Данные:

Данные представлены четырьмя образцами (sample_1, sample_2, …). Для каждого образца есть набор измерений определенных характеристик (маркеров) по каждой из клеток. Данные измерения представлены tsv файлами в папке дата. Строки соответствуют клеткам, столбцы – маркерам. Также, каждому образцу соответствует файл разметки, представленный tsv файлом в папке labels. Данный файл содержит названия типов клеток, которые будем считать ground truth.

Задачи:

1. Построить и обучить произвольный классификатор типов клеток по всем образцам. Провести тестирование на тех же образцах, оценить качество распознавания. Качество распознавания оценить с помощью confusion matrix и f1-score.
2. Провести кросс-валидацию по файлам: взять первые три образца, обучить на них модель, тестировать на оставшемся образце (confusion matrix + f1-score). Далее взять следующие три образца, обучить на них и тестировать на оставшемся. На выходе должно быть четыре матрицы неточностей и четыре характеристики f1-score для каждого образца.
3. Выбрать образец с наивысшим f1-score. Провести снижение размерности данных до двух методом TSNE. Визуализировать результат классификации модели в двухмерном представлении.
4. Оценить качество кластеризации в двухмерном пространстве любым из существующих методов (например, Rand Index, Adjusted Rand Index, Mutual Information).