МЕНЕДЖМЕНТ

УДК 005.95.96

Б. М. Бойко[1], І. С. Процик[2]
ORCID ID: [1] 0009-0002-9850-2764, [2]0000-0002-6370-1344
Національний університет "Львівська політехніка"
Кафедра менеджменту та міжнародного підприємництва

АВТОМАТИЗАЦІЯ ФОРМУВАННЯ ПРОФЕСІОГРАМИ СПЕЦІАЛІСТА НА ОСНОВІ АНАЛІЗУ ВАКАНСІЙ ТА ТЕКСТОВИХ ДАНИХ

https://doi.org/10.23939/smeu2025.02.001

Представлено комплексне дослідження проблеми автоматизації процесу професійного відбору фахівців у сфері штучного інтелекту шляхом побудови професіограм із використанням інформаційних технологій. Основну увагу зосереджено на розробці інноваційної моделі, яка поєднує підходи HR-аналітики та методи обробки неструктурованих текстових даних. Актуальність теми зумовлена стрімким розвитком цифрових технологій, зростанням попиту на спеціалістів з IT та потребою у стандартизованому описі компетенцій, необхідних для успішної професійної діяльності. Автори аналізують сучасний ринок праці, описують ключові вимоги роботодавців, здійснюють глибоке дослідження наукових джерел і застосовують такі алгоритми машинного навчання, як TF-IDF (Term Frequency-Inverse Document Frequency) і LDA (Latent Dirichlet Allocation) для тематичного моделювання даних. На основі зібраної інформації створюється структурована професіограма, яка включає класифікацію знань, навичок і компетенцій відповідно до їх значущості. Запропонована система має практичне значення для кадрового менеджменту, освітніх програм та процесів профорієнтації. Вона дозволяє скоротити час на формування професійних профілів, мінімізувати суб'єктивність у прийнятті кадрових рішень, покращити відповідність кандидатів посадовим вимогам, а також адаптуватися до швидких змін ринку праці. Стаття також містить UML-діаграми та приклади реалізації веб-застосунку, що підкреслює практичну реалізацію запропонованого підходу. Результати дослідження можуть бути корисними для HR-фахівців, керівників організацій, аналітиків, науковців і розробників інформаційних систем. Водночас окреслюються напрями подальшого розвитку, зокрема масштабування системи на інші професійні напрями, впровадження психологічних параметрів у моделі оцінювання та створення мобільних платформ для самодіагностики..

Ключові слова: професіограма, інформаційна система, штучний інтелект, автоматизація, HR-аналітика, компетенції, TF-IDF, LDA, управління персоналом, цифровізація.

Формулювання проблеми

Завданням даного дослідження є розробка ефективного, науково обґрунтованого підходу до побудови професіограми фахівця у сфері штучного інтелекту шляхом використання методів автоматизованого аналізу текстових даних. Зважаючи на стрімке зростання попиту на спеціалістів зі штучного інтелекту та одночасну нестачу стандартизованих описів професійних компетенцій, дана робота набуває особливої актуальності як у контексті кадрового менеджменту, так і для освітніх установ, що займаються підготовкою таких фахівців [1, 2].

Дослідження спрямоване на створення цифрового інструменту, що дозволяє ідентифікувати, систематизувати та структурувати ключові знання, навички й компетенції, необхідні для успішної професійної діяльності в галузі штучного інтелекту. Застосування такого інструменту забезпечує об'єктивне оцінювання кваліфікаційних характеристик, сприяє підвищенню ефективності процесів підбору персоналу, оптимізації освітніх програм та розробці індивідуальних траєкторій професійного розвитку [1, 3].

Особливої наукової й практичної цінності набуває впровадження у дослідженні алгоритмів машинного навчання та сучасних методів обробки неструктурованого тексту, зокрема TF-IDF (Term Frequency-Inverse Document Frequency) і LDA (Latent Dirichlet Allocation). Ці алгоритми дають змогу здійснити глибинний аналіз великого масиву текстової інформації, зокрема описів вакансій, статей з професійних ресурсів та експертних думок, для виявлення найбільш релевантних професійних термінів. Отримані результати формують інформаційне ядро професіограми, що відображає реальні потреби роботодавців та актуальні тенденції на ринку праці [4].

Суттєва особливість запропонованого підходу полягає в його комплексній автоматизації, що дозволяє не лише значно прискорити процес збору та обробки інформації, але й зменшити вплив людського чинника на формування опису професійних компетенцій. Такий підхід підвищує рівень достовірності та універсальності отриманих результатів, забезпечує масштабованість рішення та його адаптивність до різних спеціалізацій у межах сфери штучного інтелекту.

Таким чином, реалізація зазначеного підходу відкриває нові можливості для стратегічного управління людськими ресурсами, формування освітніх стандартів нового покоління та розробки інструментів персоналізованої оцінки й розвитку професійних компетенцій.

Актуальність дослідження

У сучасному бізнес-середовищі, що характеризується високим рівнем конкуренції, динамічними змінами технологій та глобалізацією ринку, ефективне управління людськими ресурсами набуває стратегічного значення для досягнення цілей організації. Однією з найперспективніших галузей економіки є сфера штучного інтелекту, що стрімко розвивається та активно інтегрується у корпоративні процеси – від автоматизації операцій до прийняття управлінських рішень на основі аналітики даних [1].

Водночас компанії стикаються з проблемами при наборі фахівців з цієї галузі. Відсутність чітких стандартів кваліфікації, розпорошеність уявлень про необхідні компетенції та суб'єктивність у процесі відбору кадрів призводять до неефективного використання людського капіталу, високого рівня плинності персоналу та додаткових витрат на адаптацію та навчання.

У цьому контексті розробка професіограми фахівця зі штучного інтелекту є вкрай актуальною з позиції бізнес-адміністрування та менеджменту. Вона дозволяє:

- сформувати об'єктивну основу для рекрутингових рішень;
- стандартизувати вимоги до посад у межах корпоративної структури;
- створювати індивідуальні траєкторії розвитку персоналу;
- знижувати витрати на підбір та адаптацію працівників;
- підвищувати ефективність освітніх програм у корпоративному навчанні.

Особливої цінності набуває використання методів машинного навчання та аналізу неструктурованих даних (зокрема TF-IDF та LDA), які дозволяють здійснити масштабований, автоматизований і точний аналіз великого масиву інформації з відкритих джерел: вакансій, статей, відгуків, експертних досліджень.

Таким чином, з точки зору сучасного менеджменту, актуальність дослідження полягає в тому, що воно забезпечує підприємствам можливість приймати більш зважені кадрові рішення, покращувати HR-аналітику, підвищувати адаптивність до змін ринку праці та розвивати стратегії управління персоналом на основі даних (data-driven management).

Формулювання мети та завдань статті

Метою дослідження є розробка інноваційного підходу до формування професіограми фахівця зі штучного інтелекту шляхом автоматизованого аналізу текстових даних з використанням алгоритмів машинного навчання. Це дозволить створити ефективний інструмент для управління компетенціями, оптимізації процесу підбору персоналу та вдосконалення кадрових стратегій у межах сучасної бізнес-моделі.

Для досягнення поставленої мети необхідно вирішити такі завдання:

1. Проаналізувати сучасні тенденції розвитку ринку праці у сфері штучного інтелекту та визначити основні компетенції, які висуваються до фахівців.
2. Вивчити наукові підходи до побудови професіограм та адаптувати їх до умов цифрового бізнес-середовища.
3. Розробити алгоритм автоматизованого аналізу текстових джерел (вакансій, експертних статей, публікацій) для виявлення релевантних знань і навичок.
4. Застосувати методи TF-IDF та LDA для класифікації та структуризації ключових компетенцій фахівців у галузі штучного інтелекту.
5. Сформувати професіограму у вигляді структурованої моделі, що включає ранжування компетенцій за вагою та їх тематичну класифікацію.
6. Оцінити можливості практичного застосування професіограми у кадровому менеджменті, HR-аналітиці, системах навчання персоналу та рекрутингових платформах.

Реалізація цих завдань дозволить створити ефективний інструмент для аналізу та оцінки професії фахівця зі штучного інтелекту, що сприятиме розвитку галузі та підвищенню якості підготовки кадрів.

Аналіз останніх досліджень і публікацій

У сфері бізнес-адміністрування розвиток технологій штучного інтелекту (ШІ) розглядається як один із ключових факторів трансформації бізнес-моделей, оптимізації операційної діяльності та забезпечення конкурентоспроможності компаній. Зростання обсягів даних, розвиток машинного навчання і цифровізація процесів ведуть до появи нових вимог до людського капіталу, особливо у сегменті висококваліфікованих фахівців зі ШІ.

Сучасний ринок праці демонструє зростаючий попит на таких спеціалістів, однак відсутність уніфікованих професійних стандартів та систематизованих підходів до опису їх компетенцій ускладнює процес кадрового планування, підбору персоналу та адаптації працівників до корпоративних цілей. Для вирішення цих проблем у практиці управління персоналом все більшого значення набуває розробка професіограм – інструментів, які дозволяють стандартизувати і формалізувати ключові компетенції відповідно до стратегічних завдань бізнесу.

Класичні теорії управління трудовими ресурсами, зокрема праці Г.М. Климова [1] та Є.А. Клімової [2], акцентують увагу на необхідності глибокого аналізу трудових функцій, особистісних якостей та кваліфікаційних вимог до професій. У контексті сучасного менеджменту ці підходи трансформуються у цифрову площину. Роботи таких авторів, як І. В. Захарова [3], демонструють важливість адаптації професіограм до динамічного ринку праці та потреб бізнесу в умовах цифровізації. Застосування великих даних і аналітичних алгоритмів дозволяє створювати адаптивні, актуальні моделі професійних профілів, орієнтовані на реальні потреби організацій.

За аналітичними звітами компаній Gartner та LinkedIn, посади, пов'язані зі штучним інтелектом, входять до числа найбільш затребуваних і високооплачуваних у світі. Їх основними споживачами є великі технологічні корпорації, фінансові установи, медичні компанії та промислові підприємства, які активно впроваджують рішення на основі ШІ у власні бізнес-процеси [8].

З точки зору HR-аналітики, визначення ключових компетенцій таких фахівців є критично важливим. Наприклад, у дослідженнях Девенпорта та Патіла [7] особливу увагу приділено міждисциплінарним навичкам, які поєднують технічну експертизу зі стратегічним мисленням та аналітичними здібностями.

Таким чином, сучасна література й аналітичні джерела однозначно підтверджують міждисциплінарний характер професії фахівця зі штучного інтелекту та важливість постійного оновлення знань. Формування професіограм на основі актуальних даних з відкритих джерел та застосування технологій аналітики тексту стає стратегічним інструментом для HR-менеджменту. Це дозволяє не лише оптимізувати процес найму і навчання, але й підвищити узгодженість між очікуваннями бізнесу та можливостями людського капіталу.

Виклад основного матеріалу

Організації постійно зазнають змін у структурі: це може бути реструктуризація, масштабування, скорочення або релокація. У великих компаніях це супроводжується трансформацією відділів і команд, впровадженням нових технологій та зміною стратегічних пріоритетів. Такі зміни часто впливають на функціональні обов'язки працівників, їхній робочий ритм та перелік завдань.

З точки зору бізнес-аналізу, професіограма є інструментом, що дозволяє систематизувати ключові характеристики посади або ролі, використовуючи консолідовані джерела інформації. Її формування – ресурсомісткий процес, що вимагає участі експертів і збору якісних даних.

Першим кроком у цьому процесі є організація збору релевантної інформації. Бізнес-аналітик або HR-фахівець (переважно з психологічною підготовкою) здійснює глибинний аналіз даних, отриманих від працівників, а також з відкритих джерел. Це створює базу для подальшої побудови детального профілю професії чи посади.

На рис. 1 подано найпоширеніші методи збору даних про професійні ролі – вони служать відправною точкою для бізнес-аналізу функціональних вимог.

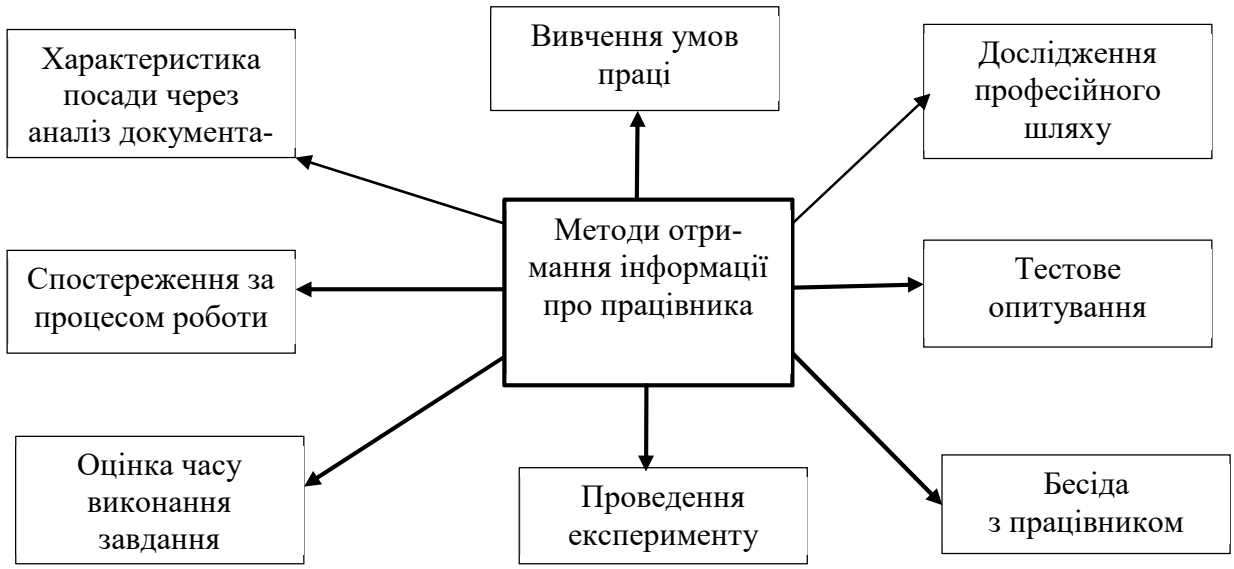


Рис. 1. Способи збору інформації

До основних методів збору даних, які використовуються у процесі побудови професіограми, належать наступні:

- Аналіз документації, що описує функціональні обов'язки посади (посадові інструкції, оголошення про вакансії, матеріали спеціалізованої літератури).

- Пряме спостереження за виконанням працівником своїх обов'язків.
- Визначення тривалості виконання окремих завдань.
- Опитування (анкетування) представників відповідної професійної групи.
- Проведення інтерв'ю з працівником з метою збирання особистих вражень та побажань.
- Аналіз професійного розвитку та кар'єрного шляху.
- Вивчення умов праці.
- Організація контрольованого експерименту для моделювання робочих ситуацій.

На другому етапі здійснюється централізоване збереження зібраних даних у відповідному сховищі. Зокрема, після проведення інтерв'ю або анкетування фахівця накопичується інформація про витрати часу, професійний досвід, мотиваційні аспекти тощо.

Третій етап передбачає дослідження форматів представлення професіограми. Завдання аналітика – вибрати найбільш ефективний спосіб візуалізації даних, зручний для кінцевих користувачів.

На четвертому етапі проводиться глибокий аналіз усієї зібраної інформації. Його мета – виокремити ключові характеристики посади та сформулювати їх у доступній і структурованій формі, орієнтованій на просте сприйняття.

Узагальнену модель побудови професіограми представлено у вигляді діаграми діяльності UML (рис. 2).

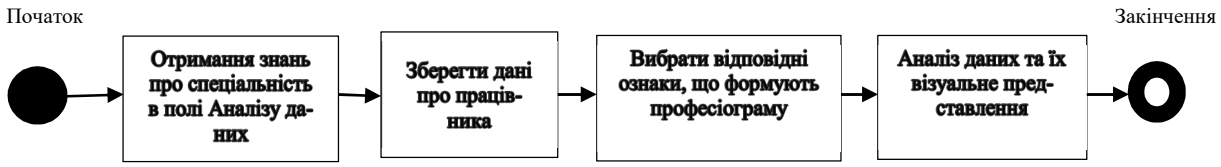


Рис. 2. UML-діаграма побудови професіограми

Кожен із перелічених етапів можна автоматизувати, що дозволяє значно підвищити ефективність процесу та мінімізувати вплив людського фактора. Автоматизація забезпечує швидке опрацювання великих обсягів даних і, що особливо важливо, виключає суб'єктивність дослідника у процесі формування професіограми.

Основна ідея автоматизації полягає у застосуванні алгоритмів для збору та аналізу релевантної інформації з відкритих джерел. Зокрема, для дослідження навичок спеціалістів у сфері штучного інтелекту можливо використовувати пошукові запити на кшталт "data scientist skills" з подальшим аналізом контенту веб-сторінок, які пропонує пошукова система.

На рис. 3 представлено приклад реалізації автоматизованого процесу збору інформації, що ілюструє, як систематизований підхід дозволяє ефективно агрегувати професійні характеристики з численних джерел.

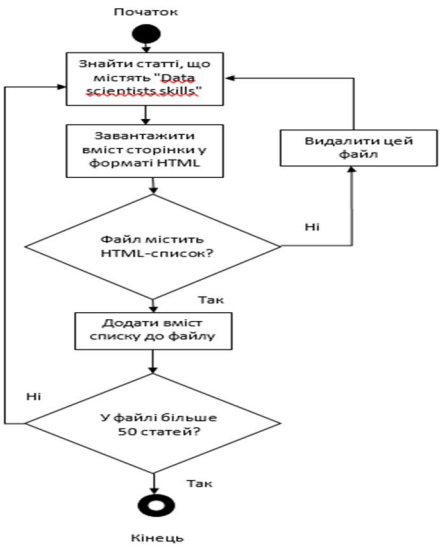


Рис. 3. UML-діаграма першого етапу

Дана схема представляє собою процес автоматизованого збору галузевої інформації з метою формування бази знань про професійні компетенції спеціалістів у сфері аналізу даних. Такий підхід дозволяє менеджменту швидко отримувати актуальну та структуровану інформацію для прийняття стратегічних HR-рішень.

1. Пошук релевантного контенту – система ініціює пошук веб-статей за ключовими словами “Data scientists skills”.
2. Завантаження веб-сторінок – сторінки з результатів пошуку завантажуються у форматі HTML для подальшого опрацювання.
3. Аналіз структури вмісту – перевіряється, чи містить сторінка структуровану інформацію (наприклад, списки навичок).
4. Відбір і очищення даних – якщо сторінка не відповідає вимогам, вона відсіюється. Інакше – вміст списку додається до робочого файлу.
5. Контроль обсягу даних – як тільки накопичується понад 50 релевантних записів, процес завершується.

Такий процес спрямований на підвищення ефективності інформаційного забезпечення управлінських рішень, особливо у сфері формування кадрової стратегії, оцінки компетенцій та розробки навчальних програм.

Основними даними для побудови професіограми є текстові дані – опис необхідних навичок для даної посади. Приклад професіограми зображений на рис. 4.

Зміст праці: Розбирає, ремонтує, складає та випробовує прості вузли і механізми устаткування, агрегатів та машин. Виконує слюсарне оброблення деталей, роботи із застосуванням пневматичних, електричних інструментів свердлильних верстатів тощо.

Слюсар-ремонтник повинен знати: основні заходи виконання робіт з розбирання, ремонту та складання простих вузлів і механізмів, устаткування, агрегатів та машин; призначення та правила застосування слюсарного та контрольно-вимірювального інструменту; основні механічні властивості оброблюваних матеріалів; основні поняття про допуски і посадки, квалітети і параметри шорсткості; найменування, маркування і правила застосування мастил, мийних речовин, металів.

Умови праці: працюють слюсарі стоячи, положення тіла нахилене. В роботі приймають участь м'язи верхніх кінцівок, плечового поясу, спини, нижніх кінцівок.

Області застосування: Переробна промисловість. Ремонт і монтаж машин і устаткування.

Професійна спрямованість: по типу «людина – техніка».

Домінуючі інтереси: техніка; *супутні* - математика, фізика, хімія, металообробка, транспорт.

Необхідні якості: фізична витривалість і сила, хороший зір і окомір, рухливість, координація і точність рухів кистей і пальців рук, спостережливість, розвинута просторова уява, гарна образна і рухова пам'ять, технічна кмітливість.

Рис. 4. Професіограма слюсаря

Текстові дані – це послідовності з підмножин знаків, які містять в собі тільки друковані символи та керуючі знаки (наприклад табуляція чи пробіл).Отже множина всіх символів (P) – це об’єднання:

$$P = A \cup B \cup C \cup D, \quad (1)$$

де A – підмножина літер, B – підмножина розділових знаків; C – підмножина символів, D – підмножина цифр.

Вони можуть бути представлені різними мовами (формальними чи природніми) та форматами. Деякі з них можуть бути зрозумілими людині, а інші – лише комп’ютеру. Формальні мови використовуються в певних галузях науки, наприклад, алгоритмічна (блок-схеми), логічна (дискретна математика), математична (формули, рівняння) і т.д. Природні мови використовуються у спілкуванні людей, прикладами є англійська, українська, іспанська мови тощо.

Враховуючи те, що найбільша кількість даних з обраної теми є англійською мовою, тому було прийнято рішення використовувати саме англійські тексти для подальшої обробки. Для збереження інформації, що підлягатиме обробці використовуватимуться статті, які видає пошукова система по визначеному запиту. Схема структури даних зображена на рис. 5.

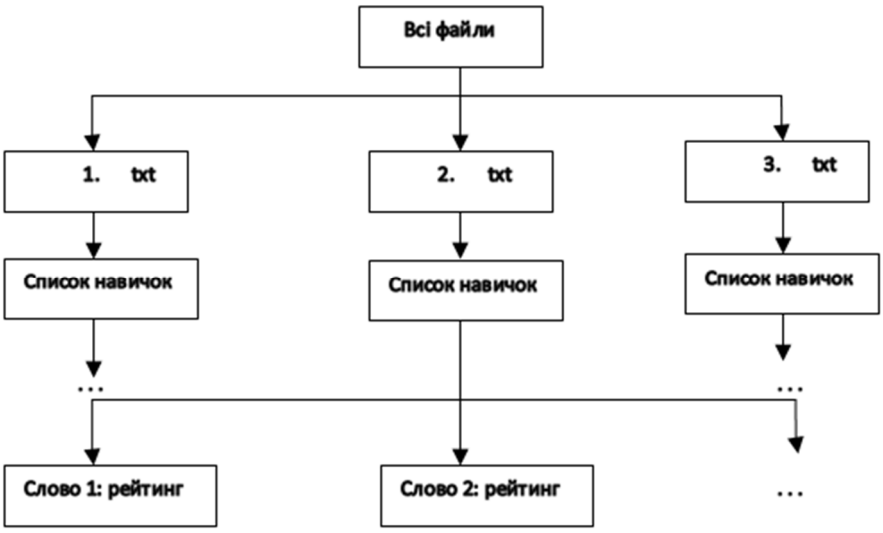


Рис. 5. Структура даних

Як видно з Рис. 5, кожен зібраний текст зберігається як окремий документ, що забезпечує структуроване й гнучке управління інформаційними потоками. У кожному документі міститься набір списків, які формуються у вигляді асоціативних масивів. Такі масиви виконують роль ключових інформаційних одиниць, де кожне слово поєднується з відповідним числовим значенням – показником його значущості або ваги у контексті тематики.

З управлінської точки зору, асоціативний масив – це ефективний інструмент представлення даних, який дозволяє зручно організовувати інформацію за ключовими категоріями (наприклад, ключові компетенції, терміни, пріоритети), без прив'язки до послідовної нумерації. Це дає змогу швидше аналізувати зміст документів, виявляти тренди, та підтримувати прийняття стратегічних рішень.

На відміну від класичних масивів, де елементи зберігаються у фіксованому порядку та доступні через числові індекси, асоціативні масиви використовують гнучку структуру – наприклад, хеш-таблиці. Такий підхід є особливо корисним у побудові систем знань, баз даних компетенцій, або контент-аналітики в HR та маркетингових відділах.

На рис. 6 наведено приклад такого масиву для одного з тематичних списків – він ілюструє, як можна швидко зіставити ключові поняття з їхньою вагою в аналізованому матеріалі.

```
Counter({'analysis': 1, 'more': 1, 'degree': 0.0368, 'master': 0.01577, 'sciences': 0.01051, 'statistics': 0.01051, 'mathematics': 0.01051, 'programme': 0.01051, 'learned': 0.01051, 'classroom': 0.01051, 's': 0.00998, 'big': 0.00775, 'field': 0.00775, 'or': 0.00675, 'computer': 0.00614, 'any': 0.00614, 'enable': 0.00614, 'are': 0.00552, 'have': 0.00552, 'education': 0.00526, 'highly': 0.00526, 'educated': 0.00526, '88': 0.00526, '46': 0.00526, 'phds': 0.00526, 'while': 0.00526, 'notable': 0.00526, 'exceptions': 0.00526, 'educational': 0.00526, 'background': 0.00526, 'usually': 0.00526, 'could': 0.00526, 'earn': 0.00526, 'bachelor': 0.00526, 'physical': 0.00526, 'fields': 0.00526, '32': 0.00526, 'followed': 0.00526, '19': 0.00526, 'engineering': 0.00526, '16': 0.00526, 'courses': 0.00526, 'give': 0.00526, 'process': 0.00526, 'analyze': 0.00526, 'after': 0.00526, 'done': 0.00526, 'yet': 0.00526, 'truth': 0.00526, 'ph': 0.00526, 'undertake': 0.00526, 'querying': 0.00526, 'therefore': 0.00526, 'enroll': 0.00526, 'program': 0.00526, 'astrophysics': 0.00526, 'related': 0.00526, 'during': 0.00526, 'transition': 0.00526, 'apart': 0.00526, 'practice': 0.00526, 'app': 0.00526, 'starting': 0.00526, 'exploring': 0.00526, 'learn': 0.00499, 'at': 0.00388, 'least': 0.00388, 'there': 0.00388, 'very': 0.00388, 'develop': 0.00388, 'depth': 0.00388, 'necessary': 0.00388, 'become': 0.00388, 'social': 0.00388, 'common': 0.00388, 'study': 0.00388, 'd': 0.00388, 'online': 0.00388, 'training': 0.00388, 'special': 0.00388, 'skill': 0.00388, 'building': 0.00388, 'an': 0.00388, 'science': 0.00352, 'scientists': 0.00338, 'most': 0.00338, 'by': 0.00338, 'strong': 0.00307, 'required': 0.00307, 'knowledge': 0.00307, 'like': 0.00307, 'other': 0.00307, 'easily': 0.00307, 'blog': 0.00307, 'skills': 0.00276, 'they': 0.0025, 'hadoop': 0.0025, 'learning': 0.0025, 'will': 0.00223, 'these': 0.00205, 'also': 0.00205, 'how': 0.00205, 'your': 0.00176, 'what': 0.00169, 'not': 0.00138, 'use': 0.00138, 'from': 0.00138, 'scientist': 0.00134, 'and': 0.00103, 'be': 0.00067, 'need': 0.00067, 'for': 0.00067, 'can': 0.00061, 'of': 0.00059, 'in': 0.00059, 'is': 0.0003, 'data': 0.0, 'a': 0.0, 'to': 0.0, 'the': 0.0, 'you': 0.0})
```

Рис. 6. Приклад асоціативного масиву

Хеш-таблиця – це ефективна структура даних, у якій доступ до елементів організовано за допомогою спеціальної хеш-функції, що генерує унікальний індекс (ключ) для кожного значення. Це дозволяє значно прискорити процеси пошуку та вставки інформації, оскільки сам ключ використовується як індекс для звернення до значення в масиві. У практиці управління даними та інформаційних систем хеш-таблиці широко застосовуються для швидкого доступу до структурованої інформації, зокрема в HR-системах, CRM, ERP та ін.

Вивчення професій – це стратегічно важлива діяльність для формування ефективної кадрової політики підприємства. Основою для такого аналізу слугує професіограма – структурований опис ключових характеристик конкретної посади. Вона використовується в менеджменті персоналу, рекрутингу, профорієнтації та освітньому плануванні.

Слід зазначити, що значна частина підприємств проводить внутрішній аналіз професій та формує унікальні професіограми, які, як правило, не публікуються у відкритому доступі через комерційну цінність таких даних. В ході дослідження були опрацьовані відкриті джерела, які можна умовно розділити на кілька категорій.

Перша категорія – це текстові професіограми, розроблені фахівцями в галузі психології на основі психограм. Як приклад – професіограма програміста, знайдена у джерелі "Психологія праці". Вона охоплює такі управлінсько-важливі аспекти:

- тип і клас професії;
- профіль робочої діяльності та освітні вимоги;
- умови праці;
- ключові завдання і сфери компетенцій;
- перелік особистісних якостей, що сприяють або перешкоджають професійній ефективності;
- варіанти кар'єрного зростання та професійного розвитку;
- рекомендації щодо закладів освіти, які забезпечують підготовку за даним напрямом.

Цей формат професіограми можна використовувати для стратегічного планування кадрів, формування навчальних програм та оцінки відповідності персоналу до очікувань бізнесу. На рис. 7 представлено приклад такої професіограми.

Професіограма "програміст"

Історія професії. Першим програмістом була жінка - Ада Лавлейс.

У 1940-х рр. з'явилися цифрові ЕОМ. Ідея їх створення належить американському математику фон Нейману. Для машин першого покоління складалися гранично докладні програми, що передбачають кожен крок, кожну операцію обчислень. Причому ніякої мови, крім свого, машина ще не розуміла.

Пізніше створюються спеціальні мови програмування, що дозволяє звести процес складання програми до запису алгоритму у спеціальній символічній формі згідно з правилами цієї мови.

В даний час ведуться численні розробки в області обчислювальної техніки і програмування, і вже досягнуті неймовірні успіхи.

Про майбутнє комп'ютеризації йдуть жваві суперечки серед вчених, але, безсумнівно, результати прогресу в цій галузі перевершили всі наші очікування.

Тип професії людина - знак.

Клас професії - евристичний.

Професійна область - інформатика.

Базова освіта - вища професійна освіта.

Умови праці. Робота відбувається в офісі, переважно сидячи, з використанням комп'ютера.

Домінуючі види діяльності:

1) розробка на основі аналізу математичних алгоритмів програм, що реалізують рішення різних завдань;

Рис. 7. Текстова професіограма

Професіограми, як аналітичний інструмент у сфері управління персоналом, мають низку ключових переваг. Передусім вони забезпечують багатовимірну оцінку професії, охоплюючи такі аспекти, як історичний розвиток, тип і клас професій, функціональну сферу, освітні вимоги, робочі умови, а також переваги та потенційні обмеження тієї чи іншої спеціалізації. Це дозволяє менеджерам отримати повну картину про посадові ролі і сформувати ефективні кадрові стратегії.

Особливу цінність становить детальний опис психологічних характеристик, що дозволяє точно узгодити тип особистості з конкретними професійними вимогами. Такий підхід ґрунтується на багатофакторній чотириярусній класифікації професій. Також у професіограмі міститься перелік навчальних закладів, які готують фахівців відповідного профілю, що є корисним для планування підвищення кваліфікації або підбору молодих спеціалістів.

Ще однією перевагою є наявність вагових характеристик у параметрах оцінки – це допомагає стандартизувати підхід до оцінювання кандидатів, зменшуючи вплив суб'єктивних чинників.

Разом із тим, використання традиційних професіограм має і певні недоліки. Зокрема, вони часто формуються окремими фахівцями, тому в їхній структурі присутній елемент суб'єктивної інтерпретації. Деякі з описаних у професіограмі функцій можуть бути застарілими через швидкі зміни в технологіях та автоматизацію бізнес-процесів. Крім того, професіограма часто створюється під конкретну організацію, що ускладнює її універсальне використання.

Ще одним сучасним інструментом у сфері управління персоналом є онлайн-платформи самодіагностики навичок. Яскравим прикладом є система Youth4work, яка створена для підтримки молоді у розвитку професійних компетенцій. Користувач проходить низку тестів на знання з різних галузей, і за результатами формується профіль його навичок. Ця система генерує персоналізоване резюме, яке можна надати потенційним роботодавцям, сприяючи прозорому та ефективному підбору персоналу. Це дозволяє не лише самостійно оцінити власні професійні можливості, але й автоматично інтегруватись у ринок праці, орієнтований на об'єктивні показники знань та вмінь (рис 8).

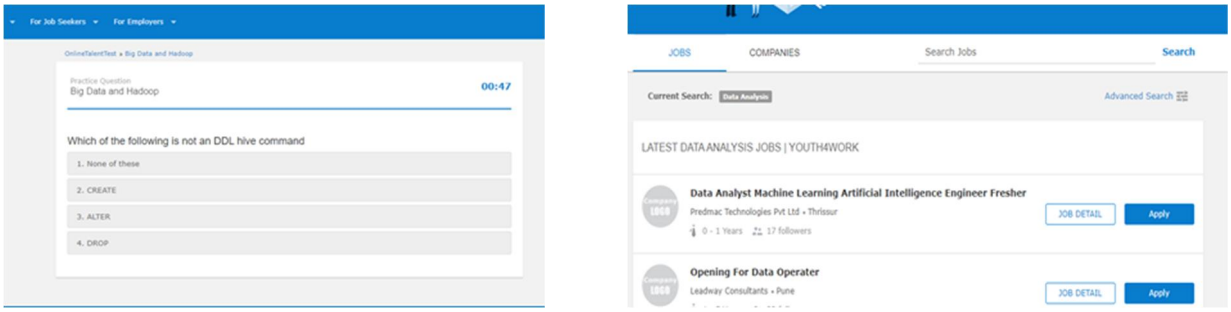


Рис. 8. Користувацький інтерфейс Youth4work

Сучасні онлайн-платформи для аналізу професійних навичок і підбору кар'єрного шляху відіграють важливу роль в управлінні людськими ресурсами, особливо в умовах цифрової трансформації бізнесу. Одним із прикладів таких рішень є система Youth4work, яка орієнтована на молодіжну аудиторію, але має великий потенціал і в корпоративному середовищі.

Ця платформа дозволяє користувачам проходити тести на знання у певній галузі та, на основі результатів, отримувати персоналізоване резюме, яке надалі розповсюджується серед компаній, що зацікавлені у талановитих кадрах. Завдяки цьому формується прозорий та ефективний механізм найму, що базується на перевірених навичках кандидата.

Серед переваг системи слід відзначити її багатофункціональність, можливість самостійно оцінити рівень знань у конкретній сфері, а також пряму інтеграцію з ринком праці, що дозволяє миттєво зреагувати на запити бізнесу. Проте існують і певні обмеження. Зокрема, необхідність попереднього тестування може стримувати частину користувачів, відсутність чіткого списку професійних вимог обмежує глибину аналітики, велика кількість реклами впливає на зручність роботи з платформою, а також не враховуються особистісні характеристики спеціаліста, які є важливими для цілісної оцінки кандидатів у системах управління талантами.

Ще одним напрямом, що активно використовується в освітніх і профорієнтаційних цілях, є системи визначення майбутньої професії. Яскравим прикладом є платформа What Career is Right for Me, яка після проходження спеціального тесту надає рекомендації щодо найбільш відповідних професій. Система враховує чотири ключові блоки інформації: індивідуальні сильні й слабкі сторони, переваги у видах діяльності та навчальних предметах, очікування від робочого середовища, а також кар'єрні цінності. Такий підхід дозволяє створити індивідуальний профіль, що може бути корисним як для особистого планування кар'єри, так і для HR-фахівців, які працюють з молодими кадрами або здійснюють стратегічне планування розвитку персоналу (рис 9).

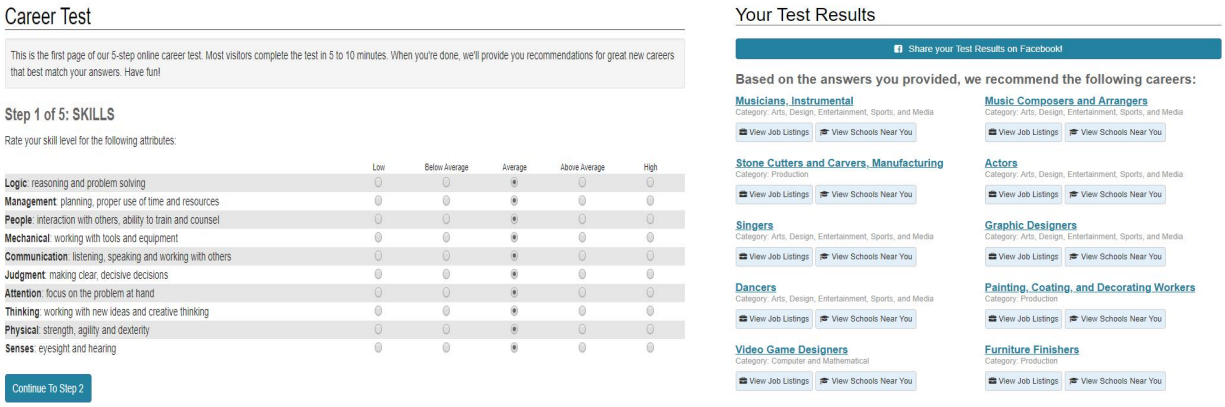


Рис. 9. Користувацький інтерфейс What career is right forme

Система, що була проаналізована, демонструє низку ключових переваг, особливо у контексті професій, пов'язаних з інформаційними технологіями. Насамперед, вона охоплює широкий спектр IT-напрямків – від розробки програмного забезпечення до кібербезпеки та управління IT-проєктами. Це дозволяє користувачам оцінити власні перспективи у різних секторах цифрової індустрії.

Однією з важливих переваг є те, що після проходження тестування користувач отримує доступ до даних щодо середнього рівня заробітної плати в кожному з IT-напрямів, що сприяє формуванню обґрунтованого уявлення про ринок праці та можливості професійного зростання.

Інша сильна сторона системи полягає у багатофакторному підході до аналізу відповідності професії: вона враховує не лише технічні навички (як-от програмування, знання баз даних, системного адміністрування), а й низку м'яких навичок, включаючи критичне мислення, командну роботу, адаптивність до швидкозмінного IT-середовища та здатність до вирішення задач під тиском – що є критично важливими рисами успішного IT-фахівця.

Разом із тим, було виявлено недоліки, що обмежують ефективність системи. Наприклад, відсутній деталізований опис особистісних характеристик, необхідних для окремих IT-ролей. Система не дає змоги миттєво визначити, чи пасує кандидат для ролі, скажімо, DevOps-інженера або data-аналітика. Крім того, повний доступ до професійної інформації можливий лише після проходження тесту, що зменшує гнучкість та зручність взаємодії. Інформація про професії іноді подається загально, без розкриття специфіки діяльності, що ускладнює прийняття рішень.

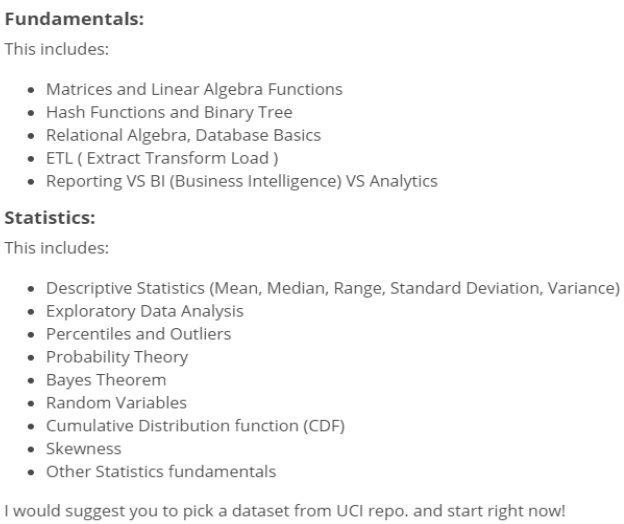
На основі аналізу аналогічних систем були визначені ключові напрями вдосконалення платформи IT-профорієнтації. По-перше, необхідно використовувати велику вибірку актуальних ринкових даних, що забезпечить високу точність результатів. По-друге, слід включати не лише технічні компетенції, а й психологічні характеристики, як-от ініціативність, гнучкість мислення, вміння швидко навчатися.

Також важливим є застосування інтерактивної візуалізації результатів – графіків, рейтингів, діаграм – що значно спрощує сприйняття даних. Важливість кожної компетенції доцільно оцінювати за шкалою від 1 до 10, що дозволяє чітко встановити рівень відповідності кандидата до тієї чи іншої IT-спеціальності.

Водночас важливо уникати перенасичення інтерфейсу інформацією – кількість критеріїв має бути збалансованою: достатньою для обґрунтованого вибору, але не надмірною, щоб не створювати інформаційне перевантаження.

У сучасних умовах постійного зростання обсягів даних особливої актуальності набуває аналіз зворотного зв'язку з відкритих джерел, зокрема – технічних форумів, блогів, професійних спільнот, соціальних мереж. Для системи профорієнтації IT-фахівців такий підхід дозволяє формувати динамічний портрет спеціаліста, адаптований до вимог ринку.

Процес інтелектуального аналізу текстів передбачає виділення ключових термінів та фраз, що описують актуальні компетенції. Для цього використовуються англомовні статті з ключовими словами “IT skills”, “tech careers”, “programming abilities”, обсягом понад 300 рядків, опубліковані після 2022 року (рис. 10). Одним із прикладів є матеріал з платформи edureka.co під назвою *“Top IT Skills You Need in 2023”*, який містить структуровані дані про затребувані IT-навички. Аналіз таких джерел дозволяє формувати об'єктивні рекомендації для кар’єрного розвитку у сфері інформаційних технологій.



Fundamentals:

This includes:

- Matrices and Linear Algebra Functions
- Hash Functions and Binary Tree
- Relational Algebra, Database Basics
- ETL (Extract Transform Load)
- Reporting VS BI (Business Intelligence) VS Analytics

Statistics:

This includes:

- Descriptive Statistics (Mean, Median, Range, Standard Deviation, Variance)
- Exploratory Data Analysis
- Percentiles and Outliers
- Probability Theory
- Bayes Theorem
- Random Variables
- Cumulative Distribution function (CDF)
- Skewness
- Other Statistics fundamentals

I would suggest you to pick a dataset from UCI repo. and start right now!

Рис. 10. Приклад неструктурованих даних

Вихідна інформація системи представлена у вигляді професіограми спеціаліста зі штучного інтелекту, що містить повний опис необхідних знань, умінь і навичок, актуальних для роботи у сфері ШІ. Супровідними є проміжні файли результатів, які включають таблиці з оцінками компетенцій, а також візуалізації у вигляді графіків та діаграм.

Зокрема, на рис. 11 показано фрагмент HTML-файлу, що ілюструє розподіл ключових слів по темах. У цьому інтерфейсі реалізовано можливість інтерактивного налаштування порогу релевантності у межах від 0 до 1, тобто

$$0 \leq \text{релевантність} \leq 1, \tag{2}$$

де значення цього коефіцієнта дозволяє фільтрувати ключові слова відповідно до їх значущості у кожній темі.

Цей файл також демонструє кластеризацію інформації за тематичними групами, в результаті чого було виявлено провідні терміни, серед яких:
“python”, “machine learning”, “visualization”, “statistics”, “sql”, “programming” тощо.

З математичної точки зору, така класифікація може ґрунтуватися на методах Latent Dirichlet Allocation (LDA) або інших моделях тематичного моделювання, де ймовірність P(w|t)P(w|t)P(w|t) — це ймовірність появи слова w у темі t, а показник релевантності можна трактувати як:

$$\text{релевантність}_{w,t} = \frac{P(w \mid t)}{\sum_{w' \in T} P(w' \mid t)}, \tag{3}$$

що дозволяє нормалізувати внесок окремих термінів у межах теми.

Вихідними даними системи є професіограма спеціаліста зі штучного інтелекту – інформація про необхідні навички для роботи в цій галузі та файли з проміжними результатами. До них належать декілька файлів з оцінками та діаграми.

На рис. 11 зображено частину html-файлу, в якому можна подивитись розподіл ключових *с*лів по темах з можливістю змінювати показник релевантності в межах від 0 до 1. Також представлені основні виділені теми. З рис. 11 можна побачити, що знайдено такі основні слова як “python”, “machine learning”, “visualization”, “statistics”, “sql”, “programming” і т.д.

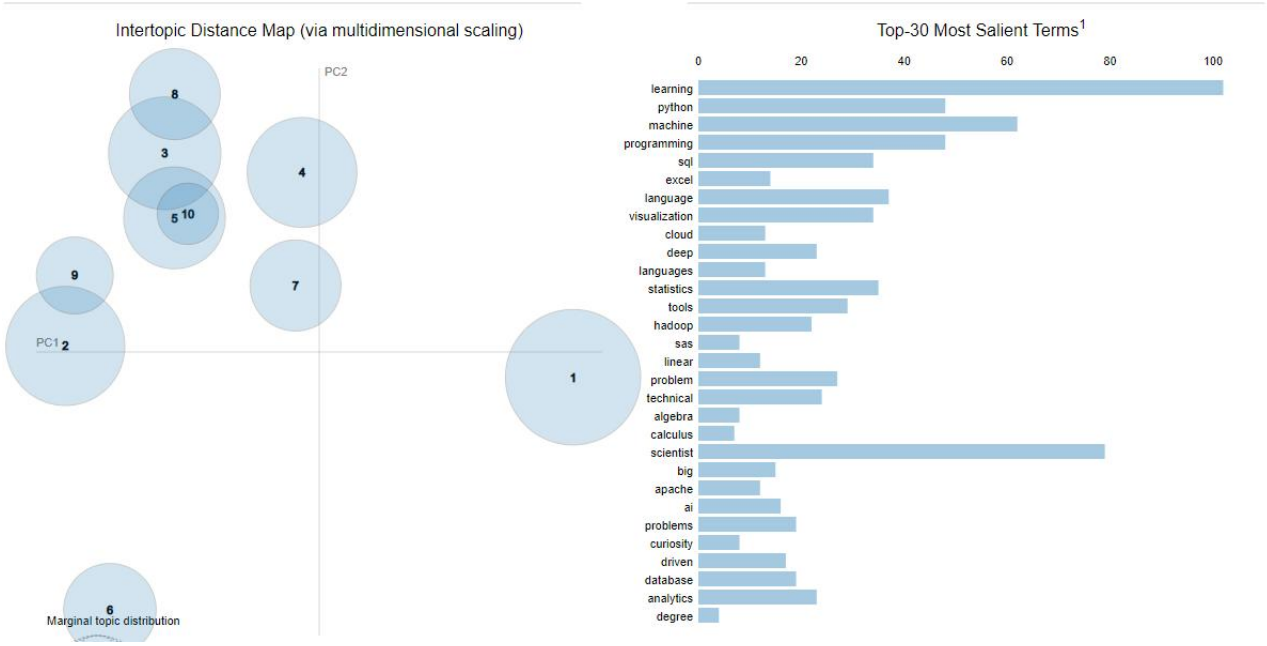


Рис. 11. Визначення ключових слів по темах

Другим за значенням є файл, який містить розподіл тем у документі (див. рис. 12). Цей файл є результатом тематичного моделювання текстового корпусу та дозволяє оцінити, які теми і з якою вагою присутні в кожному фрагменті документа. Таке подання особливо корисне для подальшого аналізу релевантності тексту до конкретних професійних напрямів.

З математичної точки зору, розподіл тем у документі описується **вектором ймовірностей тем** $\vec{\theta}_d$, де кожен елемент $\theta_{d,t}$ відповідає ймовірності того, що тема t присутня у документі d:

$$\theta_d = [\theta_{d,1}, \theta_{d,2}, \dots, \theta_{d,T}], \text{де} \sum_{t=1}^{T} \theta_{d,t} = 1 \quad (4)$$

Тут:

- T – загальна кількість виділених тем,
- $\theta_{d,t} \in [0,1]$ – частка, яку займає тема t у документі d.

У випадку використання моделей, як-от LDA (Latent Dirichlet Allocation), ці ймовірності отримуються шляхом апроксимації багатовимірного розподілу тем у тексті. Графічне представлення на рис. 12 може бути реалізоване у вигляді горизонтальних гістограм або stacked bar charts, що дозволяє візуально оцінити домінантні теми в окремих частинах тексту.

Крім того, для підвищення точності оцінки корисно використовувати метрики, як-от перплексія або кохерентність тем (topic coherence), що дозволяє кількісно оцінити якість класифікації тем.

```
Topic #0: visualization driven tools statistics problem make things analysis technical want able communication help communicating companies dataset solving includes analytical
Topic #1: learning machine deep algorithms regression supervised unsupervised works techniques decision good knowledge linear different ai logistic random nearest advanced moc
Topic #2: big software apache engineering hadoop loading processing spark process strong tools development sources variety missing generated handling background lot involves
Topic #3: linear calculus algebra function analysis skill python large effective source best ai multivariable prep problems approach answers desire functions identify
Topic #4: python programming language excel sql knowledge tools good languages tableau ms tool expertise certain suggest pretty hadoop think box spark
Topic #5: thinking critical problem skill questions analyze look problems results constantly resources trends verbal technical action order able recognize non insights
Topic #6: problems solve industry database management skill great field manipulate systems programs translate manage able languages support different operate analyze deeply
Topic #7: hbase learning communication general tools sas apply replace running accurate monitoring domain search having solutions communicate literally supervised combines rec
Topic #8: cloud computing skill drive services models end aws answer hand order azure company platforms answers analyzing results effective includes easily
Topic #9: networks neural learning deep creating used models keras fundamentals library tensorflow curiosity visualization knowledge ask questions maps artificial able curious
```

Рис. 12. Фрагмент файлу з розподілом тем

На основі результатів програми будується діаграма фахівця зі штучного інтелекту, гілки якої пронормовані оцінками певних навичок. Професіограма зображена на рис. 13.

Рис. 13. Професіограма для фахівців зі штучного інтелекту

Упродовж останніх років було створено колосальні обсяги текстової інформації, і ефективна робота з нею стала ключовою навичкою для сучасного фахівця в галузі аналізу даних. Щоб здійснити змістовний аналіз текстів, необхідно попередньо підготувати дані, адже "сирий" текст часто містить зайві або нерелевантні елементи.

На рис. 14 схематично представлено етапи попередньої обробки текстових даних, які є обов'язковими перед подальшим аналізом чи побудовою моделей машинного навчання.

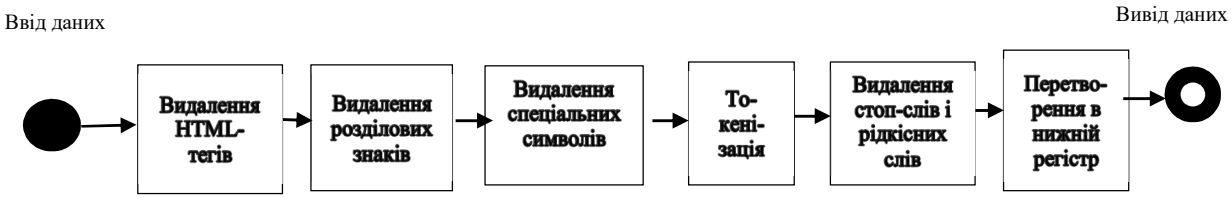


Рис. 14. Етапи підготовки вхідних даних

Основні з них:

- Видалення HTML-тегів та спеціальних символів
- Видалення пунктуації
- Токенізація – розбиття тексту на окремі слова (токени)
- Нормалізація – приведення слів до нижнього регістру
- Видалення стоп-слів (наприклад, "і", "але", "це") та рідкісних слів
- Виправлення орфографії або лематизація/стемінг (зведення слова до базової форми)

Алгоритм процесу, представленого на Рис. 14 має таке математичне пояснення:

Нехай маємо текстовий документ D, який складається з послідовності слів: $D = [w_1, w_2, \dots, w_n]$.

Процес попередньої обробки можна математично представити як функцію перетворення:

$$f_{preprocess}(D) \to D'$$

Цей процес може включати кілька послідовних операцій:

1. Фільтрація символів:

$$D_1 = \{w_i \in D : w_i \notin HTML, \text{знаки, спец. символи}\}$$

2. Видалення стоп-слів:

$$D_2 = \{w_i \in D_1 : w_i \notin StopWords\}$$

3. Застосування лематизації або стемінгу:

$$D_3 = \{Lemma(w_i) : w_i \in D_2\}$$

4. Нормалізація регістру:

$$D_4 = \{Lowercase(w_i) : w_i \in D_3\}.$$

У результаті отримуємо компактне й осмислене представлення тексту: $D' = D_4$

Розглянемо один фрагмент тексту зі статті (наприклад, перший елемент списку, як показано на рис. 15) і застосуємо до нього всі етапи попередньої обробки, відповідно до схеми на рис. 14. Це дозволить продемонструвати на практиці, як "сирий" текст перетворюється на структуровану та придатну до аналізу інформацію.

```
<li style="text-align: justify;">
    Data Science is about using capital processes, algorithms, or systems to extract knowledge, insights, and make informed decisions
    In that case, making inferences, estimating, or predicting form an important part of Data Science.
    Probability with the help of statistical methods helps make estimates for further analysis.
    Statistics is mostly dependent on the theory of probability. Putting it simply, both are intertwined.
    What can you do with Probability and Statistics for Data Science?
    Explore and understand more about the data
    Identify the underlying relationships or dependencies that may exist between two variables
    Predict future trend or forecast a drift based on the previous data trends
    Determine patterns or motive of the data
    Uncover anomalies in data
</li>
```

Рис. 15. Фрагмент статті до обробки

І етап – видаляються HTML-теги.

Data Science is about using capital processes, algorithms, or systems to extract knowledge, insights, and make informed decisions

In that case, making inferences, estimating, or predicting form an important part of Data Science.

Вилучаємо `<li>`, `</li>`, `style`, переноси рядків – залишаємо лише текст**.**

Data Science is about using capital processes, algorithms, or systems to extract knowledge, insights, and make informed decisions. In that case, making inferences, estimating, or predicting form an important part of Data Science. Probability with the help of statistical methods helps make estimates for further analysis.

Якщо дані отримані з обробки веб-документів, часто в наборі даних можуть залишитись HTML-теги. HTML призначений для оформлення текстів на веб-сторінках, що не допомагає при побудові моделі, яка буде використовуватись для подальшого аналізу. В результаті видалення тегів стаття матиме такий вигляд (рис. 16).

```
Data Science is about using capital processes, algorithms, or systems to extract knowledge, insights, and make informed decisions
In that case, making inferences, estimating, or predicting form an important part of Data Science.
Probability with the help of statistical methods helps make estimates for further analysis.
Statistics is mostly dependent on the theory of probability. Putting it simply, both are intertwined.
What can you do with Probability and Statistics for Data Science?
Explore and understand more about the data
Identify the underlying relationships or dependencies that may exist between two variables
Predict future trend or forecast a drift based on the previous data trends
Determine patterns or motive of the data
Uncover anomalies in data
```

Рис. 16. Фрагмент статті з видаленими HTML-тегами

II та III етапи – видаляють знаки пунктуації та спеціальні символи – вони мають сенс для людей, але не для машин. Видалення спеціальних символів означає збереження тільки алфавіту в текстових даних. Фрагмент статті після видалення знаків на рис. 17.

T=[“Data”, “Science”, “is”, “about”, “using”, “capital”, “processes”,..., “data”]

```
Data Science is about using capital processes algorithms or systems to extract knowledge insights and make informed decisions from data
In that case making inferences estimating or predicting form an important part of Data Science
Probability with the help of statistical methods helps make estimates for further analysis
Statistics is mostly dependent on the theory of probability Putting it simply both are intertwined
What can you do with Probability and Statistics for Data Science
Explore and understand more about the data
Identify the underlying relationships or dependencies that may exist between two variables
Predict future trend or forecast a drift based on the previous data trends
Determine patterns or motive of the data
Uncover anomalies in data
```

Рис. 17. Фрагмент статті з видаленими символами

IV етап – Зведення всіх слів до маленьких літер.

T_{lower}= [“data”, “science”, “is”, “about”, “using”,..., “data”]

V етап – токенізація, тобто переробка тексту в нормовані списки слів. Для цього застосовано алгоритм TF-IDF (Рис. 18).

data	0,026
science	0
is	0
about	0,00545
using	0,00685
capital	0,00685
processes	0,00685
algorithm	0,00358
or	0,00332
systems	0,00273
to	0
extract	0,00685
knowledg	0,00358
insights	0,00273
and	0
make	0,0137
informed	0,00685
decisions	0,00685
from	0,00152
in	0,00063
that	0,00212
case	0,00479
making	0,00716
inference	0,00685
estimatin	0,00685
predicting	0,00685
form	0,00479
an	0,00273
important	0,00479
part	0,00358
of	0
with	0
the	0
help	0,00479
statistical	0,00685
methods	0,00479
helps	0,00479
estimates	0,00685
for	0
further	0,00479
analysis	0,00273
mostly	0,00685
depender	0,00685
on	0,00818
theory	0,00685
putting	0,00685
it	0,00106
simply	0,00685
both	0,00685
are	0,00304
intertwine	0,00685
what	0,00031
can	0

Рис. 18. Фрагмент статті, переробленої в нормований список

VI етап – видалення стоп-слів, слів-паразитів та рідкісних слів – в англійському тексті це слова is, am, are, and so, on. Такі слова також мають бути видалені, адже частота їхнього повторення заважає визначенню дійсно важливих даних. Для визначення таких слів існують спеціальні алгоритми, за допомогою яких за вагою слів визначається важливість певного слова в документі. За порахованими по TF_IDF значеннями були видалені всі слова, що мають або дуже низьку, або дуже високу вагу. Отже із зображеного на Рис. 18 залишаються тільки такі слова: “data about using capital processes algorithms systems extract knowledge insights make informed decisions case making form important part statistical methods helps estimates further analysis mostly dependent”.

VII етап – стемінг. При роботі з природною мовою може настати момент, коли необхідно, щоб програма розпізнавала, що слова “ask” і “asked” - це просто різні часи одного і того ж дієслова. Слова, похідні одне від одного, можуть бути співставлені з центральним словом або символом, особливо якщо вони мають спільне основне значення. Це може бути використано при пошуку інформації, коли необхідно прискорити отримання результатів роботи алгоритму або при спробі проаналізувати використання слів в документі може виникнути бажання стиснути спільнокореневі слова, щоб не виникало сильної мінливості. У будь-якому випадку, ці техніка нормалізації тексту є корисною. Стемінг – це ідея скорочення різних форм слова до основного кореня. Алгоритми стемінгу зазвичай засновані на правилах. Їх можна розглядати як евристичний процес, що відсікає закінчення слів. Слово проходить через ряд умов, які визначають, як його скоротити. Стемінг для англомовних слів був виконаний за допомогою видалення визначених заздалегідь закінчень, таких як “-s”, ” –es”, “-ed” тощо. Отже після всіх обробок текст матиме такий вигляд: “data about using capital process algorithm system extract knowledge insight make informed decisions case make form important part statistical method help estimate further analysis mostly dependent”.

Для проєктування функціональної моделі системи в першу чергу необхідно побудувати контекстну діаграму. Контекстна діаграма є своєрідною чорною скринькою, що показує систему як єдиний процес вищого рівня та зображає зв’язки з іншими об'єктами із зовнішнього середовища (системами, зовнішніми сховищами даних тощо).

Контекстна діаграма розробки професіограми за допомогою використання алгоритмів обробки неструктурованих текстових даних зображена на рис. 19.

Основні інформаційні потоки, що відображаються в побудованій контекстній діаграмі, можна описати наступним чином:

1. Вхідна інформація — це дані, зібрані з веб-ресурсів, що містять актуальні відомості про спеціалістів у галузі штучного інтелекту. Ці дані слугують основою для подальшого аналізу.
2. Керуючі параметри (або критерії відбору) включають:

- Обсяг джерела: статті обираються лише за умови, що кількість сторінок перевищує 50;
- Актуальність: дата публікації має бути пізніше 2018 року;
- Семантична релевантність: у статтях обов’язково мають зустрічатися ключові слова “data scientist skills”.

Рис. 19. Контекстна діаграма побудови професіограми

3. Вихідна інформація – це професіограма спеціаліста з AI (штучного інтелекту), побудована на основі проаналізованих даних і виявлених зв’язків між ключовими компетенціями.

Контекстна діаграма виконує роль узагальненої моделі системи, що дозволяє побачити, як відбувається обробка та розповсюдження інформації між системою, її модулями та зовнішнім середовищем. Це спрощене графічне представлення взаємозв'язків і потоків даних.

Основна ідея побудови такої діаграми полягає в тому, що всі потоки мають починатися та завершуватись у межах певного процесу обробки, оскільки дані самостійно не трансформуються без участі функціонального елемента.

Результатом є зрозумілий інструмент для комунікації між розробниками системи та її кінцевими користувачами, представлений на рис. 20.

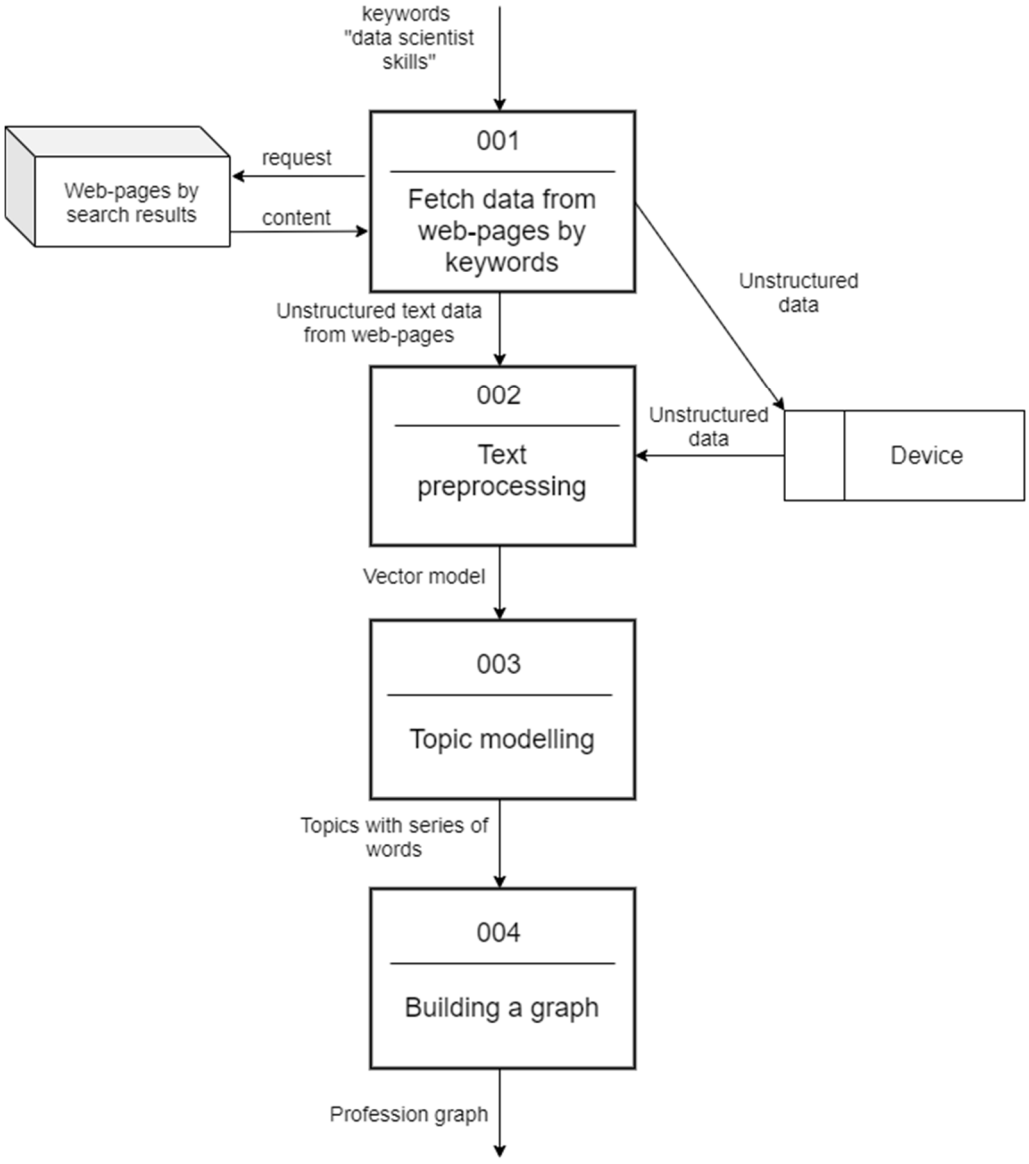


Рис. 20. Діаграма потоків даних системи

З діаграми можна побачити, що система побудови професіограми має такі процеси:

1) Видобування даних з веб-сторінок – відправляється запит з ключовими словами до API та в результаті приходить відповідь з даними згідно результатів пошуку;
2) Попередня обробка текстових даних – до неструктурованої інформації застосовуються методи обробки даних, які дозволяють нам в кінці отримати структуровані дані у вигляді векторної моделі;

3) Тематичне моделювання – до структурованих даних у вигляді вектору застосовується алгоритм, в результаті роботи якого буде отримано список тем з наборами слів;

Побудова графу – на основі тем з наборами слів визначено найвагоміші слова та побудовано професійну діаграму для фахівця штучного інтелекту

Одним з найосновніших процесів у системі є перший процес – видобування даних з веб-сторінок.

Детальніша діаграма процесу збирання даних для побудови професіограми з веб-сторінок зображена на рис. 21.

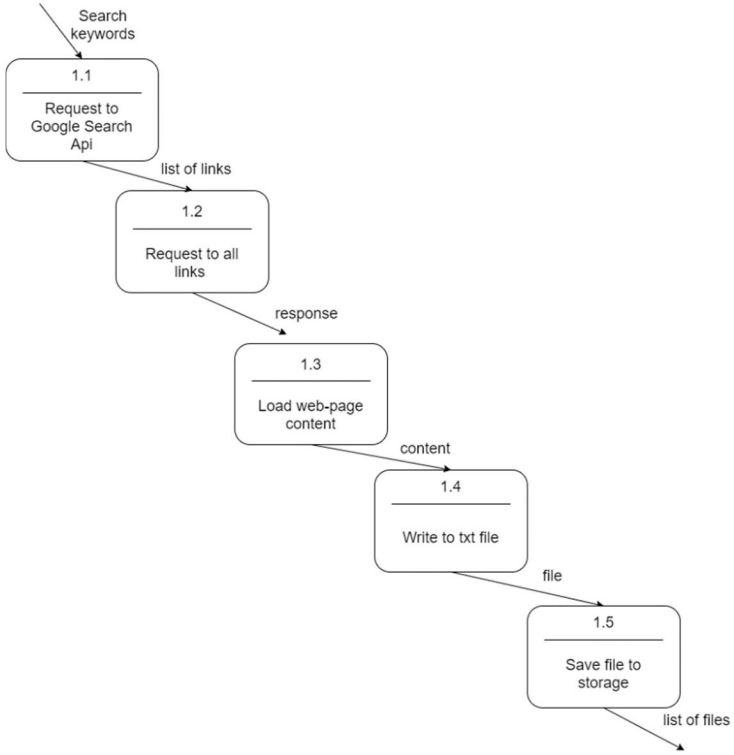


Рис. 21. Деталізація першого процесу

Основними підпроцесами є:

1) Надсилання запиту на отримання списку всіх веб-сторінок до API за словами “data scientist skills” та отримую 220 мільйонів сторінок, з них обирається 50 перших результатів, де статті написані не раніше 2022 року;
2) Виконується запит до обраного посилання, в результаті отримується відповідь від серверу;
3) Завантажується зміст веб-сторінок з серверу;
4) Записуються завантажені з серверу дані до текстового файлу;
5) Зберігаються текстові файлу у сховище.

Кроки 2-4 виконуються циклічно відповідно до обраної кількості файлів.

Для реалізації поставленого завдання першочерговим етапом є перетворення неструктурованого тексту у форму, зручну для математичної обробки – а саме, у векторне представлення. Це дозволяє застосовувати статистичні або машинні методи для аналізу текстових даних.

Одним із найефективніших способів такої трансформації є TF-IDF (Term Frequency – Inverse Document Frequency) – метод статистичного зважування, який оцінює важливість слова у межах одного документа та в контексті колекції документів (корпусу).

Основи математичного апарату TF-IDF

1. TF (Частота терміну):

Це частота появи терміну в документі, нормалізована на загальну кількість слів у цьому документі:

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}}, \quad (5)$$

де:

- $f_{t,d}$ – кількість входжень терміну t у документі d,
- $\sum_k f_{k,d}$ – загальна кількість усіх термінів у документі d.

2. IDF (Обернена частота документа):

Оцінює рідкість терміну в колекції документів:

$$IDF(t,D) = log\left(\frac{N}{1 + |\{d \in D: t \in d|}\right), \quad (6)$$

де:

- N – загальна кількість документів у корпусі D,
- $|\{d \in D: t \in d|$ – кількість документів, у яких зустрічається термін t.

3. TF-IDF:

Інтеграція двох складових – оцінка важливості слова в межах документа і корпусу загалом:

$$TF - IDF(t,d,D) = TF(t,d) \cdot IDF(t,D). \quad (7)$$

Частота (TF) – використовується для знаходження кількості разів використання певного терміну в документі. Є загальновідомим фактом, що довжина документів може варіюватися від дуже маленької до великої, тому ймовірність того, що будь-який термін зустрічатиметься частіше у великих документах в порівнянні з невеликими документами. Отже, для вирішення цієї проблеми, входження будь-якого терміна в документ ділиться на загальну кількість термінів, представлених в цьому документі, щоб знайти частоту. Розрахована частота фрагменту даних для побудови професіограми зображена на рис. 22.

data	0,04128	statistics	0,0274	you	0,01724	mathematical	0,33333	wikipedia	0,01818	is	0,02564
scientists	0,00917	data	0,08219	need	0,00431	reasoning	0,33333	defines	0,01818	central	0,01282
are	0,01835	science	0,0274	to	0,03879			it	0,03636	to	0,03846
highly	0,00459	is	0,0137	have	0,00431			as	0,01818	the	0,07692
educated	0,00459	about	0,0137	the	0,04741	Paragraph number 1		the	0,03636	growth	0,01282
88	0,00459	using	0,00685	knowledge	0,00862			study	0,01818	of	0,03846
have	0,01835	capital	0,00685	of	0,03448	business	0,25	of	0,03636	ai	0,03846
at	0,00459	processes	0,00685	languages	0,00862	communicatio	0,25	collection	0,01818	and	0,07692
least	0,00459	algorithms	0,00685	like	0,00431	leadership	0,25	analysis	0,03636	data	0,05128
a	0,04587	or	0,03425	python	0,00862	and	0,25	interpretation	0,01818	science	0,02564
master	0,01376	systems	0,00685	perl	0,00431			presentation	0,01818	arguably	0,01282
s	0,01835	to	0,0137	c	0,00862			and	0,03636	most	0,01282
degree	0,03211	extract	0,00685	sql	0,00431	Paragraph number 2		organization	0,01818	important	0,01282
and	0,03211	knowledge	0,00685	and	0,03448			data	0,07273	skill	0,01282
46	0,00459	insights	0,00685	java	0,00431	programming	0,03846	therefore	0,01818	learn	0,03846
phds	0,00459	and	0,03425	with	0,01293	you	0,03846	shouldn	0,01818	for	0,03846
while	0,00459	make	0,0137	being	0,00862	ll	0,03846	t	0,01818	exploration	0,01282
there	0,00459	informed	0,00685	most	0,00862	often	0,03846	be	0,01818	programming	0,02564
notable	0,00459	decisions	0,00685	common	0,00431	be	0,03846	a	0,03636	language	0,02564
exceptions	0,00459	from	0,00685	coding	0,00431	tasked	0,03846	surprise	0,01818	a	0,01282
very	0,00459	in	0,0137	language	0,00431	with	0,03846	that	0,01818	favorite	0,01282
strong	0,00459	that	0,0137	required	0,01293	leading	0,03846	scientists	0,01818	scientists	0,01282
educational	0,00459	case	0,00685	in	0,00862	data	0,07692	need	0,01818	web	0,01282
background	0,00459	making	0,0137	data	0,05172	science	0,03846	to	0,01818	developers	0,01282
is	0,00917	inferences	0,00685	science	0,00431	projects	0,03846	know	0,01818	experts	0,01282

Рис. 22. Розрахована частота для текстових даних

Для того щоб оцінити інформативність термінів у корпусі документів, важливо не лише враховувати, як часто слово з'являється в одному документі, але й наскільки воно унікальне в межах усього набору документів.

Частота в документах (Document Frequency) підраховує, у скількох документах загалом зустрічається кожне слово. Але, якщо певний термін трапляється майже в кожному документі, він втрачає здатність відображати специфіку окремих текстів. Наприклад, загальновживані слова типу “on” чи “and” часто не несуть корисного змісту.

Щоб вирішити цю проблему, вводиться поняття зворотної частоти документа (IDF) – воно допомагає оцінити унікальність терміну в межах всієї колекції документів. Ідея проста: якщо слово зустрічається великій кількості документів, воно отримує меншу вагу, а якщо лише в кількох – більшу.

Таким чином, IDF дозволяє зменшити вплив тривіальних або повторюваних слів і навпаки – підкреслити терміни, які можуть краще ідентифікувати вміст документа.

Частота зворотного документа призначає меншу вагу словам, які часто з'являються у всіх документах і призначають більшу вагу для слів, які не часто повторюються серед всіх документів. Інакше твердження про IDF полягає в тому, що він встановлює механізм оцінки корисності обраного терміну – більш високі бали означають, що більший вклад інформації в зміст несе дане слово. Зворотну частоту можна розрахувати як:

$$IDF = log\_e\,(n/k), \tag{8}$$

де n – загальна кількість документів, k – кількість документів, в яких зустрічається слово w. Розрахована зворотня частотат для фрагменту одної статті на рис. 23.

education	1,9432		probability	1,9459
data	0		statistics	2,6391
scientists	0,84729		data	0,3364
are	0,6931		science	0,3364
highly	2,6391		is	1,0296
educated	2,6391		about	1,2527
have	0,6931		using	2,6391

Рис. 23. Розрахована зворотна частота

Сама формула TF-IDF – це не що інше, як множення частоти терміну (TF) і зворотної частоти цього терміну (IDF). TF-IDF можна визначити як:

$$TF - IDF = TF * IDF \tag{9}$$

Наступним кроком є використання LDA (алгоритм латентного розміщення Дирихле) – це навчання без учителя, при якому документи розглядаються як набори слів (де їхній порядок не має значення). LDA працює, спочатку припускаючи існування певного ключа: початок створення документа полягав у виборі набору певних тем, а потім в кожній темі береться певний набір термінів.

Спочатку розглядаю розподіл Дирихле. Він визначається як:

$$f = (x_1,\ldots,x_k;\alpha_1,\ldots,\alpha_k) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\Pi_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}X_i^{\alpha}i^{-1}, \tag{10}$$

де гама-функція

$\Gamma(n) = (n-1)!$ – для дискретних змінних

$\Gamma(z) = \int_0^{\infty} x^{z-1}e^{-z}\,dx$ – для багатовимірних змінних

Для спільної ймовірності з m змінними вихідні дані Дирихле є m-мірними, і для їх моделювання потрібно m параметрів. Наприклад, модель $p\ (x_1, x_2, x_3, x_4)$ матиме параметри моделі α_1, α_2, α_3 і α_4. У LDA моделюється розподіл слів для кожної теми та їх коефіцієнтів для кожного документа, використовуючи розподіл Діріхле. Наприклад, вивчаючи α_i, вивчаються пропорції тем, які може охоплювати документ. Чим вище значення для α_i щодо інших, тим більша ймовірність, що така тема буде обрана.

У розподілі Дирихле вибіркове значення з p є розподілом ймовірності (наприклад, $x_1 = 0.5$, $x_2 = 0.3$, $x_3 = 0.1$ і $x_4 = 0.1$), яке завжди в сумі дає одиницю. Саме тому розподіл Дирихле називається розподілом розподілів. У цьому розподілі є окремий випадок, який називається симетричним розподілом Дирихле. Всі α_i дорівнюватимуть тоді і тільки тоді, коли використовуємо один скаляр α для представлення моделі. Коли значення α зменшується, розрідженість збільшується, тобто більшість значень в вибіркових даних дорівнюватиме нулю або буде близькою до нуля.

Ця розрідженість може бути параметризована в будь-якому розподілі Дирихле, а не тільки в симетричному. Це дозволяє контролювати розподіли слів або розрідженість пропорцій тем. Також це дозволяє побудувати модель, в якій документ повинен охоплювати лише невелику кількість тем (висока розрідженість).

Кроки в створенні документа наступні: спочатку є V-мірний розподіл Дирихле для розподілу слів по кожній темі (β_k). V - розмір словника, і всього наявні k тем. На другому кроці зображаю розподіли пропорцій тем для кожного документа. На третьому кроці визначаю теми для кожного слова і малюю слово з розподілу слів для відповідної теми. Для виведення загальної ймовірності, враховуючи всі спостереження,виводяться приховані змінні або приховані фактори θ, z, β, спостерігаючи за набором документів, тобто знаходячи $p(\theta, z, \beta \mid w)$.

$$p\left(\vec{\theta}_{1:0}, z_{1:B,1:N}, \vec{\beta}_{1:K} \middle| w_{1:D,1:N}, \alpha, \eta\right) \tag{11}$$

Для кожного документа відстежую пропорції теми θ і найбільш ймовірне призначення теми z для кожного слова документа. У подальших кроках фіксується розподіл слів β для кожної теми. Знайдемо найбільш ймовірну пропорцію теми θ і z для кожного слова документа. Для цього визначено розподіл ключових термінів по темам (рис. 24).

```
Topics in LDA model:
Topic 0:
[('visualization', 48.68926203785327), ('driven', 30.18134449114038), ('tools', 27.673879770912418), ('statistics', 27.235100980963207),
 ('problem', 21.883769120417803), ('make', 20.720611273851407), ('things', 18.036616756570627), ('analysis', 17.23680408005126),
 ('technical', 17.03944183275965), ('want', 14.974295066689933), ('able', 14.659454519138055), ('communication', 14.553629275853464),
 ('help', 13.74104907070067), ('communicating', 13.48640345387339), ('companies', 13.321558846312133), ('dataset', 13.236773511523609),
 ('solving', 13.217976212177186), ('includes', 13.043442776241786), ('analytical', 12.862635722299439), ('tableau', 12.289883412107706)]
Topic 1:
[('learning', 87.47114763038992), ('machine', 44.57812193002455), ('deep', 22.961152593760755), ('algorithms', 13.281534430944038),
('regression', 13.193194793310752), ('supervised', 10.92709264878274), ('unsupervised', 9.897143623017087), ('works', 8.426590421300581),
('techniques', 7.582088228914892), ('decision', 6.439413406506067), ('good', 6.395072119317846), ('knowledge', 6.193608976081453),
('linear', 6.043753439366162), ('different', 5.746248118745537), ('ai', 5.67416104301938), ('logistic', 5.654908215613883),
('random', 5.624552344860397), ('nearest', 5.609634280907769), ('advanced', 5.551633682371046), ('models', 5.527194097357434)]
Topic 2:
[('big', 15.03914075914663), ('software', 15.037228283347723), ('apache', 13.406472777155422), ('engineering', 12.538851536226117),
 ('hadoop', 10.794112616652674), ('loading', 8.841783927425874), ('processing', 6.579220793255257), ('spark', 6.539298082039184),
 ('process', 5.677040227566277), ('strong', 5.649730036406812), ('tools', 5.621949465204662), ('development', 5.531173504189919),
 ('sources', 5.038510632748061), ('variety', 4.939393675440552), ('missing', 4.925982969717612), ('generated', 4.878706792732975),
 ('handling', 4.778777622958517), ('background', 4.776964132085832), ('lot', 4.756280832049772), ('involves', 4.582147842061654)]
Topic 3:
[('linear', 4.369954556703928), ('calculus', 4.046104852572404), ('algebra', 3.7909898502187396), ('function', 2.8394980736545303),
('analysis', 2.157543280625041), ('skill', 1.8844187938958714), ('python', 1.7176687809175464), ('large', 1.6262490529600107),
('effective', 1.5825445636379438), ('source', 1.5461055329692552), ('best', 1.4638720063148216), ('ai', 1.438224197183175),
('multivariable', 1.4213644493497024), ('prep', 1.383670460701581), ('problems', 1.3247405381003894), ('approach', 1.2931758607950643),
('answers', 1.2773780951176985), ('desire', 1.2312906761975744), ('functions', 1.228881529089087), ('identify', 1.2062163607930334)]
Topic 4:
[('python', 30.42639806804373), ('programming', 29.758037612446874), ('language', 16.153732934381424), ('excel', 12.002643597186008),
('sql', 8.76371047493149), ('knowledge', 7.5250479586657), ('tools', 7.3268608948141), ('good', 6.937336587789713), ('languages', 5.7988112
('tableau', 5.446813579702937), ('ms', 5.317119026435173), ('tool', 5.301444702663723), ('expertise', 4.856199687138079),
('certain', 4.682447545567969), ('suggest', 4.498550935769889), ('pretty', 4.459183813939342), ('hadoop', 4.190786388307931),
('think', 4.169600309590668), ('box', 4.021688106123531), ('spark', 3.9701935205619785)]
Topic 5:
```

Рис. 24. Розподіл термінів в темах

Потім таблиця перегортається і об'єднуються результати всіх документів, щоб оновити розподіл слів β для кожної теми. Оскільки θ, β і z взаємозалежні, їх не можна оптимізувати всі відразу. Тому оптимізується по одній змінній за раз, залишаючи інші змінні фіксованими. Однак в LDA θ, β і z будуть моделюватися за допомогою імовірнісних моделей замість точкової оцінки, яка містить тільки найбільш ймовірну оцінку. Ці моделі відстежують всю ймовірність та її достовірність. Таким чином, ми зберігаємо більше інформації в кожній ітерації. Формула загальної ймовірності має такий вигляд:

$$p(\beta, \theta, z, w) = (\prod_{i=1}^{k} p(\beta_i|\eta))(\prod_{d=1}^{D} p(\theta_d|\alpha)\prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:K}, z_{d,n})), \tag{12}$$

де β_k – розподіл слів за темою k, θ – розподіл тем в документі, z – тема, призначена слову в документі.

На рис. 25 зображена вагова характеристика кожного слова в загальному наборі документів. Дана оцінка була розрахована за частотою використання слова в одному елементі списку (на рисунку – paragraph) та кількістю разів, що дане слово зустрілось в інших абзацах документа. З прикладу

можна побачити, що такі слова як “a”, “of”,“to”, “and” (їхня оцінка 0 або близька до 0) є взагалі не важливими і їх можна не враховувати, а слова як “python” (0.57), “algorithms” (0.31), “algebra” (0.52), “analysis” (0.82) мають високі коефіцієнти, отже є важливими для вирахування кінцевого результату.

AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD
skill	0		Paragraph number 1			Paragraph number 1						to	0,01957		is	0,00091		driven	0,00601		solving	0,15904
you	0								computer	0,78989		be	0,02651		a	0,01		organizati	0,00601		skills	0,04255
will	0		constructi	0,54449		creating	0,26035		skills	0,63938		really	0,02651		no	0,0054		your	0,0051			
objectivel	0,016		predictive	0,54449		and	0,18509					good	0,05301		brainer	0,0076		stakehold	0,00457			
analyze	0,00837		models	0,54449		maintainii	0,26035					in	0,03629		you	0,00202		depend	0,01149		Paragraph number 1	
questions	0,00482					algorithm	0,38509		Paragraph number 2			at	0,02651		need	0,00412		on	0,00346			
hypothes	0,008											least	0,05301		to	0,00254		scientist	0,00534		experienc	0,0734
and	0		Paragraph number 2						java	0,27978		one	0,05301		be	0,01081		skills	0		in	0,0367
results	0,00637					Paragraph number 2						also	0,05301		an	0,0076		to	0		database	0,0734
understar	0,00155		creating	0,23335								learn	0,02651		expert	0,0076		help	0,00803		interrogat	0,0734
what	0,00318		controls	0,23335		informati	0,26035		Paragraph number 3			reuse	0,05301		in	0,00505		them	0,00915		and	0,01679
resources	0,00559		to	0,16519		retrieval	0,26035					your	0,02651		either	0,0076		with	0,00053		analysis	0,05025
are	0,00318		assure	0,23335		data	0,18509		matlab	0,51978		code	0,05301		some	0,00412		decision	0,00803		tools	0,0734
to	0		accuracy	0,23335		sets	0,26035								jobs	0,01081		making	0,01149		such	0,0734
solve	0,00418		of	0,23335											will	0,00384		provide	0,01149		as	0,0734
a	0,0011		data	0,08458					Paragraph number 4			Paragraph number 1			list	0,0076		the	0		hadoop	0,0734
problem	0,00482					Paragraph number 3									sas	0,0076		necessary	0,01149		sql	0,0734
look	0,008								microsoft	0,63938		sql	0,10045		or	0,0075		methods	0,00803		sas	0,0734
at	0,00318		Paragraph number 3			linear	0,5207		excel	0,78989		99	0,05022		other	0,01081		dig	0,01149			
problems	0,00482					algebra	0,5207					of	0,03438		obscure	0,0076		deeper	0,01149			
from	0,00559		critical	0,81673								firms	0,05022		languages	0,0076		into	0,01607		Paragraph number 2	
differing	0,008		thinking	0,66622					Paragraph number 5			still	0,10045		but	0,00384		and	0			
views	0,008					Paragraph number 4						use	0,03438		was	0,0152		gain	0,01149		exception	0,04337
perspecti	0,008								perl	0,457978		relational	0,05022		constant	0,0076		valuable	0,01149		communi	0,04337
is	0,00037		Paragraph number 4			machine	0,24679					databases	0,05022		and	0,0006		insights	0,01149		and	0,00496
valuable	0,008					learning	0,18809					and	0,00927		mandator	0,0054		from	0,01149		presentat	0,04337
that	0,00372		data	0,29604		models	0,24679		Paragraph number 6			interview	0,05022		requireme	0,0076		in	0,00223		skills	0,0116
easily	0,00418		analysis	0,81673								include	0,05022		100	0,0076		addition	0,01149		in	0,02169
transfers	0,008								python	0,57978		coding	0,05022		of	0,01236		more	0,02299		order	0,02969

Рис. 25. Таблиця ваг для кожного слова

В результаті роботи програми було створено декілька файлів з результатами, для розуміння яких користувачу необхідна інструкція. Першим файлом є файл з професіограмою. Його побудовано на основі даних з файлу, фрагмент якого зображений на рис. 26.

```
ır_file.txt
Topics in LDA model:
Topic 0:
[('visualization', 48.68926203785327), ('driven', 30.18134449114038), ('tools', 27.673879770912418), ('statistics', 27.235100980963207),
 ('problem', 21.883769120417803), ('make', 20.720611273851407), ('things', 18.036616756570627), ('analysis', 17.23680408005126),
 ('technical', 17.03944183275965), ('want', 14.974295066689933), ('able', 14.659454519138055), ('communication', 14.553629275853464),
 ('help', 13.74104907070067), ('communicating', 13.48640345387339), ('companies', 13.321558846312133), ('dataset', 13.236773511523609),
 ('solving', 13.217976212177186), ('includes', 13.043442776241786), ('analytical', 12.862635722299439), ('tableau', 12.289883412107706)]
Topic 1:
[('learning', 87.47114763038992), ('machine', 44.57812193002455), ('deep', 22.961152593760755), ('algorithms', 13.281534430944038),
('regression', 13.193194793310752), ('supervised', 10.92709264878274), ('unsupervised', 9.897143623017087), ('works', 8.426590421300581),
('techniques', 7.582088228914892), ('decision', 6.439413406506067), ('good', 6.395072119317846), ('knowledge', 6.193608976081453),
('linear', 6.043753439366162), ('different', 5.746248118745537), ('ai', 5.67416104301938), ('logistic', 5.654908215613883),
('random', 5.624552344860397), ('nearest', 5.609634280907769), ('advanced', 5.551633682371046), ('models', 5.527194097357434)]
Topic 2:
[('big', 15.03914075914663), ('software', 15.037228283347723), ('apache', 13.406472777155422), ('engineering', 12.538851536226117),
 ('hadoop', 10.794112616652674), ('loading', 8.841783927425874), ('processing', 6.579220793255257), ('spark', 6.539298082039184),
 ('process', 5.677040227566277), ('strong', 5.649730036406812), ('tools', 5.621949465204662), ('development', 5.531173504189919),
 ('sources', 5.038510632748061), ('variety', 4.939393675440552), ('missing', 4.925982969717612), ('generated', 4.878706792732975),
 ('handling', 4.778777622958517), ('background', 4.776964132085832), ('lot', 4.756280832049772), ('involves', 4.582147842061654)]
Topic 3:
[('linear', 4.369954556703928), ('calculus', 4.046104852572404), ('algebra', 3.7909898502187396), ('function', 2.8394980736545303),
('analysis', 2.157543280625041), ('skill', 1.8844187938958714), ('python', 1.7176678789175464), ('large', 1.6262490529600107),
('effective', 1.5825445636379438), ('source', 1.5461055329692552), ('best', 1.4638720063148216), ('ai', 1.438224197183175),
('multivariable', 1.4213644493497024), ('prep', 1.383670460701581), ('problems', 1.3247405381003894), ('approach', 1.2931758607950643),
('answers', 1.2773780951176985), ('desire', 1.2312906761975744), ('functions', 1.228881529089087), ('identify', 1.2062163607930334)]
Topic 4:
[('python', 30.42639806804373), ('programming', 29.758037612446874), ('language', 16.153732934381424), ('excel', 12.002643597186008),
('sql', 8.76371047493149), ('knowledge', 7.5250479586657), ('tools', 7.3268608948141), ('good', 6.937336587789713), ('languages', 5.7988112
('tableau', 5.446813579702937), ('ms', 5.317119026435173), ('tool', 5.301444702663723), ('expertise', 4.856199687138079),
('certain', 4.682447545567969), ('suggest', 4.498550935769889), ('pretty', 4.459183813939342), ('hadoop', 4.190786388307931),
('think', 4.169600309590668), ('box', 4.021688106123531), ('spark', 3.9701935205619785)]
Topic 5:
[('thinking', 2.871192614510327), ('critical', 2.8329684493697043), ('problem', 1.3033802708640916), ('skill', 1.2708704624070952),
('questions', 1.2692483671070984), ('analyze', 1.2429214455593351), ('look', 1.224890338667755), ('problems', 1.195961745298627),
('results', 1.1877471240991428), ('constantly', 1.1612756580673365), ('resources', 1.156479573280077), ('trends', 1.113609563783567),
('verbal', 0.9488108607608124), ('technical', 0.937571506674553), ('action', 0.9272061517264588), ('order', 0.9251369711215849),
('able', 0.923194583666617), ('recognize', 0.9205707577862438), ('non', 0.9187512358969869), ('insights', 0.9131413497907596)]
```

Рис. 26. Розподіл термінів по темам

На професіограмі, що зображена на рис. 27, можна побачити:

- весь набір навичок;
- максимальну кількість слів, які відповідають темам;
- при наведені на певний ключовий термін видно скільки слів йому відповідають.

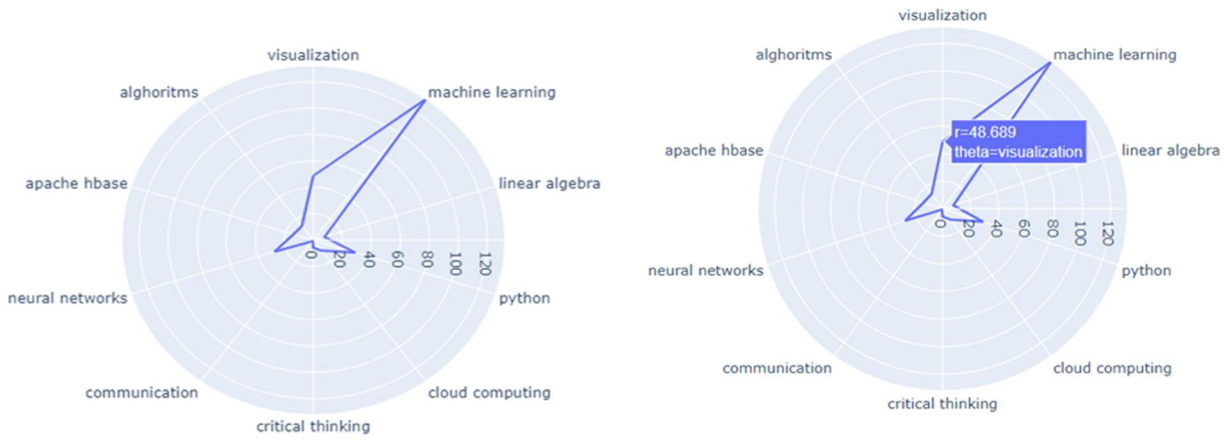


Рис. 27. Робота з професіограмою

Другим файлом є html-файл з розподілом тем на діаграмі. Загальний вигляд файлу зображено на рис. 28.

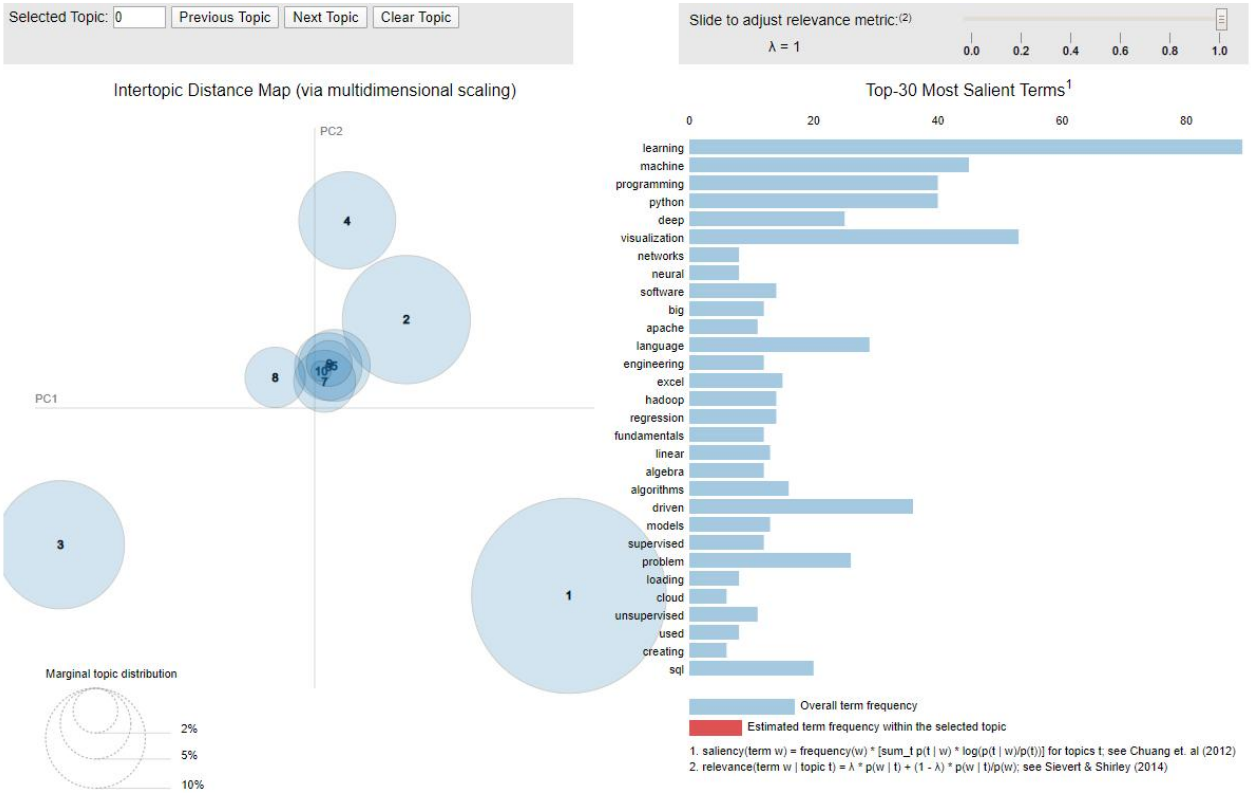


Рис. 28. Діаграма розподілу тем

При роботі з цим файлом у користувача є такі можливості:

- подивитись загальний розподіл термінів по темам;
- навести на термін та подивитись в яких темах він зустрічається;

Третім файлом є інформація зі словами, які найчастіше зустрічаються в документі. В цьому файлі можна подивитись на терміни (по горизонталі) та скільки разів ці терміни зустрічаються в документі (по вертикалі). Їхню кількість можна змінювати в залежності від параметрів (рис. 29).

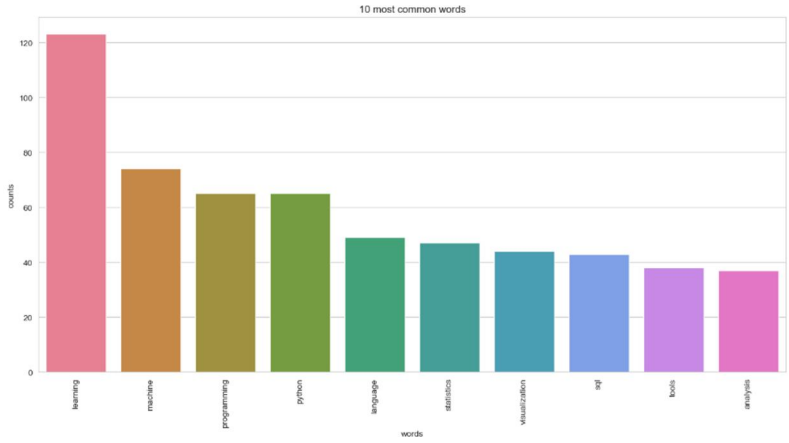


Рис. 29. Діаграма кількості входжень терміну в документі

Було розроблено веб-додаток для формування та аналізу професіограми спеціалістів зі штучного інтелекту. Додаток містить систему тестування, що дозволяє оцінювати рівень знань і навичок користувачів, а також підтримує категоризацію тестів та авторизацію. Після проходження тестування користувачу надається оцінка у відсотках, що дозволяє визначити відповідність його компетенцій вимогам професії. Всі особисті дані користувачів надійно зашифровані.

Таким чином, було створено платформу для оцінки професійних компетенцій фахівців зі штучного інтелекту. У майбутньому можливий розвиток системи з розширенням категорій тестувань для інших професій і напрямів, що дозволить більш комплексно оцінювати професійні навички спеціалістів у різних галузях.

Висновки

З позиції сучасного менеджменту, дослідження та впровадження автоматизованої системи побудови професіограми фахівця зі штучного інтелекту є вагомим кроком до формування стратегічно орієнтованої моделі управління персоналом. У контексті зростаючої конкуренції та цифровізації бізнес-процесів, саме точність, адаптивність і об’єктивність у прийнятті кадрових рішень стають визначальними чинниками успішності компаній.

Запропонована система дозволяє не лише стандартизувати вимоги до IT-посад, але й побудувати персоналізовані траєкторії розвитку працівників, ефективно використовувати HR-аналітику для прогнозування потреб у компетенціях, а також зменшити ризики плинності персоналу. Особливої цінності набуває можливість інтеграції отриманих результатів у системи корпоративного навчання, що сприяє формуванню єдиного освітнього простору всередині компанії.

Таким чином, отримані в роботі результати створюють основу для прийняття зважених управлінських рішень, підтримують реалізацію принципів data-driven HR і посилюють стратегічну позицію організації на ринку праці. Упровадження таких підходів є не лише сучасним трендом, а й практичною необхідністю для компаній, орієнтованих на сталий розвиток та інновації.

У ході дослідження було розглянуто процес створення професіограми для фахівців зі штучного інтелекту, що є актуальним завданням у сучасних умовах стрімкого розвитку технологій. Проведений аналіз літературних джерел, наукових досліджень і ринку праці дозволив сформувати чітке уявлення про вимоги до спеціалістів у цій сфері, а також визначити основні компетенції, необхідні для успішної професійної діяльності.

У роботі досліджено основні підходи до побудови професіограм, а також особливості професії спеціаліста зі штучного інтелекту. Було проаналізовано сучасний ринок праці, наукові публікації та тенденції розвитку галузі, що дозволило визначити ключові технічні та аналітичні навички, важливі для фахівців.

Розроблено програму побудови професіограми, що дозволяє систематизувати та структурувати інформацію про професію. На основі отриманих даних створено веб-додаток, який надає можливість тестування користувачів для оцінки їхніх компетенцій. Додаток підтримує категоризацію тестів, авторизацію користувачів та безпечне збереження особистих даних, а також формує результати у відсотковому вираженні.

Таким чином, у роботі досягнуто поставленої мети – сформовано професіограму фахівця зі штучного інтелекту, що містить опис основних обов'язків, компетенцій та умов праці. Це сприятиме покращенню процесу відбору спеціалістів, оптимізації освітніх програм і розвитку галузі в цілому.

Перспективи подальших досліджень включають розширення функціоналу веб-додатку, додавання нових категорій тестувань для інших професій, а також вдосконалення моделі оцінювання професійних компетенцій.

Перспективи подальших досліджень

Із точки зору сучасного менеджменту, розвиток інструментів для автоматизованого формування професіограм відкриває нові горизонти у сфері стратегічного управління людськими ресурсами. Професіограма, створена на основі реальних ринкових даних, стає не просто інформаційним ресурсом, а повноцінним аналітичним інструментом для прийняття зважених кадрових рішень.

У майбутньому однією з ключових тенденцій стане інтеграція професіограм у корпоративні HR-системи та платформи управління талантами (talent management systems). Це дозволить не лише стандартизувати вимоги до посад, але й реалізовувати моделі внутрішньої мобільності персоналу, а також формувати індивідуальні програми навчання та розвитку працівників.

Крім того, з позиції менеджменту важливою перспективою є перехід до data-driven HR – коли всі рішення щодо підбору, навчання та розвитку персоналу базуються на глибокій аналітиці. У цьому контексті автоматизовані професіограми можуть стати основою для прогнозування кадрових потреб, оцінки ризиків плинності персоналу, виявлення дефіциту компетенцій у команді тощо.

Іншим важливим вектором є формування адаптивних професіограм, які змінюються залежно від стратегічних цілей компанії, зовнішніх викликів ринку та динаміки розвитку технологій. Такі рішення сприятимуть підвищенню гнучкості організаційної структури та кращому узгодженню людського капіталу з бізнес-стратегією.

Загалом, розвиток цієї теми забезпечує управлінцям потужний інструмент для оптимізації витрат на підбір кадрів, зниження ризиків неправильного найму, підвищення продуктивності працівників та зміцнення конкурентоспроможності організації.

Список літератури

1. Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis.* Sage Publications.
2. Chan, D., & Schmitt, N. (2004). An agenda for future research on personnel selection. *Journal of Applied Psychology, 89*(4), 627–643. https://doi.org/10.1037/0021-9010.89.4.627
3. Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.
4. Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*(1), 100–112. https://doi.org/10.1037/0021-9010.88.1.100
5. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323. https://doi.org/10.1145/331499.331504
6. MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.
7. Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262
8. Thibodeaux, T., & Moore, S. (2015). The rise of skill testing in the hiring process. *Human Resource Management International Digest, 23*(4), 22–25. https://doi.org/10.1108/HRMID-04-2015-0064

References

1. Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Sage Publications.
2. Chan, D., & Schmitt, N. (2004). An agenda for future research on personnel selection. *Journal of Applied Psychology, 89*(4), 627–643. https://doi.org/10.1037/0021-9010.89.4.627
3. Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.
4. Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*(1), 100–112. https://doi.org/10.1037/0021-9010.88.1.100
5. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323. https://doi.org/10.1145/331499.331504
6. MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.
7. Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262
8. Thibodeaux, T., & Moore, S. (2015). The rise of skill testing in the hiring process. *Human Resource Management International Digest, 23*(4), 22–25. https://doi.org/10.1108/HRMID-04-2015-0064

B. M. Boiko, I. S. Protsyk
Lviv Polytechnic National University
Department of Management and International Entrepreneurship

AUTOMATION OF FORMATION OF A SPECIALIST'S PROFESSIONAL CURRICULUM BASED ON ANALYSIS OF VACANCIES AND TEXT DATA

The article explores the development and implementation of an automated system for constructing professionograms of artificial intelligence (AI) specialists using modern information technologies. Against the backdrop of an ever-evolving digital economy and rising demand for qualified AI professionals, the study highlights the need for accurate and scalable tools to identify and assess relevant competencies.

The main objective of the study is to design an innovative digital platform that enables automated generation of professionograms by analyzing unstructured textual data—such as job descriptions, expert publications, and scientific articles—through machine learning algorithms. Special emphasis is placed on the use of TF-IDF (Term Frequency-Inverse Document Frequency) and LDA (Latent Dirichlet Allocation), which allow for effective identification and classification of key terms, skills, and knowledge areas relevant to the AI domain.

The methodology involves a comprehensive review of the labor market, an analysis of expert opinions, and the processing of large-scale textual data to develop a structured model of professional competencies. The system is designed to function as a decision-support tool in HR analytics, enabling data-driven recruitment, optimization of corporate training programs, and personalized career planning.

One of the core advantages of the proposed solution is the reduction of subjectivity in the recruitment and evaluation process, as well as its adaptability to various specializations within the AI field. The system architecture includes a web-based interface, an interactive testing module, and visualization tools such as profession graphs and thematic distribution diagrams. This allows stakeholders—from HR managers to educators and researchers—to intuitively explore competency models and apply them in real-world scenarios.

Furthermore, the article discusses the broader implications of using data-driven approaches in strategic human capital management. It argues that automated professionogram systems can significantly improve workforce planning, reduce turnover rates, and align employee skill sets with evolving business needs. The research also opens pathways for extending the system to other professional domains, incorporating soft skills and psychological profiling, and developing mobile applications for continuous self-assessment and skill tracking.

In conclusion, the study presents a scalable and versatile framework that not only enhances the efficiency of personnel selection and development but also supports long-term organizational growth in the age of artificial intelligence.

Keywords: professionogram, information system, artificial intelligence, automation, HR analytics, competencies, TF-IDF, LDA, personnel management, digitalization.