МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Систем обработки информации и управления»

ОТЧЕТ

**Лабораторная работа №\_\_9\_\_**

по дисциплине«Методы машинного обучения в автоматизированных системах обработки информации и управления»

Тема: « Категоризация текста»

ИСПОЛНИТЕЛЬ: \_\_Се Цзявэнь\_\_\_\_

ФИО

Группа : ИУ5И-22М\_

подпись

"\_\_"\_\_\_\_\_\_\_\_\_202\_ г.

ПРЕПОДАВАТЕЛЬ: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ФИО

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_22\_"\_\_\_\_\_05\_\_\_\_2024\_ г.

Москва - 2024

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

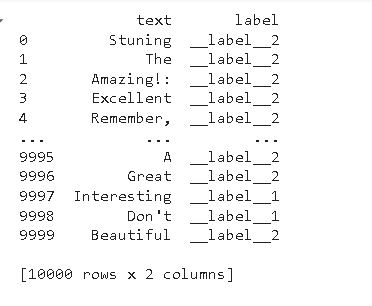
# Задания

* Для произвольного набора данных, предназначенного для классификации текстов, решите задачу классификации текста двумя способами:
* Способ 1. На основе CountVectorizer или TfidfVectorizer.
* Способ 2. На основе моделей word2vec или Glove или fastText.
* Сравните качество полученных моделей.

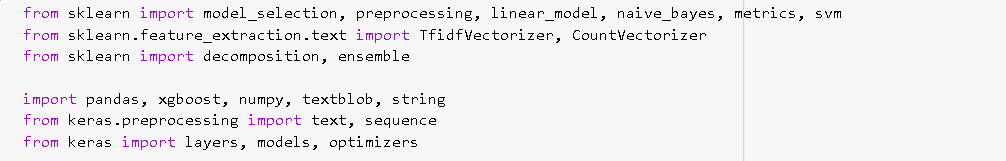
# Программы

* Текстовое описание набора данных:

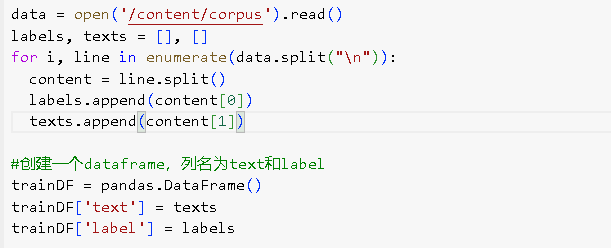
Набор данных обзора Amazon.



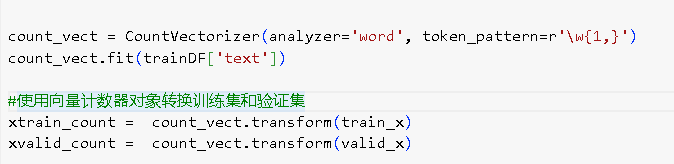
Импорт библиотек



* Загрузка данных



* CountVectorizer

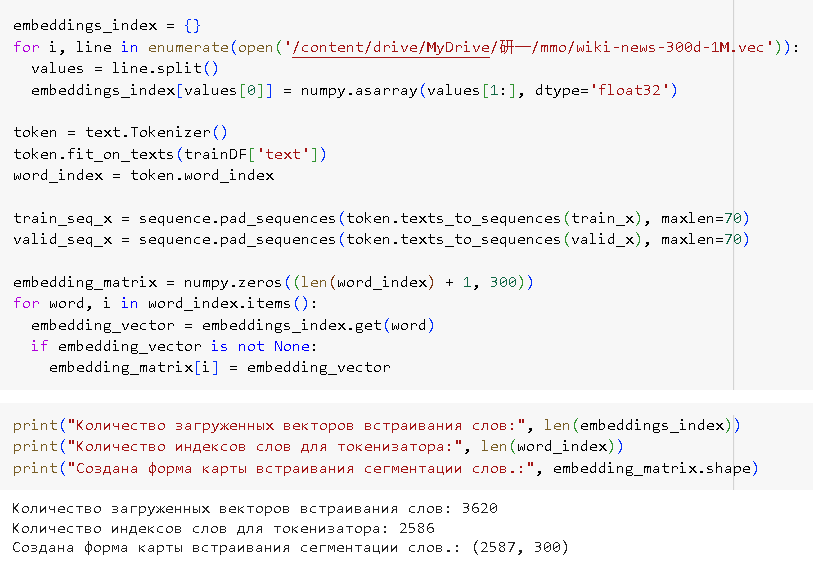


* Векторы TF-IDF как признаки

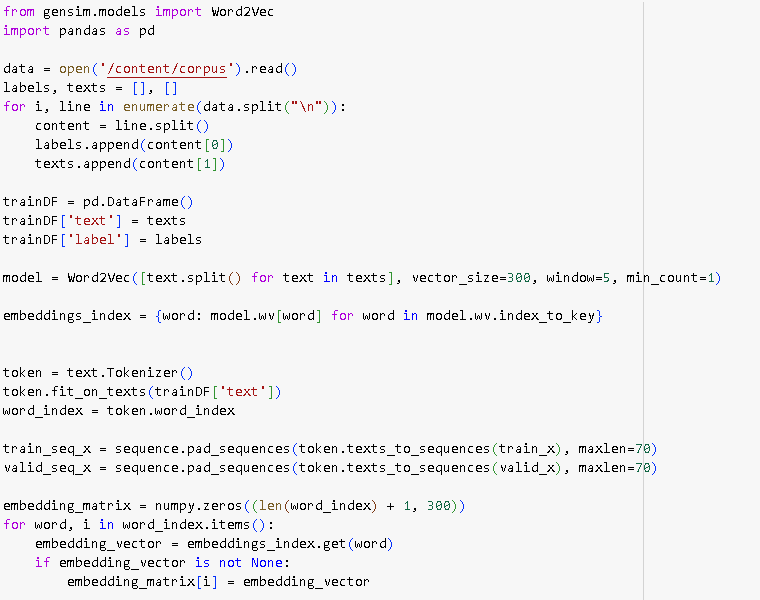


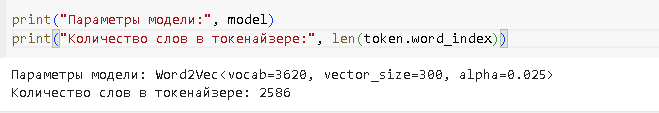
* встраивание слов

Встраивание слов — это форма использования плотных векторов для представления слов и документов. Положение слова в векторном пространстве определяется из контекста слова в тексте. Встраивание слов можно обучить с использованием самого входного корпуса или сгенерировать с использованием предварительно обученной модели встраивания слов. Модели встраивания слов включают в себя: Перчатка, FastText, Word2Vec

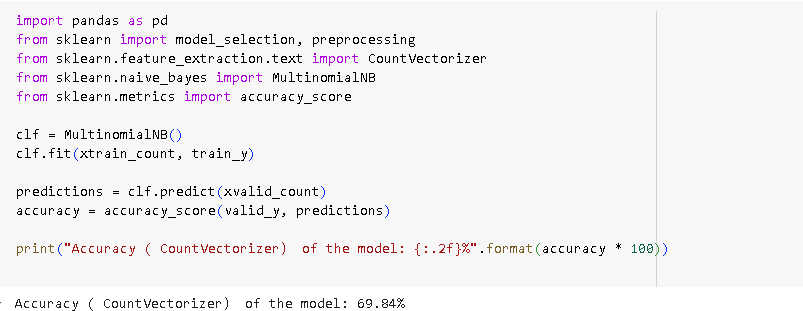


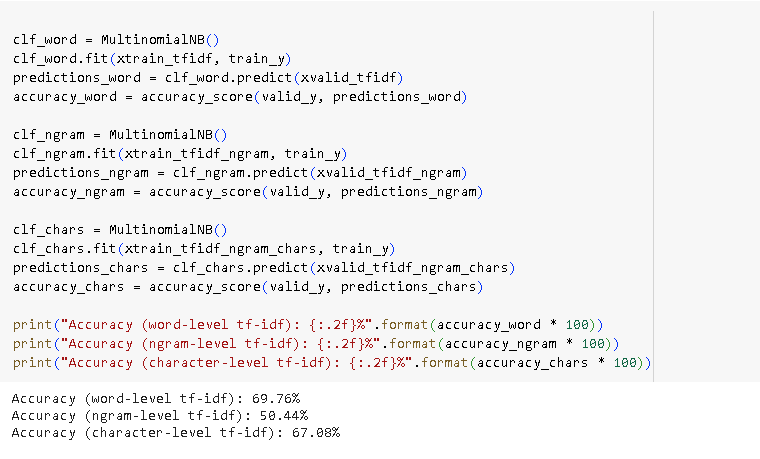
* Встраивание текста с использованием модели Word2Vec





* Используйте наивный байесовский классификатор (MultinomialNB) для обучения и прогнозирования функций, а также расчета точности модели.





# Вывод

Два часто используемых метода решения задач классификации текста основаны на CountVectorizer или TfidfVectorizer и на моделях word2vec, Glove или fastText. Методы CountVectorizer и TfidfVectorizer просты и удобны в использовании и подходят для небольших наборов данных, но они не могут уловить семантические отношения между словами. Методы, основанные на моделях word2vec, Glove или fastText, могут преобразовывать текст в плотные векторы слов, лучше улавливать семантическую и контекстную информацию слов и подходят для моделей глубокого обучения и крупномасштабных наборов данных.