ISSN 2222-2944. Інформаційні технології: наука, техніка, технологія, освіта, здоров'я. 2023

МОДЕЛІ ТА АЛГОРИТМИ ПОШУКУ КЛЮЧОВИХ СЛІВ У ТЕКСТІ

Богряшов А. О., Пікалова В. В.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Кожну хвилину в пошукову систему Google робиться близько 3 мільйонів запитів. Ці запити оброблюються надзвичайно швидко, і вже за декілька мілісекунд користувач отримує результати пошуку. Це можливо завдяки алгоритмам та моделям пошуку ключових слів у тексті, головною метою яких є швидке та ефективне виділення головної теми та ідеї з тексту. Це корисно для автоматичного аналізу великих обсягів тестової інформації, таких як соціальні мережі, блоги, новинні статті, наукові дослідження та інші.

Існує безліч моделей та алгоритмів пошуку ключових слів у тексті [1, 2]. Ось деякі з них.

TF-IDF (term frequency-inverse document frequency) – це статистична модель, яка використовується для визначення важливості кожного слова в документі в порівнянні з іншими документами у корпусі. Слова, які відображаються в більшості документів, отримують менше ваги, ніж ті, які відображаються лише в кількох.

TextRank – це алгоритм, який використовує граф зв'язків між реченнями у тексті, щоб визначити важливість кожного слова. Він дозволяє виділяти найбільш важливі слова, які з'являються в багатьох реченнях, а також ті, які з'являються в реченнях, які є важливими.

RAKE (Rapid Automatic Keyword Extraction) – це алгоритм, який використовується для автоматичного виділення ключових слів у тексті. Він працює шляхом визначення частин слів, які відповідають певним критеріям, таким як частота зустрічі інших слів в тому ж реченні.

LSA (Latent Semantic Analysis) – це модель, яка використовується для знаходження семантичних зв'язків між словами у тексті. Вона дозволяє знаходити слова, які використовуються в близькому контексті з іншими важливими словами.

Моделі та алгоритми пошуку ключових слів в тексті мають значний практичний потенціал: пошук інформації, рекомендації контенту, пошук схожих документів та інше. Кожен з розглянутих методів має свої переваги та недоліки, та їх ефективність залежить від типу тексту та мети, для якої вони використовуються. Метою нашої роботи є проведення порівняльного аналізу описаних алгоритмів та створення підсумкової таблиці, що дозволить розробнику зорієнтуватися у виборі найоптимальнішого алгоритму пошуку ключових слів для поставленої перед ним задачі.

Література:

1. Jones K. S. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation: журнал. – MCB University : MCB University Press, 2004. – Т. 60, № 5. – С. 493-502.

2. Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411). https://www.aclweb.org/anthology/W04-3252.pdf