

Московский Государственный Технический Университет имени Н.Э. Баумана

Факультет Информатика и системы управления

**Кафедра ИУ-5 «Системы обработки информации и управления»**

**Отчёт по рубежному контролю № 2**

# По дисциплине

**«**Методы Машинного Обучения**»**

Выполнил студент Люй Чжэ

Группа ИУ5И-23М

**Москва 2024г**

# Задание

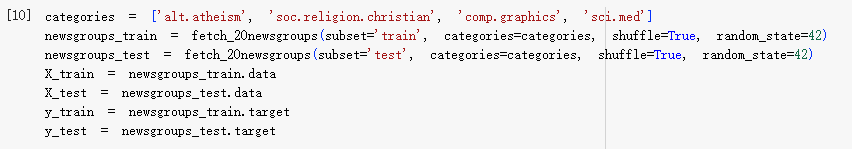
Необходимо решить задачу классификации текстов на основе любого выбранного Вами датасета. Классификация может быть бинарной или многоклассовой. Целевой признак из выбранного Вами датасета может иметь любой физический смысл, примером является задача анализа тональности текста.

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

В качестве классификаторов необходимо использовать два классификатора:"**LinearSVC**" и "**LogisticRegression**".

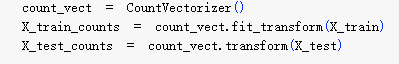
Для каждого метода необходимо оценить качество классификации. Сделайте вывод о том, какой вариант векторизации признаков в паре с каким классификатором показал лучшее качество.

# Шаг 1 Загрузка наборов данных



# Шаг 2 векторизация признаков

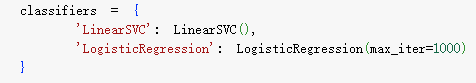
**CountVectorizer**



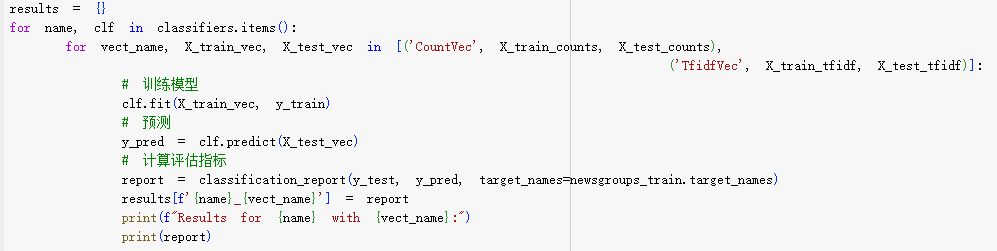
**TfidfVectorizer**

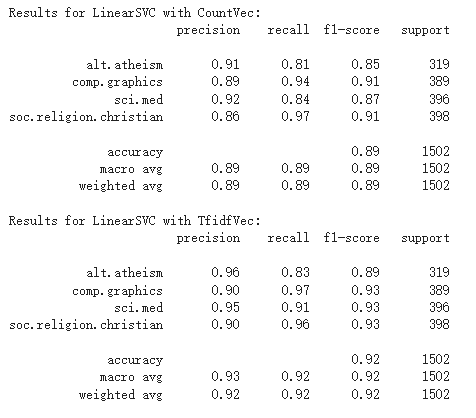


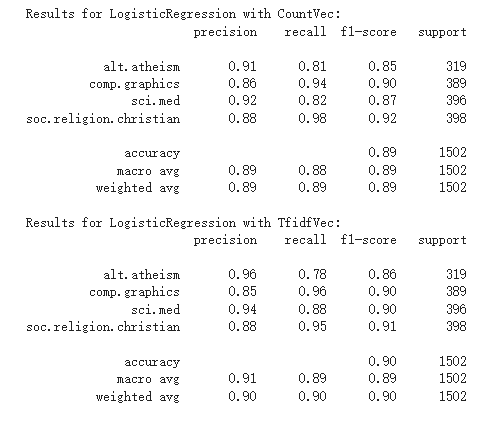
# Шаг 3 Инициализация классификатора



# Шаг 4 Испытания и оценки







# Вывод

Наилучшее качество показала связка метода вектора признаков LinearSVC с классификатором TfidfVectorizer.