Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Системы обработки информации и управления»

Лабораторная работа №6
По курсу «Методы машинного обучения»

«Классификация текста»

ИСПОЛНИТЕЛЬ:

Чичикин Тимофей Дмитриевич
Группа ИУ5-25М

_____

ПРОВЕРИЛ:

Гапанюк Ю.Е.

_____

Цель работы:

Изучение методов классификации текста.

Задание:

Для произвольного набора данных, предназначенного для классификации текстов, решите задачу классификации текста двумя способами:

1. Способ 1. На основе CountVectorizer или TfidfVectorizer.
2. Способ 2. На основе моделей word2vec или Glove или fastText.
3. Сравните качество полученных моделей.

Описание задания:

Для выполнения лабораторной работы возьмём датасет с обзорами фильмов IMDB для анализа настроений, где выделена целевая переменная: 1 – положительное мнение, а 0 – отрицательное.

Выполнение работы:

1. Классификация текста на основе CountVectorizer
2. Классификация текста на основе модели word2vec
3. Сравнение качества полученных моделей

Вывод:

Была проделана работа по изучению методов классификации текста, в результате чего можно сделать вывод, что для данного датасета наибольшая точность получилась при использовании CountVectorizer и LogisticRegression.