Лабораторна робота 3
Базові алгоритми навчання без вчителя та обробка текстових даних

1. Зниження розмірності і візуалізація даних
Застосуйте методи зниження розмірності sklearn.decomposition.PCA і sklearn.manifold.TSNE для візуалізації даних, з якими ви працювали в лабораторній № 2 (знижуючи розмірність до двох). Візуалізуйте результат.

2. Кластерний аналіз
1) За допомогою алгоритму k-means зробіть квантування зображення (видалення візуально надлишкової інформації) з глибиною 64, 32, 16 та 8 рівнів для будь-якого обраного самостійно зображення.
Приклад: https://scikit-learn.org/stable/auto_examples/cluster/plot_color_quantization.html

2) Згенеруйте набір синтетичних даних у вигляді суміші двох гаусіан за допомогою функції:
https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.random.multivariate_normal.html
(застосуйте її двічі з різними mean і cov, результат об'єднайте)
Розділіть суміш за допомогою EM алгоритму (sklearn.mixture.GaussianMixture), зверніть увагу на параметр *covariance_type*. За допомогою атрибутів weights_ і covariances_ відновіть їхні значення, порівняйте з оригінальними. Візуалізуйте результат.

3. Обробка текстових даних
Завантажте набір текстових даних (з мітками класів). Проведіть передобробку даних (видаліть стоп-слова, пунктуацію), за допомогою **wordcloud** зробіть візуалізацію найбільш поширених слів або n-gram у кожному класі. Векторизуйте тексти (наприклад за допомогою sklearn.feature_extraction.text .TfidfVectorizer). Проведіть класифікацію текстових даних, зробіть оцінку якості. Застосуйте алгоритм LDA до кожного класу, визначте декілька тематик (sklearn.decomposition.LatentDirichletAllocation)

Текстові дані для аналізу можна обирати тут:
https://analyticsindiamag.com/10-open-source-datasets-for-text-classification/
https://www.ics.uci.edu/~smyth/courses/cs175/text_data_sets.html
https://medium.com/@ODSC/20-open-datasets-for-natural-language-processing-538fbfaf8e38
або з будь-якого іншого джерела за вашим вибором
(в разі великої кількості класів достатньо залишити 2-3)

Датасет IMDB Movie Review Sentiment classification не обирайте

Як звіт – робочий код в Jupyter notebook заливаєте на свій репозиторій на https://github.com/.
Лінк відправляєте на пошту natsakh-ipt@lll.kpi.ua
Максимальний бал – 10, 6 за роботу + 4 захист.
Deadline 20.11.21, після цього терміну максимальний бал зменшується на 1 кожні 2 тижні

Нагадую: однакові або дуже схожі роботи прийматися не будуть.