УДК 004.85:004.91 DOI: https://doi.org/10.30837/IYF.IIS.2024.030

ЗАСТОСУВАННЯ ТЕХНОЛОГІЇ ML.NET ЗАДЛЯ АВТОМАТИЧНОГО СОРТУВАННЯ ТА ТЕГУВАННЯ ТЕКСТОВИХ ДОКУМЕНТІВ ДЛЯ СЕМАНТИЧНОЇ ЦИФРОВОЇ БІБЛІОТЕКИ.

Заворіна М.А.
Науковий керівник – ас. каф. ШІ Поліїт А. Г.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна
e-mail: mariia.zavorina@nure.ua

The paper proposes to develop a tool that will allow creating a semantic network of files in addition to the standard hierarchical Windows file system, which could automatically sort text files for a semantic digital library by categories based on their content. The preparation of a corpus of texts for the purpose of automatically obtaining categories, tags and attributes to fill a semantic digital library is an extremely important and relevant technology. In order to implement such an add-on to the Semantic File System, it is proposed to apply ML.NET technology, this will allow easy integration with other ML.NET components or external libraries, expanding the possibilities of sorting text documents.

У сучасному світі, коли інформаційний потік невпинно зростає, використання штучного інтелекту для сортування текстових документів стає надзвичайно важливою та актуальною технологією. Саме виходячи з цього запропоновано створити інструмент, який дозволить створити семантичну мережу файлів додатково до стандартної ієрархічної файлової системи Windows, що могла б автоматично сортувати текстові файли домашньої бібліотеки за категоріями на основі їх контенту.

Semantic File System (SFS) – це спеціальний тип файлової системи, який додає семантичні можливості до стандартної файлової системи. Він дозволяє користувачам асоціювати з файлами додаткові метадані та встановлювати між файлами семантичні зв'язки, що збагачує інформацію про файли та дозволяє більш гнучко та ефективно керувати ними.

Ключові риси Semantic File System (SFS): додавання метаданих до файлів; семантичні зв'язки між файлами; гнучка організація файлів; семантичний пошук та фільтрація; інтеграція з операційною системою.

Приклади деяких проектів Semantic File Systems включають Semantic File System for Linux (SFS-Linux), Semantic File System (SFS) та деякі інші. Кожен з них має свої особливості та реалізацію, але всі вони прагнуть збагатити стандартні файлові системи семантичними можливостями для кращого керування файлами та інформацією. деякі інструменти та програми надають семантичні можливості для керування файлами в операційній системі Windows: TMSU, TagSpaces, Tabbles, WDS (Windows Desktop Search), Semantic File Organizer

У більшості із згаданих програм потрібно вручну встановлювати асоціації та тегувати файли, хоча деякі з них також пропонують автоматичне тегування та класифікацію файлів, існує ряд обмежень, що можуть зробити їх недоцільними або недостатньо ефективними для конкретних потреб. По-перше, деякі з цих програм можуть бути занадто дорогими для бюджетів певних користувачів. Крім того, не всі програми можуть ефективно вирішувати конкретні вимоги або специфічні потреби користувача, що призводить до необхідності пошуку альтернативних рішень.

Зазвичай домашню цифрову бібліотеку можна уявити у вигляді дерева. Хоча деревоподібна файлова система зручна для операційних систем, вона не завжди відповідає потребам користувачів через обмеженість у встановленні зв'язків між папками та файлами. Для подолання цієї проблеми варто перейти до графової структури, що дозволить кожному елементу мати кілька батьків, забезпечуючи гнучке управління. Також важливо впроваджувати віртуальні папки та віртуальні каталоги, які автоматично формують вміст згідно з заданими критеріями, що спрощує організацію та пошук файлів. Імена віртуальних каталогів інтерпретуються як запити. Результатом запиту є набір файлів і/або каталогів, які містять описані сутності – текстові документи.

Запити – це логічні комбінації атрибутів, де кожен атрибут описує бажане значення поля. Застосування файлової системи на основі онтологій відкриває шлях до злиття семантичної мережі та семантичного робочого столу (Semantic Desktop), створюючи єдину мережу, яка полегшує взаємодію та доступ до інформації.

Семантичні файлові системи створюють ефективну абстракцію для зберігання інформації, надаючи гнучкий асоціативний доступ до вмісту. Це дозволяє краще відтворювати зв'язки між різними елементами та надає користувачам зручний інтерфейс для пошуку та використання інформації. Додатково, семантичні файлові системи розширюють можливості традиційних файлових систем, дозволяючи автоматично вилучати та індексувати атрибути файлів [1].

Система асоціативного доступу в семантичній файловій системі базується на концепції сутностей та запитів. Сутність може представляти собою цілий файл, окремий об'єкт у файлі або навіть каталог. Запити визначаються як логічні комбінації атрибутів, що описують бажані характеристики сутностей. Ці запити дозволяють користувачам отримувати набір файлів або каталогів, які задовольняють вказані критерії. За допомогою кон'юнктивних, диз'юнктивних та заперечувальних запитів можна точно визначити потрібні сутності та виключити з результатів непотрібні об'єкти. Система стає узгодженою із запитами, коли гарантує результати запиту, що відповідають її поточному вмісту. Використання семантичних метаданих дозволяє уточнювати запити та точно визначати

потрібні атрибути сутностей. Додатково, можливість призначення атрибутів об'єктам користувачем розширює можливості системи, дозволяючи персоналізувати та уточнювати критерії пошуку [2].

Стає питання, як автоматично розкласифікувати, згенерувати імена тегів та атрибутів для текстових файлів. Тобто, пропонується попередня підготовка корпусу текстів з метою автоматичного отримання категорій, тегів та атрибутів для наповнення семантичної цифрової бібліотеки. Задля реалізації такої надбудови до SFS пропонується застосувати технологію ML.NET. ML.NET — це платформа машинного навчання, розроблена Microsoft [3]. Однією з ключових переваг ML.NET для сортування текстових документів є його доступність і простота інтеграції в .NET-додатки. Завдяки відкритому коду та крос-платформенній сумісності ML.NET спрощує процес впровадження моделей машинного навчання в існуючі робочі процеси, незалежно від операційної системи. Крім того, розширювана архітектура ML.NET дозволяє легко інтегруватися з іншими компонентами ML.NET або зовнішніми бібліотеками, розширюючи можливості сортування текстових документів [4].

Розробка системи автоматичної класифікації та тегування текстових файлів для семантичної цифрової бібліотеки, яка допомагатиме організовувати та швидко знаходити документи за їх змістом, є одним із напрямів розвитку цієї технології. Інтеграція технології машинного навчання ML.NET може значно полегшити реалізацію подібних проектів, забезпечуючи доступність та простоту використання в різних програмних середовищах.

Список використаних джерел

1. Гіффорд Д.К., Жувело П., Шелдон М., О’Тул Д. Semantic file systems. ACM Operating Systems Review. 1991.

2. Артюшина Л.А. Методы представления информации в простых семантических сетях. Научно-технический вестник информационных технологий, механики и оптики. 2020.

3. ML.NET: Machine Learning for .NET Developers. : вебсайт URL: https://www.codemag.com/Article/1911042/ML.NET-Machine-Learning-for-.NET-Developers (дата звернення: 26.02.2024).

4. What is ML.NET? An open-source machine learning framework. : вебсайт URL: https://dotnet.microsoft.com/en-us/learn/ml-dotnet/what-is-mldotnet (дата звернення: 26.02.2024).

5. Tanvir, Q. Multi Page Document Classification using Machine Learning and NLP. Medium. 2021.