Задание 1. Сжатие изображений. Одно из самых наглядных применений сингулярного разложения.

- Выберите какое-нибудь изображение, преобразуйте его к оттенкам серого и представьте в виде матрицы.
- Выполните SVD-разложение получившейся матрицы.
- Найдите «укороченное» SVD-разложение этой матрицы, оставив только k первых (наибольших) сингулярных чисел и соответствующих им векторов. Выведите полученное изображение. Сделайте это для не менее 9 различных значений k.
- Сделайте выводы о влиянии параметра k на качество вашего изображения и степень сжатия (какую долю от исходной информации необходимо хранить в памяти). При каких значениях k картинка имеет приемлемое качество? При каких значениях качество уже не очень, но картинка всё ещё различима? Для каких k картинка становится совершенно непонятной? Оцените численно степень сжатия изображения при каждом из рассмотренных вами значений параметра k.

Задание 2. Latent Semantic Analysis. Перед выполнением задания ознакомьтесь с материалом по теме LSA.

- Выберите один из текстовых файлов, доступных по ссылке. Каждый файл представляет собой массив строк (документов), с которыми вам предстоит работать.
- Удалите знаки препинания и разделите текст на отдельные слова.
- Проведите лемматизацию текста, приведя все слова к их начальной форме. Для этого можно использовать библиотеку `pymorphy2` языка Python.
- Исключите из текста «шумовые» слова, не несущие значительной смысловой нагрузки. Такими являются *предлоги*, *союзы*, *местоимения*, а также часто встречающиеся слова. Для этого предлагается использовать библиотеку `nltk`.
- Составьте словарь из оставшихся слов и создайте терм-документную матрицу, описывающую частоту употребления терминов в коллекции документов.
- Выполните SVD-разложение получившейся матрицы.
- Постарайтесь ответить на вопросы: Какой смысл имеют столбцы/строки матриц U и V? Какой смысл имеют сингулярные числа матрицы Σ?
- Проанализируйте первые два сингулярных числа и соответствующие им правые и левые сингулярные векторы. Используя топ-5 слов с наибольшим отклонением в соответствующих векторах, восстановите основные две темы в тексте. Определите, каким документам эти темы соответствуют в наибольшей степени. В дополнение создайте диаграмму «облако слов» из словаря, где размер слова будет отражать его значимость для темы.
- Сделайте выводы по проделанной работе.