БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Кафедра компьютерных технологий и систем

Аннотация к дипломной работе

СТАТИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТА И ИССЛЕДОВАНИЕ ЕГО ЦЕЛЕВОГО НАЗНАЧЕНИЯ

Клюев Владислав Дмитриевич

Научный руководитель – старший преподаватель кафедры компьютерных технологий и систем, Лагуто Анна Андреевна

Минск 2020

РЕФЕРАТ

Дипломная работа: 52 с., 23 рис., 11 табл., 22 источника.

СТАТИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТА, МАШИННОЕ ОБУЧЕНИЕ, АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТА, АНАЛИЗ ДАННЫХ, ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, КЛАССИФИКАЦИЯ ДАННЫХ

Объектом исследования являются различные наборы текстов, использующиеся в них языковые единицы, обладающие свойством знака.

Цель исследования состоит в построении моделей машинного обучения на основании статистических признаков текста, программной реализации методов классификации текста.

Методы исследования представлены методами обработки естественного языка, статистических характеристик текстов, методами машинного обучения. Задача решалась в программной среде Jupyter Notebook на языке Python с использованием следующих библиотек: Keras, SKLearn, Pandas и других библиотек.

В результате исследования построены модели машинного обучения для анализа тональности текстов, реализованы различные классификаторы для определения эмоциональной окраски текстов, произведена оценка и сравнение результатов, полученных после обучения, сделаны выводы о проблемах существующей реализации, а также о способах ее улучшения.

Областью применения являются задачи классификации текста.

Дипломная работа выполнена автором самостоятельно.

ABSTRACT

Thesis: 52 p., 23 fig., 11 tab., 22 sources.

STATISTICAL ANALYSIS OF THE TEXT, MACHINE LEARNING, ANALYSIS OF THE TONALITY OF THE TEXT, ANALYSIS OF DATA, PROCESSING THE NATURAL LANGUAGE, CLASSIFICATION OF DATA

The object of the study is various sets of texts, the language units used in them that have the property of a sign.

The purpose of the study is to build machine-learning models based on statistical features of the text, software implementation of text classification methods.

Research methods consist of natural language processing methods, statistical characteristics of texts, machine-learning methods. The problem was solved in the Python Jupyter Notebook software environment using the following libraries: Keras, SKLearn, Pandas and other libraries.

As a result of the study, machine learning models for analyzing the tonality of texts were built, various classifiers for determining the emotional coloring of texts were implemented, the results obtained after training were evaluated and compared, conclusions were drawn about the problems of the existing implementation, as well as ways to improve it.

The scope is tasks of text classification.

Thesis is performed by the author independently.