БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Факультет радиофизики и компьютерных технологий
Кафедра интеллектуальных систем

Аннотация к дипломной работе

«Веб-приложение для оценки количества просмотров интернет-публикаций»

Михайлова Татьяна Николаевна

Научный руководитель: кандидат физ.- мат. наук, доцент Д. В. Щегрикович

2017

РЕФЕРАТ

Дипломная работа: 55 страниц, 17 рисунков, 3 таблицы, 42 использованных источника, 3 приложения.

ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, КОРПУСНАЯ ЛИНГВИСТИКА, ИСКУСТВЕННЫЙ ИНТЕЛЛЕКТ, МАШИННОЕ ОБУЧЕНИЕ, БИНАРНАЯ КЛАССИФИКАЦИЯ, ВЕБ-ПРИЛОЖЕНИЕ.

Объект исследования - новостные интернет-публикации.

Цель работы - разработка веб-приложения для оценки количества просмотров интернет-публикаций в будущем на основе содержащейся в них текстовой информации.

Методы исследования - методы обработки естественного языка.

Для проведения исследования выбран интернет-ресурс «Forbes». Сформирована исходная выборка из новостных публикаций и данных о количестве их просмотров за неделю.

Для извлечения признаков из текстов интернет-публикаций, разработаны функции с использованием методов обработки естественного языка: графематического, морфологического, синтаксического, сентиментального, пунктуационного анализов. В результате реализовано извлечение 120 признаков, характеризующих текст на естественном языке.

Проведен сравнительный анализ методов машинного обучения по точности, скорости и минимальному количеству настраиваемых параметров. В результате сравнения выбран метод «Случайный лес».

Проанализирована зависимость точности классификации методом «Случайный лес» от количества решающих деревьев в ансамбле и количества признаков, выбираемых для обучения очередного дерева. Выбраны оптимальные значения – 500 и 6. Достигнутая точность работы модели – 67% правильно классифицируемых объектов из тестовой выборки. Определены наиболее значимые и не значимые признаки для популяризации текстовых материалов.

Полученное программное решение реализовано в виде веб-приложения.

РЭФЕРАТ

Дыпломная праца: 55 старонак, 17 малюнкаў, 3 табліцы, 42 выкарыстанных крыніцы, 3 дадатка.

АПРАЦОЎКА НАТУРАЛЬНАЙ МОВЫ, КОРПУСНАЯ ЛІНГВІСТЫКА, ШТУЧНЫ ІНТЭЛЕКТ, МАШЫННАЕ НАВУЧАННЕ, БІНАРНАЯ КЛАСІФІКАЦЫЯ, ВЭБ-ПРЫКЛАДАННЕ.

Аб'ект даследавання - навінавыя інтэрнэт-публікацыі.

Мэта - распрацоўка вэб-прыкладання для ацэнкі колькасці праглядаў інтэрнэт-публікацый у будучыні на аснове тэкставай інфармацыі, якая прысутнічае ў іх.

Метады даследавання - метады апрацоўкі натуральнай мовы.

Для правядзення даследвання быў выбраны інтэрнэт-рэсурс «Forbes». Сфарміравана сыходная выбарка з публікацый навінаў і дадзеных аб колькасці іх праглядаў за тыдзень.

Для вылучэння значэння прыкметаў з тэксатаў інтэрнэт-публікацый былі расправаны функцыі з выкарыстаннем метадаў апрацоўкі натуральнай мовы: графематычнага, марфалагічнага, сінтаксічнага, сентыментальнага, пунктуацыйнага аналізаў. У выніку былі атрыманы 120 прыкметаў, якія характырызуюць тэкст на натуральнай мове.

Быў праведзены параўнаўчы аналіз метадаў машыннага навучання па дакладнасці, хуткасці і мінімальнай колькасці параметраў, якія наладжваюцца. У выніку параўнання выбраны метад «Выпадковы лес».

Прааналізавана залежнасць дакладнасці класіфікацыі метадам «Выпадковы лес» ад колькасці рашаючых дрэваў у ансамблі і колькасці знакаў, якія выбіраюцца для навучання чарговага дрэва. Выбраны аептымальныя значнні – 500 і 6. Дасягнутая дакладнасць класіфікацыі – 67% правільна класіфікаваных аб'ектаў з тэставай выбаркі. Вызначаны найбольш значныя і ня значныя прыкметы для папулярызацыі тэкставых матэрыялаў.

Атрыманае праграмнае рашэнне рэалізавана ў выглядзе вэб-прыкладання.

ABSTRACT

Thesis: 55 pages, 17 figures, 3 tables, 41 sources, 3 applications.

NATURAL LANGUAGE PROCESSING, CORPUS LINGUISTICS, ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, BINARY CLASSIFICATION, WEB APPLICATION.

The object of research is the news Internet publications.

Objective is the web application development for estimate the quantity of the news Internet publications in future based on the textual information.

The methods is the natural language processing methods.

The Internet resource "Forbes" was chosen for research. Training and test data sets were formed from the textual information of publications and number of views in a week.

Functions based on the natural language processing methods were developed for extracting the features from the textual information of publications. Graphematic, morphological, semantic, sentimental and punctuation analyzes were used. As a result the extraction of 120 features characterizing the text in natural language was realized.

Methods of machine learning were compared in accuracy, speed, and the minimum number of configurable parameters. As a result "Random Forest" method was chosen.

The dependence of the accuracy of the "Random Forest" method on the number of decision trees in the ensemble and the number of features chosen for the training of the next tree was analyzed. The optimal values are 500 and 6. The achieved accuracy of the model is 67% of correctly classified objects from the test data set. The most important and not important features for popularization of textual information were identified.

The resulting software solution was developed as a web application.