Quiz 6: NLP и автокодировщики

– Не ходи туда, там тебя ждут неприятности.
– Ну как же не ходить они же ждут...

Котенок по имени Гав (1976)

Решите все задания. Все ответы должны быть обоснованы. Решения должны быть прописаны для каждого пункта. Рисунки должны быть чёткими и понятными. Все линии должны быть подписаны. Списывание карается обнулением работы. **При решении работы можно пользоваться чем угодно.** Удачи!

[5] Задание 1

Мы пытаемся сжать картинку v_i с помощью метода главных компонент. Давайте запишем эту задачу в виде автокодировщика.

- Запишите формулы для энкодера и декодера. Из каких пространств в какие они бьют как функции?
- Выпишите функцию потерь, которая будет использоваться при решении задачи для обучения.

[3] Задание 2

Объясните что такое негативное сэмплирование (negative sampling) и для чего оно нужно при обучении w2v.

[2] Задание 3

Томаш обучает w2v для английского языка на корпусе новостей с помощью метода skip-gram. Сколько параметров ему надо будет оценить при обучении модели, если он собирается оставить в словаре 100000 самых частотных токенов?

[1] Задание 4

Мы хотим, чтобы эмбеддинги рассуждали также, как это делают люди. Давайте, наоборот, попробуем рассуждать как нейросети, обучившиеся на каком-то корпусе текстов и словившие странные артефакты. Предположите, как решаются следующие уравнения и кратко поясните почему вы так думаете.

а. ночь - темнота + свет = ?

б. сосиска - маленькая + большая = ?

в. python - язык + алкоголь = ?

г. цб - резервы + проблемы = ?