УДК 004.4

Козенко О.В., Мазурець О.В., Молчанова М.О., Собко О.В.

Хмельницький національний університет

ВИКОРИСТАННЯ МЕТРИК КОСИНУСНОЇ СХОЖОСТІ ТА ІНДЕКСУ ЖАККАРА ДЛЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ СЕМАНТИЧНОЇ ПОДІБНОСТІ ТЕКСТОВИХ ДОКУМЕНТІВ

Робота присвячена дослідженню та застосуванню метрик косинусної схожості та індексу Жаккара у контексті рекомендації текстових документів відповідно до запитань користувачів. Згідно до проведеного аналізу, обрано використання зазначених метрик у контексті порівняння векторів, що представляють текстові документи. Отримані результати аналізу можуть бути використані для подальшого вдосконалення систем рекомендацій та інформаційного пошуку, зокрема в областях, де ключовим є семантичне порівняння текстової інформації.

The work is devoted to the study and application of cosine similarity metrics and the Jaccard index in the context of recommending text documents according to user queries. According to the conducted analysis, it was chosen to use the specified metrics in the context of comparing vectors representing text documents. The results of the analysis can be used for further improvement of recommendation systems and information retrieval, in particular in areas where semantic comparison of textual information is key.

Наразі існує багато методів пошуку релевантних документів за запитом користувача. Один з найпоширеніших методів – це використання метрик подібності. Метрики подібності використовуються для вимірювання подібності між запитом та документом. Чим більша подібність між запитом та документом, тим більш релевантний документ [1].

Нижче наведено деякі поширені метрики подібності [1]:

– TF-IDF – це метрика, яка враховує частоту виникнення слів у документі та частоту виникнення слів у наборі документів.

– Косинусна схожість (Cosine similarity) – це метрика, яка враховує кут між векторами запиту та документа.

– Індекс Жаккара (Jaccard similarity) – це метрика, яка враховує кількість спільних слів у запиті та документі.

Метрики косинусної схожості та індексу Жаккара пропонується використовувати у рамках методу рекомендації текстових документів за запитаннями користувачів на базі служби психологічної підтримки призначений для пошуку релевантних рекомендацій з бази текстових порад за текстовим запитом користувача та вхідними даними має датасет із набором очищених лематизованих асоціативних запитів, векторизований корпус та користувацький запит, що

перетворює їх у вихідні дані у вигляді текстової рекомендації релевантної до запиту.

Основна ідея косинусної схожості полягає в тому, що якщо два вектори спрямовані в одному напрямку (кут між ними дорівнює 0 градусів), то їхні вектори мають найвищу косинусну схожість та вважаються дуже схожими. Навпаки, якщо кут між векторами дорівнює 90 градусів, то косинусна схожість дорівнює 0, що вказує на максимальну різницю між ними [2].

Для обчислення косинусної схожості між двома векторами тексту (наприклад, словниковими представленнями документів), використовується наступна формула:

$$Cosine\ Similarity\ (A, B) = (A \bullet B) / (||A|| * ||B||),$$

де: *A* та *B* – вектори тексту або документів. $A \bullet B$ - скалярний добуток векторів *A* і *B*. $||A||$ та $||B||$ – норми (довжини) векторів *A* і *B*.

Значення косинусної схожості зазвичай лежать в діапазоні від -1 (повна протилежність) до 1 (повна ідентичність), де 0 означає відсутність схожості.

Цей метод широко використовується в пошукових системах, рекомендаційних системах, аналізі тексту та класифікації документів для визначення ступеня схожості між текстами та відбору найбільш релевантних результатів. У роботі буде застосований як одна із метрик подібності запиту користувача до наявних шаблонів запитів з метою отримання релевантної відповіді.

Індекс Жаккара (Jaccard index), також відомий як Жаккардова схожість чи коеліцієнт Жаккара, є метрикою схожості, яка використовується для порівняння множин елементів. Ця метрика визначає ступінь схожості двох множин шляхом вимірювання кількості спільних елементів в обох множинах відносно загальної кількості унікальних елементів [3].

Індекс Жаккара обчислюється за наступною формулою:

$$J(A, B) = |A \cap B| / |A \cup B|,$$

де*: J(A, B)* – індекс Жаккара між множинами A і B.$|A \cap B|$ – кількість спільних елементів між множинами *A* і *B*. $|A \cup B|$ – кількість унікальних елементів у множинах *A* і *B* разом.

Значення Індексу Жаккара можуть лежати в діапазоні від 0 до 1, де 0 означає відсутність спільних елементів, а 1 означає повну ідентичність множин. Значення, близькі до 1, вказують на високу схожість множин, тоді як значення, близькі до 0, свідчать про низьку схожість.

Індекс Жаккара широко використовується в різних галузях, включаючи аналіз тексту, рекомендаційні системи, пошукові системи та біологічні дослідження. В роботі буде використовуватися для визначення схожості між ключовими словами користувацького запиту та наявних рекомендацій.

Перелік посилань

1. Cosine Similarity. URL: https://www.geeksforgeeks.org/cosine-similarity/
2. Cosine Similarity. URL: https://www.learndatasci.com/glossary/cosine-similarity/
3. Jaccard Similarity Made Simple: A Beginner's Guide to Data Comparison. URL: https://medium.com/@mayurdhvajsinhjadeja/jaccard-similarity-34e2c15fb524