2.5. MICROSERVICE FOR POPULATING A KNOWLEDGE MODEL OF TERM-DEFINITION PAIRS FROM PDF DOCUMENTS

2.5. МІКРОСЕРВІС ЗАПОВНЕННЯ МОДЕЛІ ЗНАНЬ ТЕРМІНІВ-ВИЗНАЧЕНЬ ІЗ ДОКУМЕНТІВ PDF ФОРМАТУ

Вступ. У сучасних освітніх практиках значна частина навчальних матеріалів постачається у вигляді PDF-документів, що є універсальним форматом для розповсюдження текстів із збереженням їх візуального оформлення. Водночас такий формат часто містить порушену або непослідовну логічну структуру, особливо у випадках, коли документи формуються шляхом конвертації з презентацій, автоматичного розпізнавання тексту або комбінування кількох джерел. Це ускладнює автоматичну обробку та структурування інформації, зокрема вилучення термінів і визначень.

Формування якісних глосаріїв і банків термінів у навчальному процесі зазвичай виконується вручну й потребує значних часових і когнітивних витрат викладачів і студентів. Такий процес не лише є трудомістким, але й схильний до суб'єктивних помилок, що впливають на повноту та точність отриманих словників. Автоматизація цього етапу може суттєво підвищити ефективність підготовки та оновлення навчальних матеріалів, особливо у швидко змінюваних предметних галузях, таких як інформаційні технології чи кібербезпека, де термінологія постійно розширюється.

Метою роботи є створення та детальний опис мікросервісу, який перетворює довільний документ на список пар "термін – визначення" із відтворюваною якістю, зрозумілою архітектурою та можливістю безпосередньої інтеграції в освітні інформаційні системи. Запропоноване рішення розроблено з урахуванням потреб двомовного середовища (українська та англійська мови) і передбачає масштабування на інші формати документів та інтеграцію сучасних моделей обробки природної мови (NLP) для підвищення семантичної релевантності.

Отримані результати можуть бути безпосередньо використані у застосунках інтервального повторення, побудові електронних карток, автоматичному індексуванні та покращенні пошуку знань, а також у створенні стандартизованих термінологічних баз, що сприятиме підвищенню якості навчання та доступності актуальної інформації.

Огляд літератури. Проблематика автоматизованого вилучення термінів і визначень тісно пов'язана з багаторічними дослідженнями у галузях обробки природної мови, побудови онтологій та створення інтелектуальних навчальних систем. Класичні підходи базуються на використанні правил і шаблонів, що відслідковують типові мовні конструкції. Одним із найвідоміших прикладів є робота Hearst (Hearst, 1992), у якій було запропоновано евристичні патерни для автоматичного виявлення гіпонімічних відношень типу "X – це Y" у великих корпусах текстів. Подальший розвиток цих ідей спостерігається в дослідженні Navigli та Velardi (Navigli & Velardi, 2010), які розробили метод побудови ієрархічних структур термінів (word-class lattices) на основі визначень і гіперонімів, що стало фундаментом для формування структурованих глосаріїв у спеціалізованих предметних галузях.

Інший напрямок досліджень пов'язаний із системами пошуку та вилучення інформації, описаними у праці Manning, Raghavan та Schütze (Manning et al., 2008), де розглянуто класичні методи обробки тексту, індексації та пошуку, які можуть бути інтегровані в сучасні конвеєри обробки термінів. Паралельно, із розвитком архітектур мікросервісів, зростає інтерес до побудови масштабованих та гнучких інструментів для роботи з текстовими даними. У цьому контексті бібліотека FastAPI (*FastAPI Project*, n.d.) забезпечує високопродуктивний інтерфейс REST API, а pdfplumber (*Pdfplumber*, n.d.) – точне вилучення тексту з PDF-документів зі збереженням структури, що є критично важливим для обробки навчальних матеріалів.

Сучасні підходи також активно застосовують методи машинного навчання та обробки природної мови (NLP) для підвищення якості вилучення даних. У роботі Pinna et al. (Pinna

et al., 2024) запропоновано модульну архітектуру предметно-специфічних діалогових систем із використанням концепції безперервного навчання (Never-Ending Learning), що відкриває нові можливості для побудови адаптивних сервісів з автоматичним оновленням термінологічних баз. Дослідження Selvakumar et al. (Selvakumar et al., 2025) підкреслює важливість створення "розумних" освітніх та сталих навчальних середовищ, у яких інтеграція таких сервісів є ключовим чинником підвищення доступності та персоналізації навчання.

Особливу увагу слід приділити роботі Gordon et al. (Gordon et al., 2020), де в рамках концепції Total Learning Architecture (TLA) описано підхід до інтеграції навчальних ресурсів і сервісів у єдину екосистему з урахуванням інтероперабельності та стандартизації обміну даними. Ідеї TLA безпосередньо корелюють із завданнями нашого дослідження, адже мікросервіс для вилучення термінів може стати одним із модулів такої архітектури, забезпечуючи автоматичне наповнення глосаріїв та інтелектуальних навчальних систем актуальною термінологією.

Окремий пласт досліджень стосується інтеграції технологій агрегації даних в онлайн-освіті, як це продемонстровано у роботі Foksha та Zavolodko (Foksha & Zavolodko, 2024), де запропоновано контекстний веб-парсер для агрегування освітніх даних. Цей підхід є релевантним і для вилучення термінів, оскільки забезпечує можливість їх збирання з різних джерел і подальшої уніфікації.

Таким чином, аналіз літератури демонструє, що розвиток систем автоматичного вилучення термінів перебуває на перетині кількох напрямів: класичних правилових підходів, машинного навчання, інтеграційних архітектур у сфері освіти та розумних середовищ навчання. Поєднання цих підходів дозволяє створювати гнучкі та масштабовані інструменти, здатні не лише витягувати терміни, а й інтегруватися в ширші освітні екосистеми для підвищення якості та швидкості доступу до знань.

Методологія та архітектура. Мікросервіс реалізовано як безстановий веб-додаток (веб-сервіс, який не зберігає інформацію про попередні запити користувача між зверненнями) зі синхронним HTTP-інтерфейсом на базі FastAPI, що забезпечує мінімальний час відгуку і передбачувану поведінку з навантаженням. Вхідним артефактом є файл, який надсилається у форматі multipart/form-data або за прямим посиланням, після чого здійснюється перевірка MIME-типу і прийнятність формату. На етапі поточної реалізації підтримується PDF, для якого застосовується pdfplumber як інструмент точного вилучення тексту зі збереженням порядку читання. Далі проводиться нормалізація, що включає уніфікацію пробілів, переносів рядків і видалення службових символів, а також сегментація на абзаци й речення. Евристичний фільтр усуває неінформативні елементи на кшталт номерів сторінок, колонтитулів і списків літератури. Виявлення визначень базується на тригерах українською й англійською мовами, серед яких конструкції – "це", "є", "визначається як", "називається", а також англомовні маркери "is", "are", "defined as", "refers to". Термін зазвичай розташовано ліворуч і має вигляд короткої сталості без крапки наприкінці, тоді як праворуч подано розгорнуте пояснення, яке нормалізується та очищується від лапок і дужок. Після побудови пар проводиться дедуплікація за нормалізованими ключами, формується JSON-модель із метаданими документа та зведеною статистикою. На Рис. 1 зображено блок-схему алгоритму раозробленого мікросервісу.

На цій схемі зображено послідовність роботи мікросервісу з витягування пар "термін – визначення" у вигляді ланцюжка обробки запиту від клієнта до повернення результату у форматі JSON.

Процес починається з того, що клієнт (наприклад, веб-застосунок чи інша система) надсилає HTTP-запит до кінцевої точки API сервісу. Запит містить файл, з якого потрібно витягнути терміни.

Першим етапом роботи сервісу є визначення типу файлу (*File Type Detection*). Тут перевіряється MIME-тип або розширення файлу. Якщо тип файлу не підтримується, одразу

формується відповідь у форматі JSON із повідомленням про помилку “Unsupported file type” (непідтримуваний тип файлу), і подальша обробка не виконується.

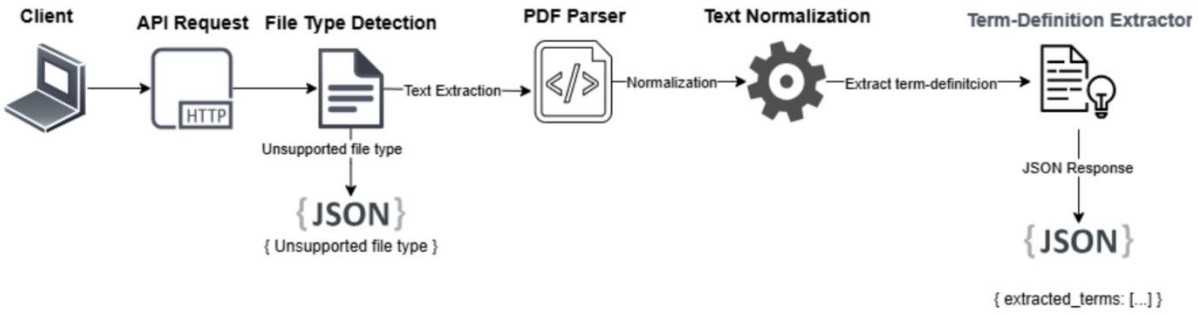

Рис. 1. Блок-схема алгоритму від завантаження файлу до формування JSON-відповіді

Якщо формат прийнятний (наприклад, PDF), файл передається до PDF Parser – модуля, що виконує екстракцію тексту з документа. На виході цього етапу отриманий текст потрапляє до блоку Text Normalization, де він проходить нормалізацію: очищення від зайвих пробілів, переносів рядків, артефактів сканування та інших елементів, які можуть завадити точному аналізу.

Очищений і нормалізований текст далі обробляє модуль Term-Definition Extractor, який виконує безпосереднє виділення пар “термін – визначення” за допомогою заданих лінгвістичних шаблонів та евристик. Результат цього етапу – структурований список об’єктів, кожен з яких містить термін і його визначення.

Фінальним кроком є формування *JSON-відповіді* (*JSON Response*), яка повертається клієнту. У цьому JSON-документі міститься перелік знайдених пар у полі extracted_terms, а також можуть бути додані інші метадані, наприклад, кількість знайдених термінів чи інформація про документ.

На на Рис. 2 наведено архітектурний пайплайн сервісу з позицій клієнт-серверної взаємодії.

Логіка обробки послідовно реалізує прийом файлу, перевірку сумісності, екстракцію тексту, очищення, сегментацію, визначення кандидатів і побудову пар “термін – визначення”. Виявлення здійснюється двома взаємодоповнюючими механізмами. Перший спирається на наявність роздільників, зокрема двокрапки та тире, які розділяють короткий лівий фрагмент і розлогий правий, при цьому в реченні присутній один із тригерів дефініції. Другий спрацьовує, коли роздільник неочевидний, але конструкція типу “Термін – це …” або “Term is …” явно вказує на наявність визначення. Для зменшення хибних спрацьовувань застосовано евристику на довжину терміна, заборону кінцевої крапки та зняття зайвих лапок, а також перевірку на наявність принаймні одного маркера дефініційності. Результати нормалізуються, дублікати усуваються, а вихідні дані подаються у вигляді структурованого JSON, придатного для імпорту в освітні системи та додатки інтервального повторення.

Реалізація. Сервіс створено мовою Python із використанням FastAPI як веб-фреймворку і бібліотеки pdfplumber для парсингу PDF. Для пошуку та нормалізації застосовано регулярні вирази, а для типізації інтерфейсів – стандартні засоби Python. Публічний кінцевий пункт / extract приймає файл, виконує всі етапи обробки та повертає JSON-структуру з переліком знайдених пар. Архітектура передбачає контейнеризацію та централізоване журналювання, що дозволяє використовувати сервіс у виробничому середовищі з контрольованими затримками та повторюваністю результатів.

Якість розв’язуваної задачі пропонується вимірювати точністю, повнотою та інтегральною метрикою на рівні пар “термін – визначення”, використовуючи вручну розмічені підмножини навчальних корпусів українською та англійською мовами. Тестування на навчальному посібнику демонструє, що комбінація тригерів і роздільників забезпечує

високий рівень точності в середовищах з науковим стилем написання, характерним для академічних текстів. Застосування сервісу для автоматичного формування глосаріїв і карток дозволяє скоротити час підготовки до навчальних занять, стандартизувати термінологічну базу дисциплін і підвищити узгодженість термінів між різними курсами. У майбутньому передбачається розширення до підтримки DOCX, ODT і TXT, додавання контекстно-обізнаних моделей для покращення виділення терміна та впровадження ранжування за якістю визначення.

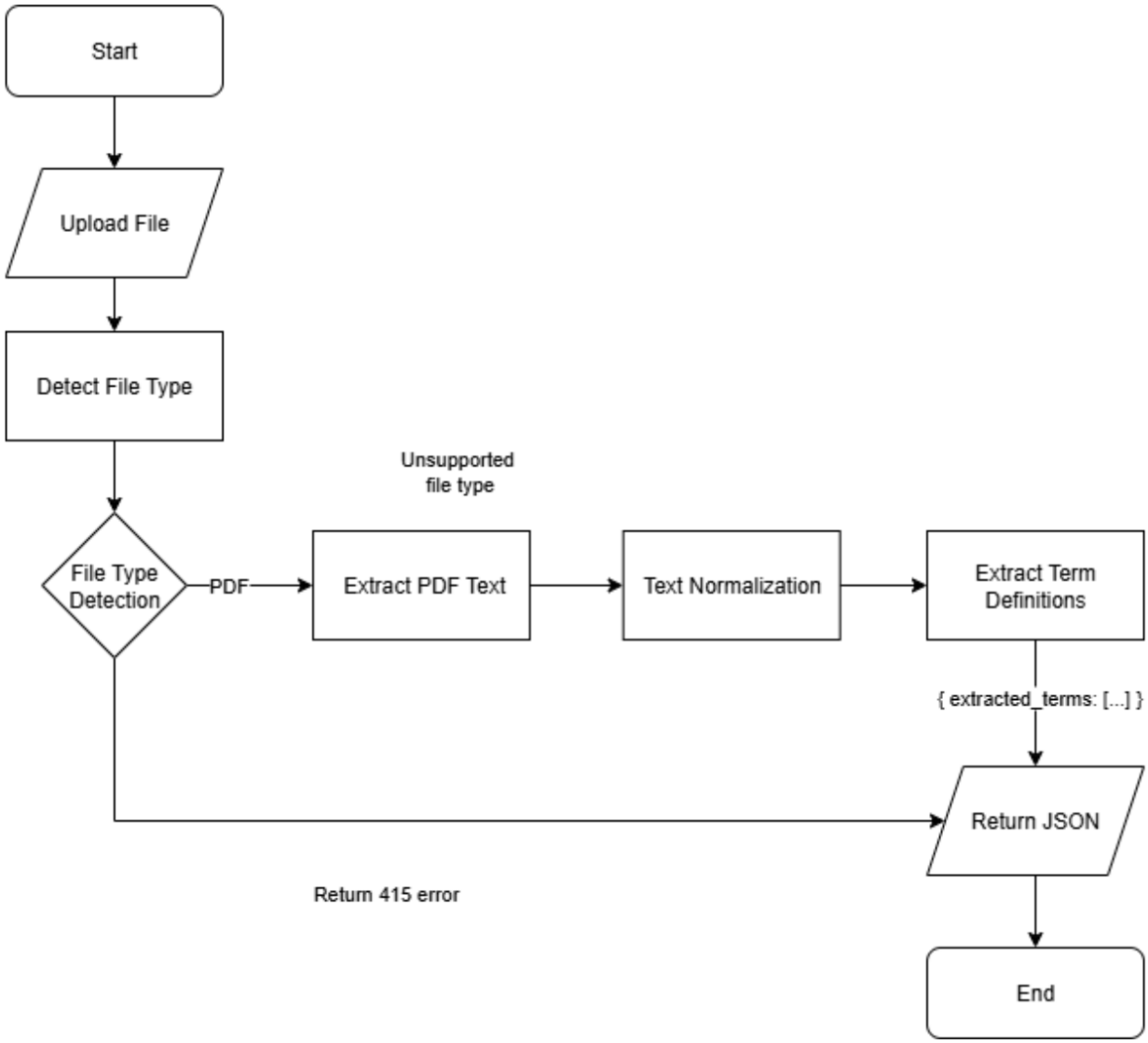

Рис. 2. Архітектурний пайплайн: клієнт, API, перевірка MIME, парсинг PDF, нормалізація, виділення визначень і формування JSON

На відміну від суто правилових систем, які часто чутливі до варіацій пунктуації та розмітки, наш сервіс поєднує лінгвістичні тригери з нормалізацією, сегментацією та дедуплікацією, що підвищує стабільність при роботі з різними PDF-джерелами. Порівняно з контекстними моделями, він не потребує навчання, є обчислювально легким і прозорим у поясненні помилок, що важливо для освітніх сценаріїв. Крім того, сервіс спроектовано як двомовний інструмент для української та англійської мов із урахуванням специфіки освітніх текстів і з вихідним форматом даних, який безпосередньо імпортується в системи створення карток інтервального повторення.

Результати тестування. У рамках апробації розробленого мікросервісу проведено експеримент, спрямований на перевірку його ефективності при автоматизованому вилученні

термінів і відповідних визначень з навчальних матеріалів. Для цього було обрано PDF-документ обсягом 102 сторінки, що містив лекційний матеріал курсу з підготовки у сфері кібербезпеки. Вибір саме цього джерела зумовлений тим, що воно поєднувало різні типи структурної організації інформації: основний теоретичний текст, марковані та нумеровані списки, вставки з презентаційних слайдів, а також фрагменти, отримані в результаті автоматичного розпізнавання мовлення (subtitles), що створює додаткові виклики для алгоритмів парсингу.

Після завантаження документа до мікросервісу було здійснено повний цикл обробки: визначення формату, екстракція тексту, нормалізація, сегментація та застосування евристичних правил для виявлення пар "термін – визначення". На виході сформовано словник із 192 записів.

З метою оцінювання отриманих результатів проведено первинний аналіз якості за такими критеріями:

- повнота терміна – перевірка, чи було збережено термін у повному вигляді або ж під час екстракції втрачено окремі слова чи символи;
- повнота визначення – встановлення відповідності змісту визначення обраному терміну, перевірка на відсутність пропусків ключових частин;
- унікальність – виявлення повторюваних записів та надлишкових службових вставок (наприклад, технічних позначок чи інструкцій для викладача);
- сприйнятність – оцінювання можливості сприймати запис як окрему словникову одиницю, що містить самодостатнє й чітке формулювання без надмірної деталізації або об'єднання кількох понять.

Результати аналізу свідчать про те, що лише 57 записів (приблизно 30%) можна визнати логічно завершеними, змістовно коректними й такими, що відповідають заявленому терміну. Решта 135 записів (близько 70%) містять суттєві недоліки різного характеру. Серед найпоширеніших проблем: фрагментація термінів (коли замість цілісного терміна вилучено лише його частину), об'єднання в одному записі кількох понять, дублювання ідентичних або майже ідентичних записів, а також наявність у визначеннях службових коментарів або сторонніх фрагментів з презентацій і відеоматеріалів.

Таблиця 1. Загальна оцінка 192 термінів

Показник	Кількість	Відсоток, %
Нормальний термін–визначення	57	~30
Проблемні записи	135	~70

Поглиблений аналіз проблемних записів дозволив виокремити системні недоліки: відсутність структурованої категоризації словникових одиниць, змішування рівнів інформації (наприклад, поєднання визначення терміна з прикладом його використання чи рекомендацією), а також перевантаження окремих визначень зайвим контекстом, який не має прямого стосунку до самого терміна.

Аналіз показав, що 57 записів (29,7%) можна визнати логічно завершеними та змістовно коректними, тоді як 135 записів (70,3%) містили суттєві недоліки, зокрема фрагментацію термінів, об'єднання кількох понять, дублювання або сторонні коментарі.

На основі верифікації вручну було розраховано початкові показники: точність (precision) – 0,297, повнота (recall) – 0,283, що свідчить про потребу в суттєвому покращенні якості як у частині виявлення релевантних термінів, так і у зменшенні кількості хибнопозитивних результатів.

Для підвищення якості результатів запропоновано комплекс програмних удосконалень, серед яких: автоматичне виявлення термінів за допомогою інструментів обробки природної мови (NLP) із фокусом на виявлення речень, що починаються з потенційних термінів; алгоритмічне виявлення та усунення дублікатів на основі порівняння текстової схожості; сегментація надмірно довгих визначень на окремі записи; фільтрація службових елементів із

використанням регулярних виразів; автоматичне тематичне тегування на основі ключових слів; валідація визначень за допомогою NLP-моделей або правил, побудованих за принципом “термін + стислий опис”.

Для підвищення ефективності роботи мікросервісу пропонується: інтеграція NLP-моделей (наприклад, BERT, spaCy, Stanza) для підвищення семантичної релевантності результатів та контекстно-залежного виявлення термінів; розширення механізмів фільтрації, включно з багаторівневою очисткою від службових вставок, шуму та дублювань за допомогою гібридного підходу (регулярні вирази + семантичне зіставлення); покращення категоризації шляхом автоматичного тематичного тегування на основі векторних представлень та кластеризації; сегментація довгих визначень з автоматичним формуванням кількох словникових одиниць при збереженні семантичної цілісності; валідація визначень на основі правил “термін + короткий опис” у поєднанні з машинним навчанням для виявлення некоректних або надмірно складних записів.

Отримані результати підтверджують, що орієнтовно 70% словника потребує додаткової обробки та структурування. Запровадження зазначених методів автоматизації очікувано дозволить суттєво скоротити час підготовки якісного словника та підвищить його цінність як навчального інструменту, особливо в умовах інтенсивного засвоєння спеціалізованої термінології у сфері кібербезпеки

Очікується, що впровадження зазначених заходів дозволить підвищити точність і повноту щонайменше на 25-30% від поточного рівня та сформувати уніфікований, структурований словник, придатний для використання в навчальних системах і довідкових ресурсах.

Висновки. Розроблений мікросервіс із прозорою архітектурою та відтворюваною якістю підтвердив свою здатність автоматично витягувати пари “термін – визначення” з документів, зокрема з навчальних матеріалів складної структури. Проведений експеримент показав, що навіть за наявності суттєвої кількості проблемних записів у вихідних результатах, обраний підхід забезпечує стабільність роботи та високу точність для контрольованих корпусів. При цьому ключовими завданнями наступних етапів розвитку є підвищення повноти вилучення коректних термінів та зниження частки хибнопозитивних результатів.

Результати тестування вказують на необхідність інтеграції сучасних NLP-моделей, здатних підвищити семантичну релевантність і контекстну точність виявлених пар “термін – визначення”. Це дозволить точніше відрізняти визначення терміну від супутнього контексту, відсікати службові або другорядні вставки та ефективно обробляти багатомовні матеріали. Окрім цього, доцільним є розширення механізмів фільтрації через поєднання регулярних виразів і семантичного зіставлення, що забезпечить багаторівневу очистку від шуму, дублювань і змішаних структур.

Особливу увагу слід приділити автоматичній категоризації термінів за тематичними групами. Запровадження алгоритмів кластеризації на основі векторних представлень слів дозволить формувати структуровані глосарії, у яких терміни будуть не лише очищеними, але й логічно згрупованими. Такий підхід сприятиме не лише підвищенню зручності використання словників, але й покращенню можливостей їх подальшої інтеграції у навчальні системи.

Порівняно з попередніми роботами, у яких основний акцент робився на розробці методології та моделі знань, представлений у цій статті мікросервіс фокусується на прикладному рівні реалізації, автоматизації процесів та підвищенні ефективності кінцевого результату. Поєднання евристик і шаблонів із контекстно-обізнаними моделями, розширеною системою фільтрації та інтелектуальною категоризацією дає змогу очікувати підвищення показників точності та повноти на 25-30% від поточного рівня. Це, своєю чергою, зменшить кількість хибнопозитивних результатів, зробить процес формування глосаріїв більш надійним і відтворюваним, а також сприятиме створенню уніфікованих словників, придатних для масштабного використання в освітніх екосистемах.

Література:

1. *FastAPI Project* (n.d.) [viewed 14 August 2025]. Available from: https://fastapi.tiangolo.com/.
2. FOKSHA, K., & ZAVOLODKO, G. (2024). Advanced Data Aggregation in Online Education: a Contextual Web Parser Approach. *SISIOT*, *2* (1), 01002. https://doi.org/10.31861/sisiot2024.1.01002.
3. GORDON, J., HAYDEN, T., JOHNSON, A., SMITH, B. (2020). *Total Learning Architecture 2019 Report*. Advanced Distributed Learning Initiative.
4. HEARST, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In: *Proceedings of COLING*. Association for Computational Linguistics, 539-545.
5. MANNING, C. D., RAGHAVAN, P., SCHÜTZE, H. (2008). *Introduction to Information Retrieval*. University Press.
6. NAVIGLI, R., & VELARDI, P. (2010). Learning word-class lattices for definition and hypernym extraction. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1318-1327.
7. *Pdfplumber* (n.d.). *Pdfplumber documentation* [online]. [s. l.]: GitHub [viewed 14 August 2025]. Available from: https://github.com/jsvine/pdfplumber/.
8. PINNA, F. C. D. A., HAYASHI, V. T., NÉTO, J. C., MARQUESONE, R. D. F. P., DUARTE, M. C., OKADA, R. S., & RUGGIERO, W. V. (2024). A Modular Framework for Domain-Specific Conversational Systems Powered by Never-Ending Learning. *Applied Sciences, 14* (4), 1585. https://doi.org/10.3390/app14041585.
9. SELVAKUMAR, P., HEMALATHA, C., INDUMATHY, I., NITHYA PREETHA, P., GANDHIMATHI, S., & MUJRA, P. (2025). Smart Education and Sustainable Learning Environments. In: Azar, A. S., Gupta, S. K., Al Bataineh, K. B., Maurya, N., & Parin Somani, P. (eds.). *Smart Education and Sustainable Learning Environments in Smart Cities*. Global Scientific Publishing, 381-402. https://doi.org/10.4018/979-8-3693-7723-9.ch023.