

УДК 004

*Груницький Д. С., магістрант,
Чижмотря О. В., ст. викладач,
Дмитренко І. А., ст. викладач*

Державний університет «Житомирська політехніка»

ВИДІЛЕННЯ І СТРУКТУРИЗАЦІЯ КЛЮЧОВИХ ПОНЯТЬ У ТЕКСТОВИХ НАВЧАЛЬНИХ МАТЕРІАЛАХ

У сучасному навчальному середовищі люди потребують нові способи обробки великої кількості тексту [1] для кращого розуміння, запам'ятовування та загалом, навчання. Як всім нам відомо, студенти навчальних закладів, користувачі на різних платформах часто працюють із книгами, статтями або презентаціями, де важливо виділити ключові поняття даного матеріалу. Автоматичне виділення та структуризація ключових слів, термінів і основних ідей допомагає покращити навчальний процес та підвищити його ефективність.

Основна ідея полягає у створенні інструменту, який автоматично аналізуватиме текстовий матеріал різних форматів, таких як книги, статті, документи або презентації. Під час аналізу система використовуватиме спеціальні алгоритми для визначення ключових понять, термінів та інших важливих фраз, які є основою змісту тексту. Виділена інформація не лише буде зібрана, але й упорядкована у структурований формат, що зробить її більш доступною та зрозумілою для будь-якого користувача. Це дозволить значно скоротити час на обробку великих обсягів текстової інформації та спростить ознайомлення з матеріалом.

Для досягнення цієї мети використовуватимуться сучасні методи обробки текстових даних. Зокрема, TF-IDF (Term Frequency-Inverse Document Frequency) [2] допомагає знаходити важливі слова у тексті, відкидаючи непотрібні і виділяючи ті, які краще підходять для розуміння змісту, використовуючи наступні формули:

$$TF(t, d) = \frac{\text{number of times } t \text{ appears in } d}{\text{total number of terms in } d} \quad (1)$$

$$IDF(t) = \log \frac{N}{1+df} \quad (2)$$

$$TF - IDF(t, d) = TF(t, d) * IDF(t), \quad (3)$$

де:

t – термін (слово або фраза, що аналізується);

d – окремий документ у колекції;

N – загальна кількість документів у колекції;

df – кількість документів, у яких зустрічається термін t .

Методи токенизації та лематизації допомагають розбити текст на окремі слова й привести слова до базової форми, що полегшує їх подальший аналіз. Додатково буде використовуватись аналіз змісту, щоб знайти основні ідеї та зв'язки в тексті. Ключові поняття будуть згруповані за темами, щоб інформація була чіткою та легкою для розуміння. Загалом, комплексне використання цих методів забезпечить ефективну обробку текстових даних і створення зручного для сприйняття результату.

Практична реалізація інструменту дозволить користувачам завантажувати текстові файли у різних форматах (DOC, DOCX, PDF тощо) або вставляти посилання на статті, книги та інші джерела. Система автоматично аналізує завантажений або введений матеріал, виділяє ключові слова, фрази та терміни, а також знаходить основні ідеї та визначення. Отримана інформація обробляється та структурується у зручному форматі, наприклад, у вигляді списку термінів, коротких підсумків або згрупованих термінів за темами. Це дозволить користувачам швидко знайти необхідну інформацію та отримати чітке уявлення про основний зміст тексту.

Розроблений інструмент дозволить ефективно обробляти великі об'єми тексту автоматично, створюючи структуровані конспекти, виділяти найважливіші терміни та поняття. Це значно спрощує процес засвоєння навчального матеріалу, оскільки користувачам стає легше знаходити необхідну інформацію та отримувати чітке уявлення про зміст документа. Завдяки такому інструменту можна економити час, уникати необхідності переглядати великі тексти та швидко зосередитися на ключових аспектах. Однак є і певні недоліки. Автоматичне виділення інформації може іноді бути неточним, особливо у складних текстах або при використанні специфічної термінології. Також інструмент може не завжди точно визначати терміни та фрази, що впливає на якість підсумованої інформації.

Отже, автоматичне виділення та упорядкування ключових понять допомагає краще обробляти текстові матеріали, роблячи навчання швидшим та зручнішим для будь-якого користувача.

Список використаних джерел:

1. Що таке НЛП? Як це працює, переваги, проблеми, приклади [Електронний ресурс] – Режим доступу до ресурсу: <https://uk.shaip.com/blog/what-is-nlp-how-it-works-benefits-challenges-examples/>.

2. Understanding TF-IDF for Machine Learning [Електронний ресурс] – Режим доступу до ресурсу: <https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/>.