторами.

Існує перспектива проведення подальших досліджень із розгляданням додаткових змінних. Наразі відомо, що популярність професій у сфері ІТ зростає. Збільшення популярності пошукових запитів почалося у 2023 році. При цьому кількість пропозицій від роботодавців зменшилася й становить лише 7172 вакансії на момент збору даних за листопад 2023 року. Для порівняння, у травні 2022 року медіанна кількість вакансій за місяць становила 15867 вакансії. Очевидно, що з часом вплив таких факторів, як

погіршення економічної ситуації, розгортання бойових дій, знищення інфраструктури, евакуація підприємств в інші країни, виїзд жителів в сусідні регіони та країни лише погіршили ситуацію для бізнесу в Україні. Стан невизначеності щодо безпеки сприяє відмовам від інвестицій у проекти на території України.

У подальших дослідженнях необхідно розширити географію для збору даних та приділити ще більше уваги супутнім показникам і обмеженням загалом. Для того, щоб комплексно дослідити природу змін на ринку ІТ потрібно розглядати економіку загалом.