|  |
| --- |
|  |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА - Московский технологический университет»**  **РТУ МИРЭА** |

Институт Информационных Технологий

Кафедра Прикладной Математики (ПМ)

**ПРАКТИЧЕСКАЯ РАБОТА № 5**

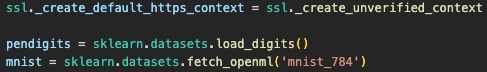
|  |  |
| --- | --- |
| Выполнил студент группы ИКБО-08-19  Борисов А.В. | (\_\_\_\_\_\_\_\_\_\_)  *подпись* |
| Принял Ассистент кафедры ПМ  Высоцкая А.А. | (\_\_\_\_\_\_\_\_\_\_)  *подпись* |
| Практическая работа выполнена | « » 2022 г. |
| «Зачтено» | « » 2022 г. |

Москва 2022

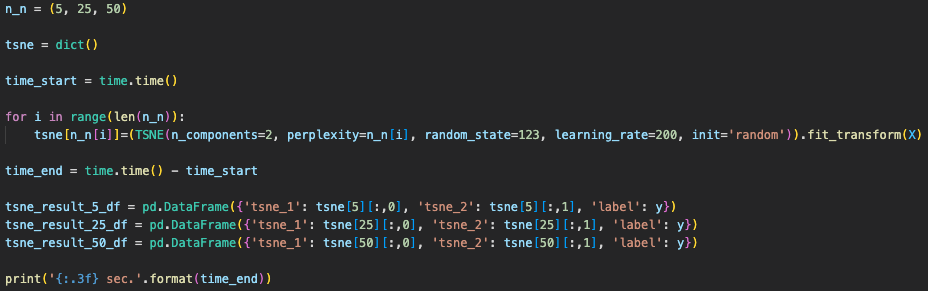
1. ***Выполнить визуализацию многомерных данных, используя t-SNE. Необходимо использовать набор данных MNIST или fashion MNIST Рассмотреть результаты визуализации для разных значений перплексии.***

Выбор набора данных

В качестве набора данных для работы был выбран “mnist\_784”



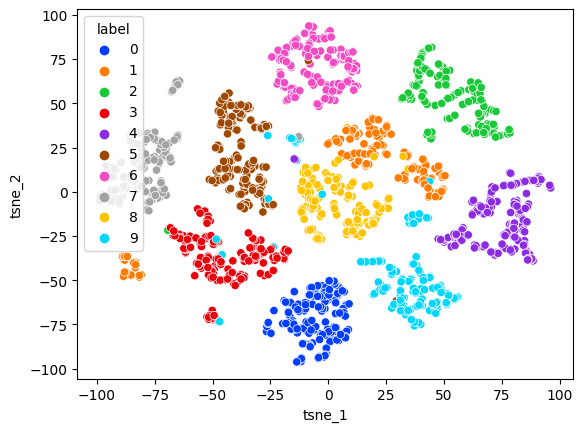
Применение t-SNE и запись результатов в дата фрейм:



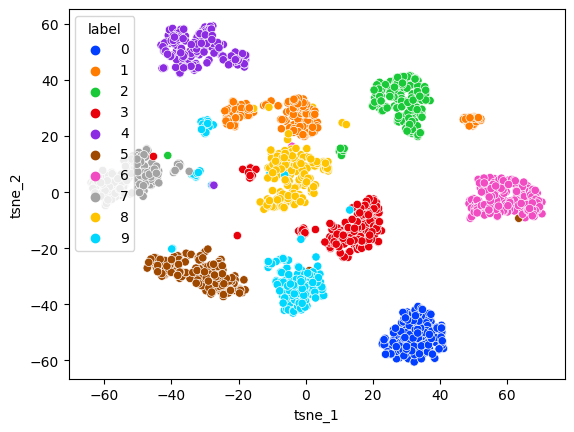
Время исполнения кода, отвечающего за t-SNE:



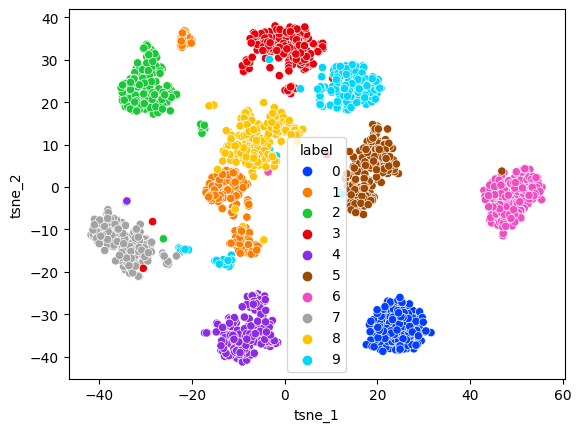
t-SNE, приплексия – 5:



t-SNE, приплексия – 25:



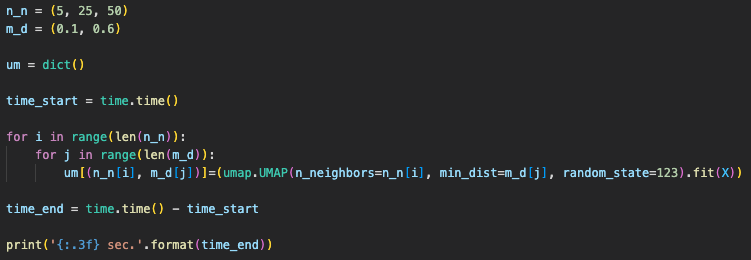
t-SNE, приплексия – 50:



Из результатов можно видеть, что при увеличении значение perplexity увеличивается плотность скопления точек, внутри кластеров, больше становится видна глобальная структура данных, однако менее заметными становятся отношения между точками внутри одного кластера.

1. ***Выполнить визуализацию многомерных данных, используя UMAP с различными параметрами n\_neighbors и min\_dist. Рассчитать время работы алгоритма с помощью библиотеки time и сравнить его с временем работы t- SNE.***

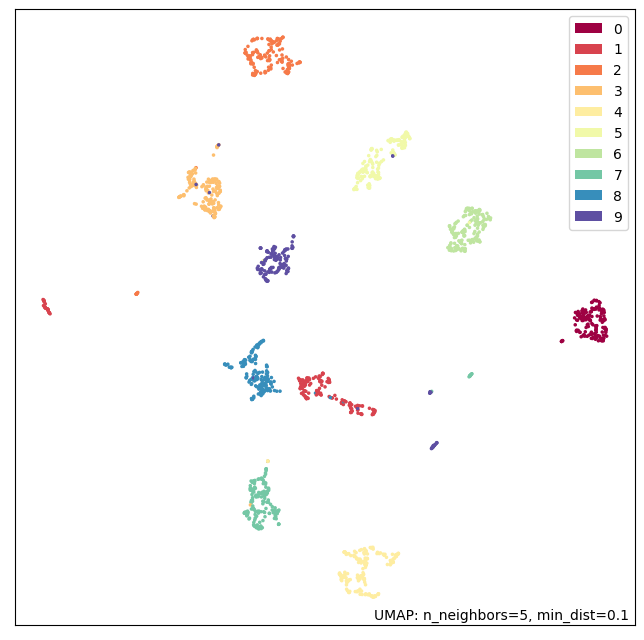
Применение UMAP:



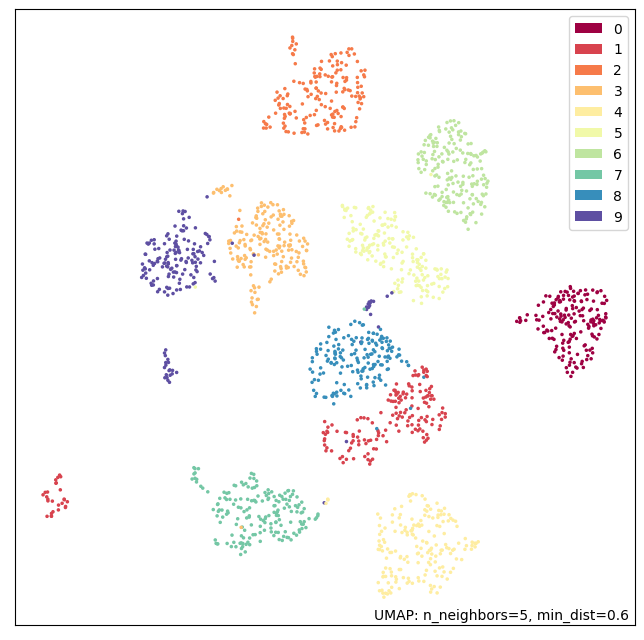
Время исполнения кода, отвечающего за UMAP:



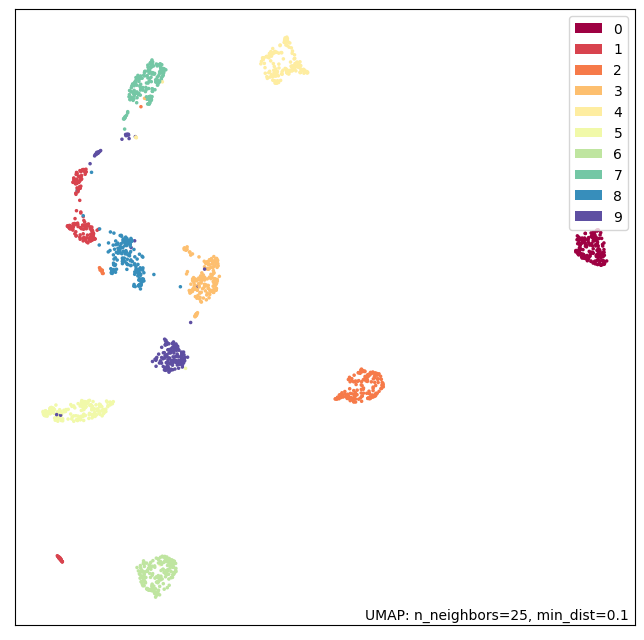
UMAP n\_neighbors = 5 и min\_dist = 0.1:



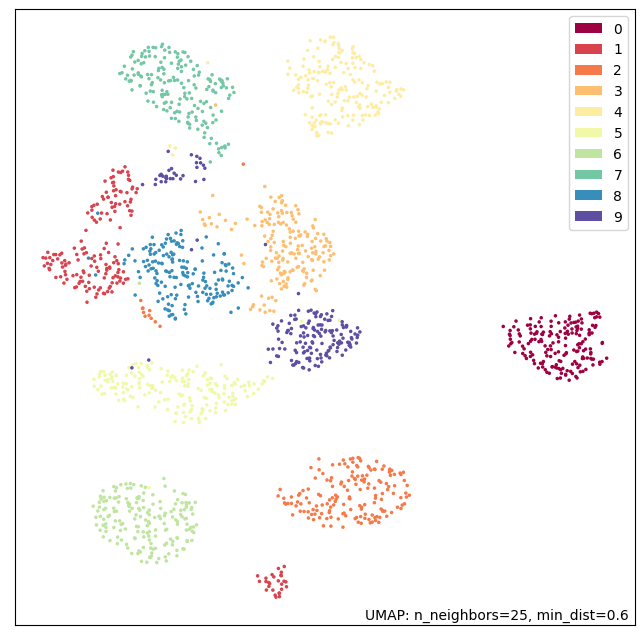
UMAP n\_neighbors = 5 и min\_dist = 0.6:



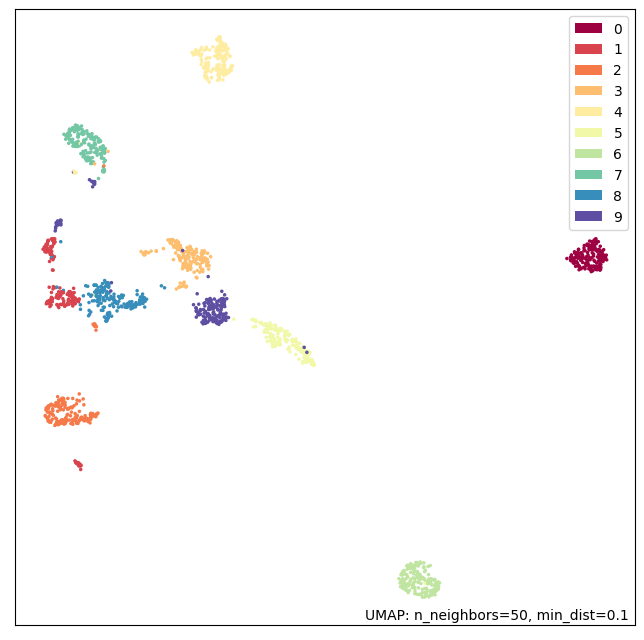
UMAP n\_neighbors = 25 и min\_dist = 0.1:



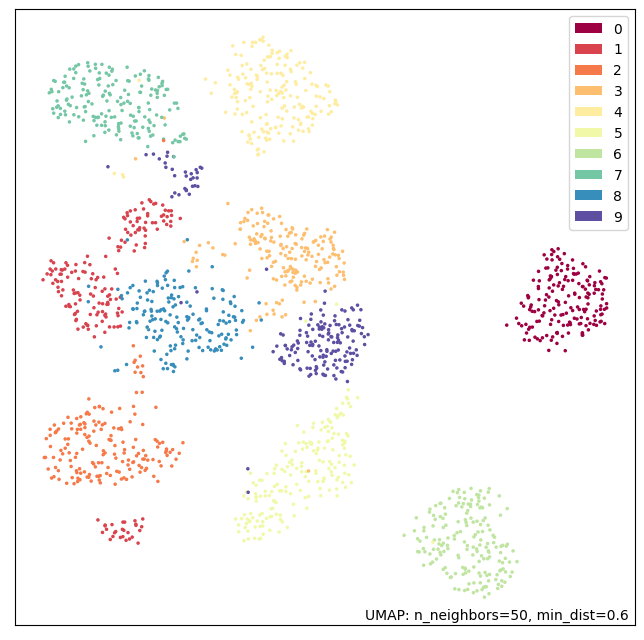
UMAP n\_neighbors = 25 и min\_dist = 0.6:



UMAP n\_neighbors = 50 и min\_dist = 0.1:



UMAP n\_neighbors = 50 и min\_dist = 0.6:



При изменении значения n\_neighbors происходит “уплотнение” структуры данных внутри кластера. При низких значениях алгоритм ограничивается малыми окрестностями вокруг точек, и наоборот при больших значениях. Грубо говоря, опять же можно выбирать что важнее в новом представлении данных: сохранить локальную или глобальную структуру данных.

Изменение параметра min\_dist ведет к изменению минимального расстояния, на котором могут находиться точки в новом пространстве. При низких значениях будет больше уделяться внимание общей структуре кластеров, при больших – структура кластеров, как единого целого.

1. ***Вывод.***

Оба способа уменьшения размерности многомерных данных достаточно эффективны и хорошо показали себя в работе. Они имеют отличия по скорости обучения (в моем примере t-SNE - +- 20 секунд, UMAP - +- 30 секунд) на все операции, однако стоит заметить что у последнего изменяется большее количество параметров и, соответственно, если считать по изменяя по одному параметру, то время UMAP будет быстрее. Так же он лучше показывает себя при работе с большими объемами информации, что может соответственно повлиять на итоговый выбор, но они оба достаточно хорошо справляются с поставленными перед ними задачами.