БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра дискретной математики и алгоритмики

Аннотация к магистерской диссертации

«Эвристические алгоритмы для выделения ключевых слов текстов веб-страниц на стороне клиента»

Кулик Сергей Дмитриевич

Научный руководитель – доктор физико-математических наук,
Профессор Котов В.М.

Минск, 2019

Реферат

Магистерская диссертация, 59 страниц, 20 источников, 2 приложения.

АНАЛИЗ ТЕКСТОВ; ВЫДЕЛЕНИЕ КЛЮЧЕВЫХ СЛОВ; МАШИННОЕ ОБУЧЕНИЕ; ВЫЧИСЛЕНИЯ НА СТОРОНЕ КЛИЕНТА.

Объект исследования – тексты на естественных языках и ключевые слова в них; записи и заметки в блогах, статьи.

*Цель рабо*ты – исследование методов для выделения ключевых слов текстов на естественных языках на стороне клиента. Цель включает в себя как исследование известных результатов в рассматриваемой задаче и смежных областях, так и поиск новых.

В ходе работы были исследованы имеющиеся подходы для выделения ключевых слов как на стороне клиента, так и на стороне сервера. Были рассмотрены проблемные вопросы построения тренировочных выборок для рассматриваемой задачи, также предложена методология для сравнения получаемых результатов с известными аналогами. Также рассмотрены проблемные вопросы выделения стоп-слов в текстах.

Результатом является метод для выделения ключевых слов на основе алгоритма логистической регрессии. Для данного метода были также проведены вычислительные эксперименты и сравнение результатов с известными аналогами. Показана переносимость данного метода: будучи обучен на текстах на английском языке, метод также показывает высокие результаты при тестировании на текстах на испанском языке.

Областью применения является разработка на стороне клиента; приложения, работающие с конфиденциальными данными; алгоритмы категоризации текстов на естественном языке.

Abstract

Master thesis, 59 pages, 20 references, 2 appendices.

TEXT MINING; KEYWORD DETECTION; MACHINE LEARNING; CLIENT-SIDE CALCULATIONS.

The object of the study are texts in natural languages and keywords in them; records and notes on blogs and articles.

The objective of the thesis is the study of methods for keywords detection in texts in natural languages on the client side. The goal includes both: the study of known results in the problem and related fields, and the research of new ones.

During the course of the research, the available approaches for detecting keywords on the client side and on the server side were researched. We've also addressed the problematic issues of building training set for the problem, also proposed a methodology for comparing the obtained results with known analogues. The problematic issues of stop words detection in the texts were also addressed.

The result is a proposed method for extracting keywords based on a logistic regression algorithm. For this method, computational experiments were also conducted and the results were compared with known analogues. The portability of this method was shown: being trained on texts in English, the method also shows good results when tested on texts in Spanish.

The field of application is client-side development; applications that work with confidential data; natural language categorization algorithms.