БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Факультет радиофизики и компьютерных технологий
Кафедра интеллектуальных систем

Аннотация к дипломной работе

«Алгоритмы анализа и извлечения структурированных в сети Интернет»

Шишко Владислав Олегович

Научный руководитель: старший преподаватель А.В. Курочкин

Минск, 2020

РЕФЕРАТ

Дипломная работа: 64 страницы, 26 использованных источников, 3 иллюстрации, 7 приложений.

АЛГОРИТМЫ АНАЛИЗА И ИЗВЛЕЧЕНИЯ СТРУКТУРИРОВАННЫХ ДАННЫХ В СЕТИ ИНТЕРНЕТ.

Объект исследования - алгоритмы анализа структурированных данных, веб-краулеры.

Цель работы - разработка приложения, использующего алгоритмы анализа и обработки данных, для поиска продуктов сети интернет.

Методы исследования - изучение литературы и применение изученного на практике.

В исследовании используются различные алгоритмы для сбора данных сети интернет, а также их последующей обработки и визуализации, в частности при помощи мессенджера Telegram.

В результате данной работы было разработано приложение, осуществляющее поиск товаров в сети интернет, и предоставляющее собранные данные в понятном для конечного пользователя виде.

Так же был разработан собственный веб-краулер, добавлено использование баз данных для сохранения исторических данных, а также реализован метод случайного леса для анализа собранного контента.

Было проведено тестирование системы и проверка работоспособности каждого этапа его работы. Результаты показали, что приложение собирает и обрабатывает продукты сети интернет и имеет потенциал для дальнейшего развития.

Так же результаты данной работы были использованы во внутренних проектах компании «ООО Айтакко».

РЭФЕРАТ

Дыпломная праца: 64 старонкі, 26 выкарыстаных крыніц, 3 іллюстрацыі, 7 прыкладанняў.

АЛГАРЫТМЫ АНАЛІЗУ І ВЫМАННІ СТРУКТУРАВАНЫХ ДАДЗЕНЫХ У СЕТЦЫ ІНТЭРНЕТ.

Аб'ект даследавання - алгарытмы аналізу структураваных дадзеных, вэб-краулеры.

Мэта даследвання - распрацоўка прыкладання, якое выкарыстоўвае алгарытмы аналізу і апрацоўкі дадзеных, для пошуку прадуктаў сеткі інтэрнэт.

Метады даследавання - вывучэнне літаратуры і прымяненне вывучанага на практыцы.

У даследаванні выкарыстоўваюцца розныя алгарытмы для збору дадзеных сеткі інтэрнэт, а таксама іх наступнай апрацоўкі і візуалізацыі, у прыватнасці пры дапамозе мессенджера Telegram.

У выніку гэтай работы было распрацавана прыкладанне, якое ажыццяўляе пошук тавараў у сеткі інтэрнэт, і якая прадастаўляе сабраныя дадзеныя ў зразумелай для канчатковага карыстальніка выглядзе.

Гэтак жа быў распрацаваны ўласны вэб-краўлер, дададзена выкарыстанне баз дадзеных для захавання гістарычных дадзеных, а таксама рэалізаваны метад выпадковага лесу для аналізу сабранага кантэнту.

Было праведзена тэставанне сістэмы і праверка працаздольнасці кожнага этапу яго працы. Вынікі паказалі што прыкладанне збірае і апрацоўвае прадукты сеткі інтэрнэт і мае патэнцыял для далейшага развіцця.

Таксама вынікі дадзенай працы былі выкарыстаны ва ўнутраных праектах кампаніі «ТАА Айтакко».

ABSTRACT

Thesis: 64 pages, 26 sources, 3 figures, 7 applications.

ALGORITHMS FOR ANALYSIS AND EXTRACTION OF STRUCTURED DATA ON THE INTERNET.

The object of research - structured data analysis algorithms, web crawlers.

Objective - to develop an application that uses data analysis and processing algorithms to search for Internet products.

Research methods - study of literature and application of the studied in practice.

The study uses various algorithms to collect data from the Internet, as well as their subsequent processing and visualization, in particular using the Telegram messenger.

As a result of this work, an application was developed that searches for goods on the Internet and provides the collected data in a way that is understandable to the end user.

We also developed our own web crawler, added the use of databases to save historical data, and also implemented a random forest method for analyzing collected content.

The system was tested and tested for operability of each stage of its operation. The results showed that the application collects and processes Internet products and has the potential for further development.

Also, the results of this work were used in the internal projects of «Aitakko LLC».