Московский государственный технический университет

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа № 8**

**По курсу «методы машинного обучения в АСОИУ»**

**«Предобработка текста»**

**Выполнил:**

студент ИУ5-24М

Ширшов А.С.

**Проверил:**

Гапанюк Ю.Е.

Подпись:

29.02.2024

Москва, 2024

**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.

Для поиска наборов данных в поисковой системе можно использовать ключевые слова "datasets for text classification".

**Ход работы**

Скачаем с сети набор данных Imdb. Применим на нём предложенные методы.

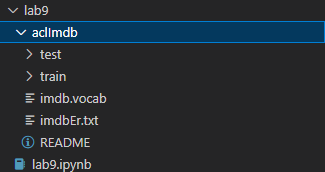


Рисунок 1 - Структура проекта

Указываем путь к данным.

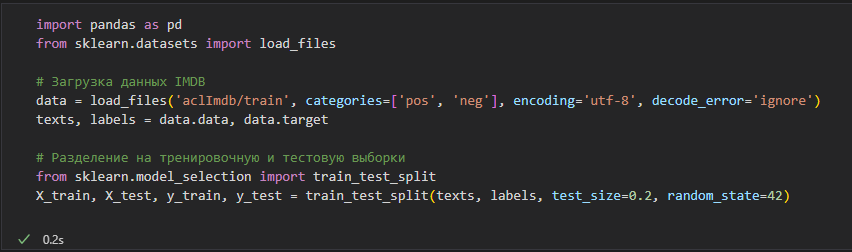


Рисунок 2 - Начало работы с данными

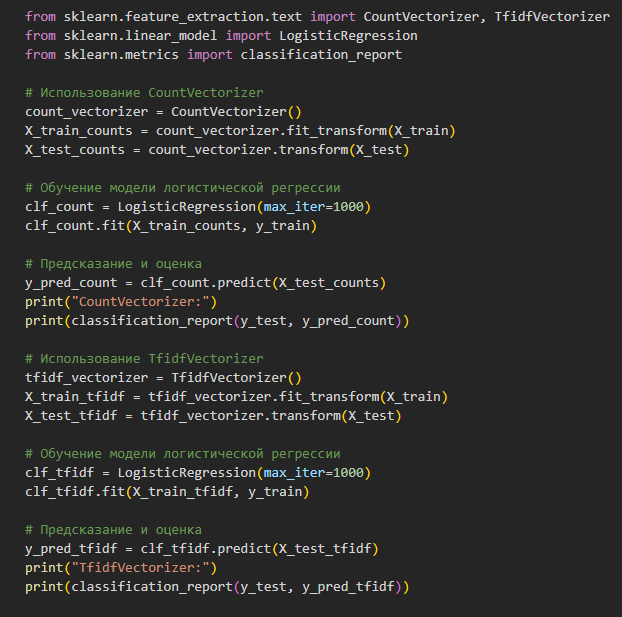


Рисунок 3 - Запускаем CountVectorizer и TfidfVectorizer

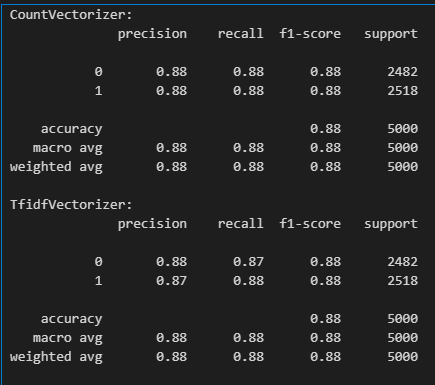


Рисунок 4 - Полученные результаты CountVectorizer и TfidfVectorizer

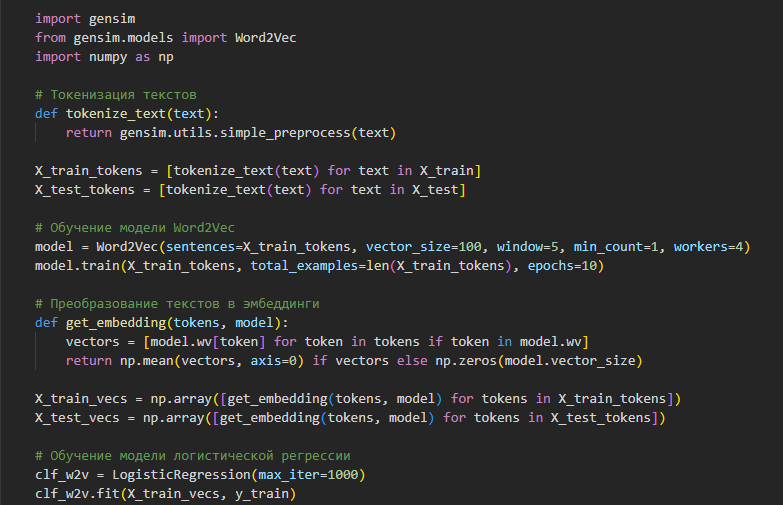


Рисунок 5 - Word2Wec

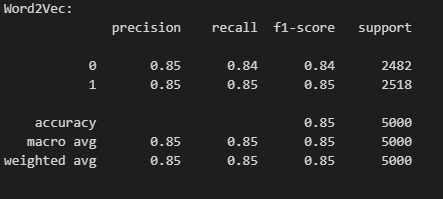


Рисунок 6 - Результаты Word2Vec