МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ
Кафедра интеллектуальных систем

Аннотация к дипломной работе

ОПРЕДЕЛЕНИЕ КЛЮЧЕВЫХ СЛОВ В ТЕКСТОВОМ СООБЩЕНИИ

Журавлёва Есения Дмитриевна

Научный руководитель: профессор кафедры интеллектуальных систем, кандидат технических наук, доцент Садов В.С

Минск, 2025

РЕФЕРАТ

Дипломная работа содержит 50 с., 4 рис., 14 источн.

ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ СЛОВ, ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, СЕМАНТИЧЕСКИЙ ПОИСК, СХОДСТВО ТЕКСТОВ, ВЕКТОРНЫЕ ПРЕДСТАВЛЕНИЯ, КОСИНУСНОЕ СХОДСТВО

Объект исследования: ключевые слова

Цель работы: разработка и реализация программного приложения для автоматического извлечения ключевых слов из пользовательского текста и поиска семантически схожих документов на основе этих ключевых слов.

В работе изучены современные подходов к извлечению ключевых слов. Также исследованы основные методы определения семантического сходства текстов. Спроектирована и реализована архитектурная модель взаимодействия компонентов приложения, включающая модули предобработки, извлечения ключевых слов, векторизации и поиска сходства. Спроектирован и реализованы модуль извлечения ключевых слов из предложенного пользователем текста и модуль поиска схожих текстов на основе извлеченных ключевых слов, что обеспечивает нахождение семантически близких документов в тестовом корпусе.

Область применения: разработанная программа может быть использовано в различных сферах, требующих автоматизации анализа больших объемов текстовых данных: информационный поиск, категоризация документов, суммаризация, анализ отзывов, мониторинг новостных лент, интеллектуальные системы документооборота и другие области, где необходимо быстрое извлечение сути информации и выявление связей между документами.

РЭФЕРАТ

Дыпломная праца змяшчае 50 с., 4 мал., 14 крыніц.

ВЫЯМЛЕННЕ КЛЮЧАВЫХ СЛОЎ, АПРАЦОЎКА ПРАЦОЎНАЙ МОВЫ, СЕМАНТЫЧНЫ ПОШУК, ПАДАБЕНСТВА ТЭКСТАЎ, ВЕКТАРНЫЯ ПРАДСТАЎЛЕННЯ, КОСІНУСНЫХ ПАДАБЕНСТВА

Аб'ект даследавання: ключавыя словы

Мэта працы: распрацоўка і рэалізацыя праграмнага прыкладання для аўтаматычнага вымання ключавых слоў з карыстацкага тэксту і пошуку семантычна падобных дакументаў на аснове гэтых ключавых слоў.

У працы даследаваны сучасныя падыходаў да вызначэння ключавых слоў. Таксама даследаваны асноўныя метады вызначэння семантычнага падабенства тэкстаў. Спраектавана і рэалізавана архітэктурная мадэль узаемадзеяння кампанентаў праграммы, якая ўключае модулі перадапрацоўкі, вызначэння ключавых слоў, вектарызацыі і пошуку падабенства. Спраектаваны і рэалізаваны модуль вызначэння ключавых слоў з прапанаванага карыстальнікам тэксту і модуль пошуку падобных тэкстаў на аснове вынятых ключавых слоў, што забяспечвае знаходжанне семантычна блізкіх дакументаў у тэставым корпусе.

Вобласць выкарыстання: распрацаваная праграмма можа быць выкарыстана ў розных сферах, якія патрабуюць аўтаматызацыі аналізу вялікіх аб'ёмаў тэкставых дадзеных: інфармацыйны пошук, катэгарызацыя дакументаў, сумарызацыя, аналіз водгукаў, маніторынг стужак навін, інтэлектуальныя сістэмы дакументазвароту і іншыя вобласці, дзе неабходна хуткае выманне сутнасці інфармацыі і выяўленне сувязяў.

ABSTRACT

The thesis contains 50 p., 4 figures, 14 sources.

KEYPHRASE EXTRACTION, NATURAL LANGUAGE PROCESSING, SEMANTIC SEARCH, TEXT SIMILARITY, VECTOR REPRESENTATIONS, COSINE SIMILARITY

Research object: Keywords

Research goal: To develop and implement a software application for automatically extracting keywords from user-provided text and finding semantically similar documents based on these keywords.

The work examines contemporary approaches to keyword extraction. It also investigates fundamental methods for determining semantic text similarity. An architectural model for component interaction within the application was designed and implemented. This model includes modules for preprocessing, keyword extraction, vectorization, and similarity search. A module for extracting keywords from user-submitted text and a module for finding similar texts based on the extracted keywords were designed and implemented. This enables the identification of semantically related documents within a test corpus.

Application area: The developed application can be utilized in various domains requiring automation of large-scale text data analysis: information retrieval, document categorization, text summarization, sentiment analysis, news feed monitoring, intelligent document management systems, and other fields where rapid extraction of essential information and identification of connections between documents is necessary.