Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика, искусственный интеллект и системы управления»

Кафедра «Системы обработки информации и управления»



**Рубежный контроль № 2**

**по дисциплине «Методы машинного обучения»**

Методы обработки текстов

Вариант 8

ИСПОЛНИТЕЛЬ:

студентка ИУ5-23М

Морозевич М.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

\_\_\_ "\_\_\_\_\_\_\_\_" 2024 г.

Москва, 2024

# **Вариант и задание**

Необходимо решить задачу классификации текстов на основе датасета, классификация может быть бинарной или многоклассовой.

Необходимо сформировать два варианта векторизации признаков - на основе CountVectorizer и на основе TfidfVectorizer.

В качестве классификаторов необходимо использовать два классификатора:

* LinearSVC;
* LogisticRegression.

Для каждого метода необходимо оценить качество классификации, затем сделать вывод о том, какой вариант векторизации признаков в паре с каким классификатором показал лучшее качество.

# **Описание набора данных**

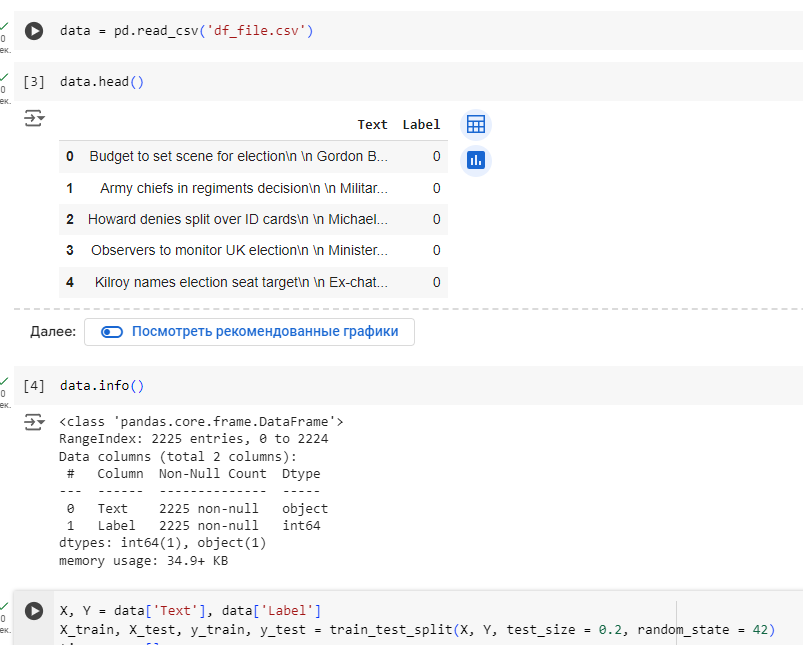
Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с использованием метода заполнения модой.

Будет использован набор данных для классификации текстовых документов, который содержит 2225 текстовых данных и пять категорий документов.

Данные содержат 2 столбца:

* Текст: Содержит различные категории текстовых данных
* Метка: Содержит метки для пяти различных категорий (0,1,2,3,4)
  1. Политика = 0
  2. Спорт = 1
  3. Технологии = 2
  4. Развлечения = 3
  5. Бизнес = 4

Добавим данные и разделим данные на обучающую и тестовую выборки.

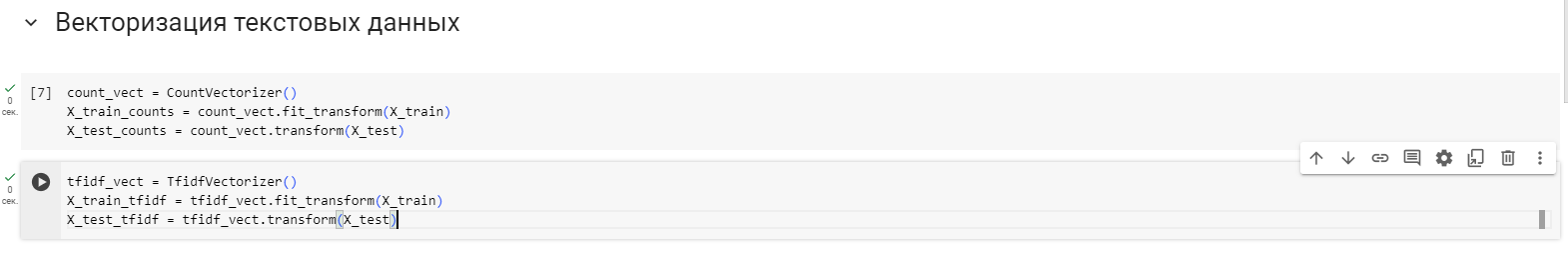


1. **Векторизация текстовых данных**

Преобразуем текстовые данные в векторы действительных чисел, которые понятны моделям машинного обучения. Она является шагом в извлечении признаков.

Используем 2 метода:

* Мешок слов: включает в себя токенизацию, создание словаря и создание вектора
* TF-IDF: числовой статистический показатель, который отражает важность слова для документа.

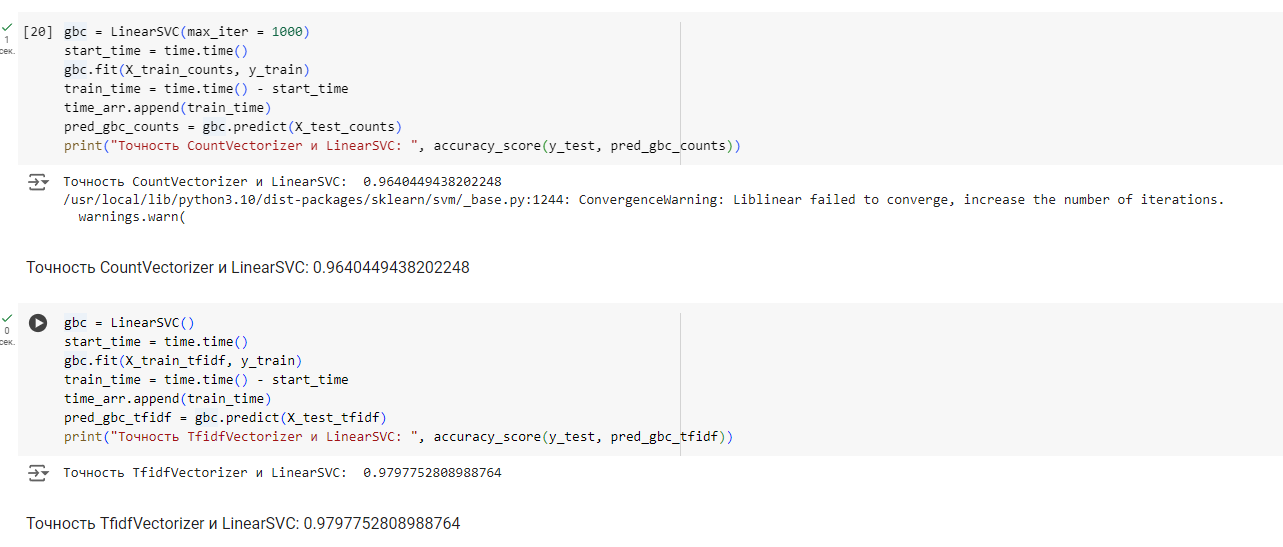


Используем их в дальнейшем при обучении различных классификаторов.

# **Обучение и сравнение классификаторов**

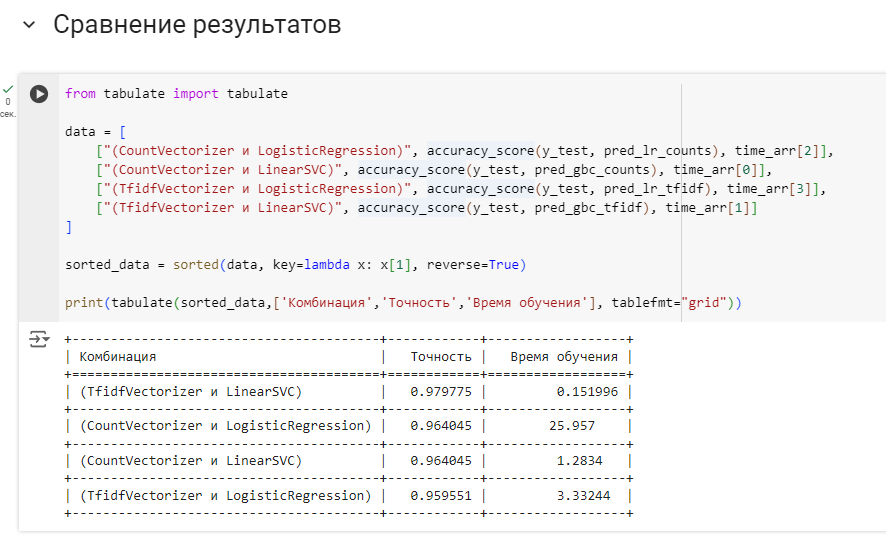
Для классификации текстов будут использованы LinearSVC и LogisticRegression.

Обучим классификаторы для данных, векторизованных каждым из способов.





Сведем результаты обучения в одну таблицу для сравнения.



Из таблицы можно отметить, что лучший результат показала комбинация «TfidfVectorizer+LinearSVC» (97,9%), а наихудший – «TfidfVectorizer+LinearSVC» (95,9%). При этом результаты обучения всех комбинаций в целом достаточно высокие, выше 95%.

С другой стороны, наименьшее время обучения продемонстрировала комбинация для «TfidfVectorizer+LinearSVC» (0,15 с), а наибольшее – «CountVectorizer+LogisticRegression» (25,96 с).

В итоге, наилучший результат у комбинации – «TfidfVectorizer+LinearSVC». Более того, оба метода с использованием CountVectorizer выполняются дольше, чем с TfidfVectorizer.