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

Лабораторная работа №9  
по дисциплине  
«Методы машинного обучения»  
на тему

# «Классификация текста»

Выполнил:  
студент группы ИУ5И-22М  
Лу Жуньда

Москва — 2024 г.

1. **Цель лабораторной работы**

Изучение методов классификации текстов.

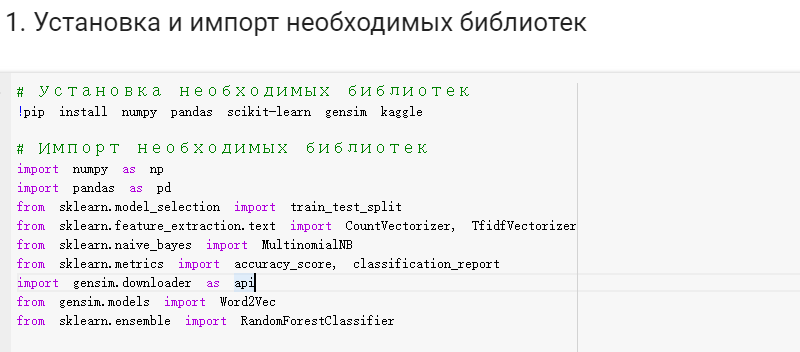
1. **Задание**

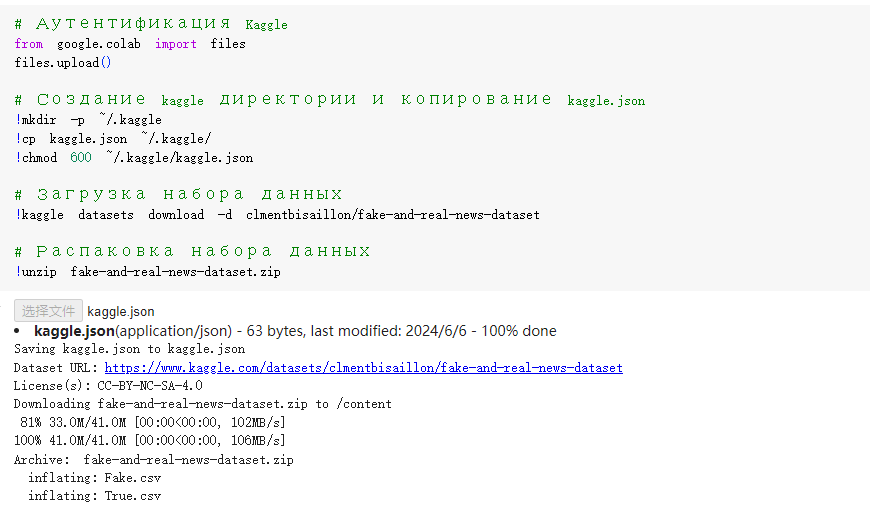
Для произвольного набора данных, предназначенного для классификации текстов, решите задачу классификации текста двумя способами:

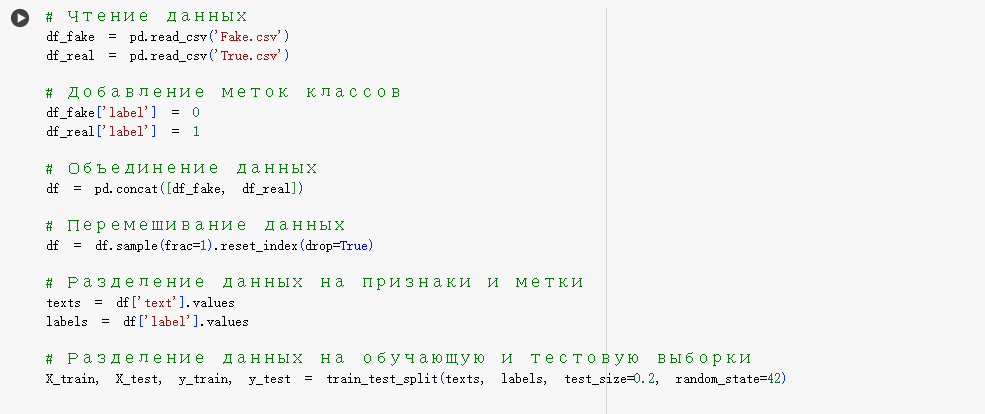
1. Способ 1. На основе CountVectorizer или TfidfVectorizer.
2. Способ 2. На основе моделей word2vec или Glove или fastText.
3. Сравните качество полученных моделей.

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

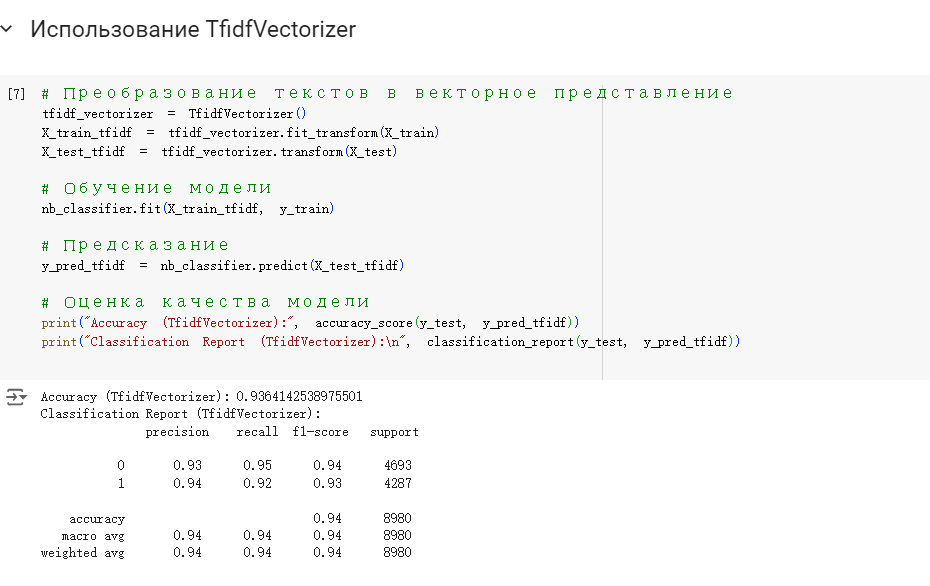
1. **Текст программы**

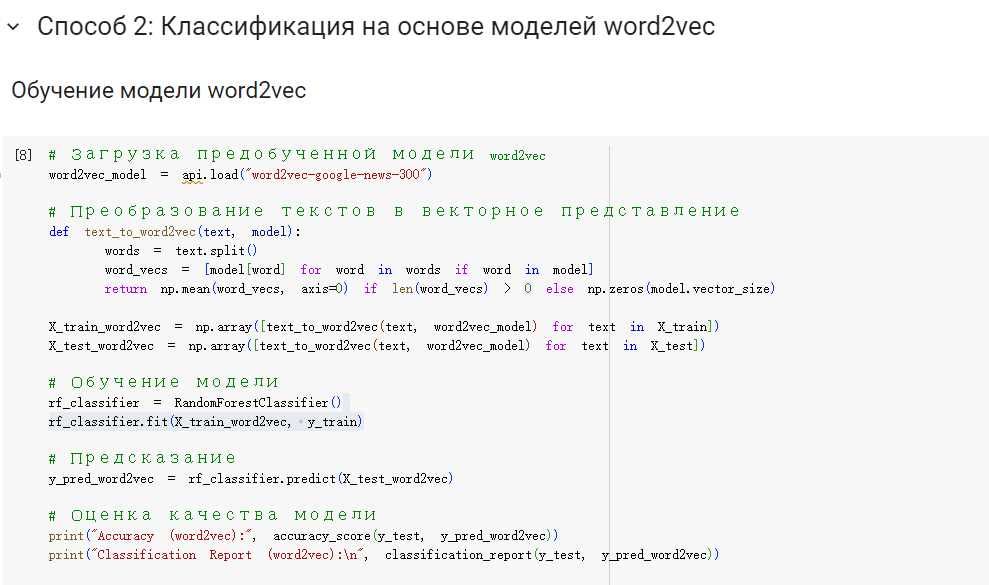
****

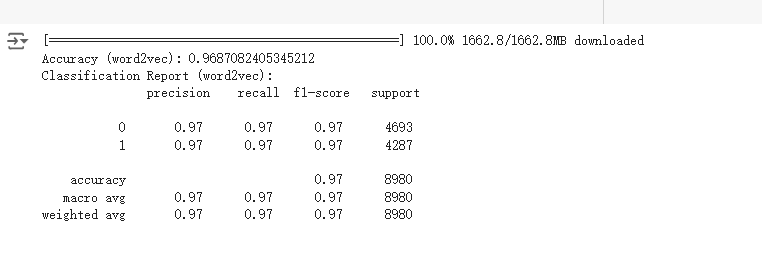
****

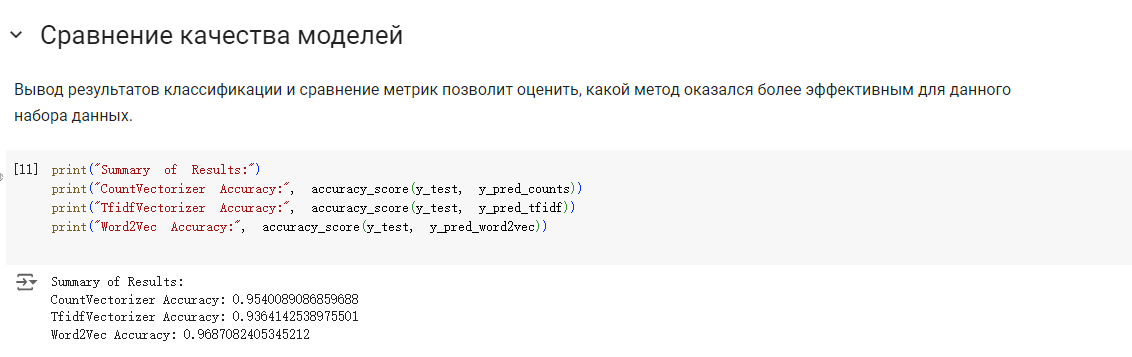
****

****

****

****

****

****

**Заключение**

В данной лабораторной работе мы рассмотрели два различных подхода к классификации текстов и сравнили их результаты. Вот сводка результатов:

**CountVectorizer:**

Точность: 95.40%

Методы на основе CountVectorizer показали хорошую точность, демонстрируя, что частотные характеристики слов могут эффективно использоваться для классификации текстов.

**TfidfVectorizer:**

Точность: 93.64%

Методы на основе TfidfVectorizer также показали высокую точность, хотя немного ниже по сравнению с CountVectorizer. Использование TF-IDF позволяет учесть важность слов в контексте всего корпуса текстов, что также является мощным инструментом для классификации.

**Word2Vec:**

Точность: 96.87%

Методы на основе Word2Vec показали наивысшую точность среди рассмотренных методов. Это может быть связано с тем, что Word2Vec учитывает семантические связи между словами, что позволяет более точно представлять тексты и улучшать качество классификации.