Московский государственный технический университет им. Н.Э. Баумана

Факультет «Головной учебно-исследовательский и методический центр

профессиональной реабилитации лиц с ограниченными возможностями здоровья (инвалидов)»

Кафедра «Системы обработки информации и управления»



**Рубежный контроль 2**

**по дисциплине «Методы машинного обучения в АСОИУ»**

**"** **Методы обработки текстов "**

СТУДЕНТ:

студент группы ИУ5Ц-21М

Москалик А.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

Москва, 2024

# Задание:

# Необходимо решить задачу классификации текстов на основе любого выбранного Вами датасета (кроме примера, который рассматривался в лекции). Классификация может быть бинарной или многоклассовой. Целевой признак из выбранного Вами датасета может иметь любой физический смысл, примером является задача анализа тональности текста.

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

# В качестве классификаторов необходимо использовать два классификатора по варианту для Вашей группы:

# Группа - ИУ5-21М, ИУ5И-21М, ИУ5Ц-21М

# Классификатор №1 – KneighborsClassifier

# Классификатор №2 – LogisticRegression

# 

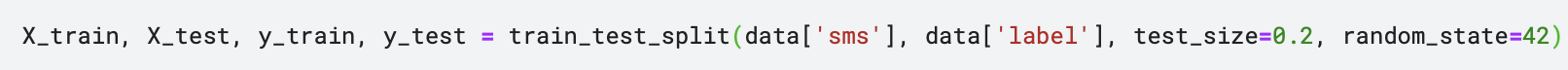
# Ход работы

# Подготовка

# Импорт библиотек и загрузка набора данных

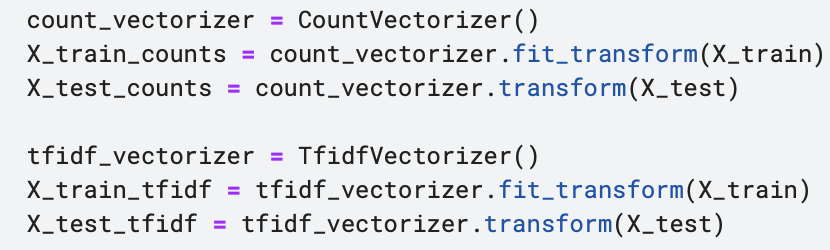


Разделение выборки на обучающую и тестовую



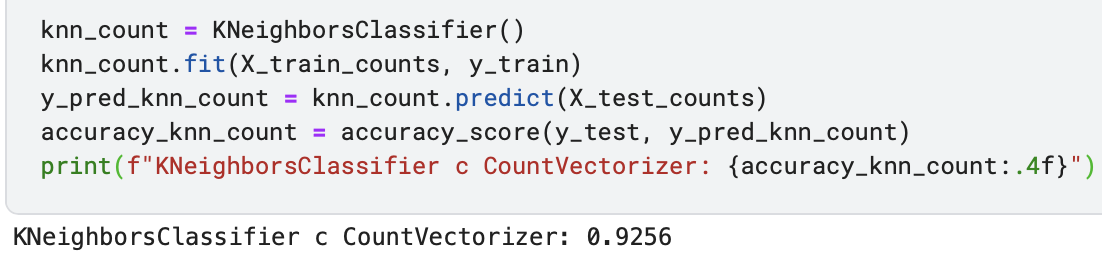
**Векторизация признаков**

В этом коде происходит процесс векторизации текстовых данных. Векторизация — это преобразование текстовых данных в числовой формат, который может быть использован моделями машинного обучения.

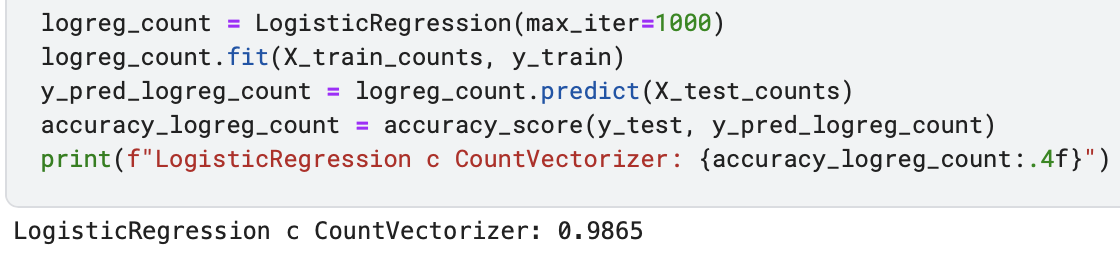


**Обучение моделей**

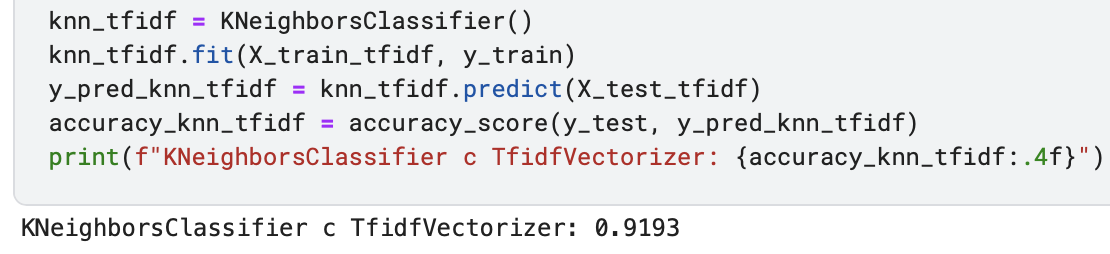
1. KNeighborsClassifier с CountVectorizer



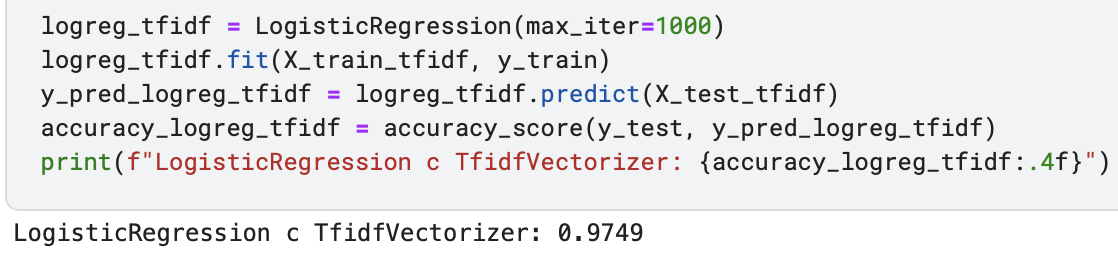
1. LogisticRegression c CountVectorizer



1. LogisticRegression c CountVectorizer

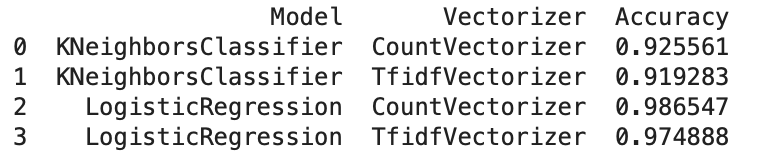


1. LogisticRegression c CountVectorizer

****

**Оценка качества классификации**



****

**Вывод:**

Наилучшие результаты показала модель LogisticRegression с использованием CountVectorizer с точностью 0.9865. Следовательно, для данной задачи классификации текстов наилучшей комбинацией является использование LogisticRegression в паре с CountVectorizer.