

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра дискретной математики и алгоритмики

Аннотация к дипломной работе

«Исследование методов выделения смысловых единиц из деловой переписки»

Гулин Кирилл Иванович

Научный руководитель – доцент Свирид Ю. В.

Минск, 2021

Реферат

Дипломная работа, 49 с., 26 рис., 8 таблиц, 19 источников.

Ключевые слова: ОБРАБОТКА ТЕКСТОВ, МАШИННОЕ ОБУЧЕНИЕ, ЛАТЕНТНОЕ РАЗМЕЩЕНИЕ ДИРИХЛЕ, МЕТОДЫ ПОНИЖЕНИЯ РАЗМЕРНОСТИ, ВЕКТОРНОЕ ПРЕДСТАВЛЕНИЕ ТЕКСТОВ, СТАТИСТИЧЕСКИЙ АНАЛИЗ, МЕТОДЫ КЛАСТЕРИЗАЦИИ, НЕЙРОННЫЕ СЕТИ.

Объект исследования — наборы данных деловых электронных переписок.

Цель работы — анализ деловых электронных переписок методами машинного обучения.

Методы исследования — латентное размещение Дирихле, методы кластеризации, методы обработки текстов, методы понижения размерности, методы получения векторных представлений.

Работа посвящена исследованию и анализу деловых электронных переписок, в частности, переписок Хиллари Клинтон и переписок сотрудников корпорации Enron. В результате работы был произведен статистический анализ электронных переписок. Были обнаружены закономерности в исходных данных. Также была разработана кластеризация содержаний электронных писем, в результате которой получились интерпретируемые результаты, что показало эффективность разработанных методов.

Abstract

Diploma thesis, 49 p., 26 fig., 8 tables, 19 sources.

Keywords: TEXT PROCESSING, MACHINE LEARNING, LATENT DIRICHLET ALLOCATION, DIMENSION REDUCTION METHODS, TEXT EMBEDDINGS, STATISTICAL ANALYSIS, CLUSTERIZATION METHODS, NEURAL NETWORKS.

The object of research is business e-mail datasets.

Objective: analysis of business e-mails using machine learning methods.

Research methods --- latent Dirichlet allocation, clustering methods, text processing methods, dimension reduction methods, methods for obtaining text embeddings.

The work is devoted to the research and analysis of business e-mails, in particular, the e-mails of Hillary Clinton and the e-mails of employees of the Enron corporation. As a result of the work, a statistical analysis of emails was carried out. Patterns were found in the original data. Also, the clustering of the contents of e-mails was developed, as a result of which interpretable results were obtained, which showed the effectiveness of the developed methods.