Лабораторная работа 3

Базовые алгоритмы обучения без учителя и работа с текстовыми данными

1. *Понижение размерности и визуализация данных*

Примените методы снижения размерности sklearn.decomposition.PCA и sklearn.manifold.TSNE для визуализации данных с которым вы работали в лабораторной № 2 (снижая размерность до двух). Отобразите полученные результаты

2. Кластерный анализ

1) С помощью алгоритма k-means сделайте квантование изображения (удаление визуально избыточной информации) с глубиной 64, 32, 16 и 8 уровней. Изображение выбираете произвольно.
Пример: https://scikit-learn.org/stable/auto_examples/cluster/plot_color_quantization.html
2) Сгенерируйте набор синтетических данных (точек на плоскости), например при помощи sklearn.datasets.make_blobs, число центров N (от 3 до 5) выберите произвольно. Постройте силуэтные графики для Kmeans (для числа кластеров N-1, N, N+1), объясните результаты
https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
3) Сгенерируйте набор синтетических данных в виде смеси двух гауссиан, для этого воспользуйтесь функцией: https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.random.multivariate_normal.html
(примените ее дважды с разными mean и cov), результат объедините в одно множество. Разделите смесь с помощью EM алгоритма(sklearn.mixture.GaussianMixture), обратите внимание на параметр *covariance_type*. С помощью атрибутов weights_ и covariances_ восстановите их значения (сравните с оригинальными). Визуализируйте результат.

3. *Обработка текстовых данных*

Загрузите набор текстовых данных, снабженных метками классов. Проведите предобработку данных (удалите стоп-слова, пунктуацию, проведите нормализацию), постройте визуализацию наиболее частых слов или n-gram в каждом классе (**wordcloud**), извлеките признаки (например sklearn.feature_extraction.text .TfidfVectorizer; или sklearn.decomposition.TruncatedSVD)
Проведите классификацию текстовых данных, сделайте оценку качества
Текстовые данные для анализа можно брать здесь:
https://lionbridge.ai/datasets/the-best-25-datasets-for-natural-language-processing/
или из любого другого источника по вашему выбору
(в случае данных с множеством классов достаточно взять 2-3 класса)