Анализ отзывов к фильмам.

Задача: Бинарная классификация отзывов к фильмам. Имеется обучающая выборка с размеченными отзывами. У нас есть только текстовые файлы, поэтому нужно подготовить матрицу, с которой будет работать классификатор, будем использовать метод “Bag of words”. Также мы хотим наглядно увидеть, какие слова имели наибольший вес при анализе.

В работе используется датасет отсюда, для тренировки есть 25000 отзывов, которые пополам разделены на положительные и отрицательные, так же и для тестирования.

Для того чтобы работать с текстовыми данными их нужно закодировать, для этого будем использовать CountVectorizer.
Принцип работы CountVectorizer: сначала он выбирает уникальные слова из всех документов, длина кода равна количеству уникальных слов, далее номера элементов будут соответствовать, количеству раз встречи данного ключа с данным номером в строке.
Пример:

Значение
раз два три
три четыре два два
раз раз раз четыре

Уникальные ключи: [раз, два, три, четыре]
раз два три -> [1, 1, 1, 0]
три четыре два два -> [0, 2, 1, 1]
раз раз раз четыре -> [3, 0, 0, 1] Это и есть итог кодирования.

В качестве **классификатора** будем использовать логистическую регрессию.

Логистическая регрессия это частный случай линейного классификатора, который признаковое пространство с помощью гиперплоскости делит на два полупространства, в каждом из которых содержатся элементы одного признака.
Самый простой линейный классификатор:

$$a(\vec{x}) = sign(\vec{w}^T x),$$

где

- $\vec{x}$ – вектор признаков примера (вместе с единицей);
- $\vec{w}$ – веса в линейной модели (вместе со смещением w_0);
- $sign(\bullet)$ – функция "сигнум", возвращающая знак своего аргумента;
- $a(\vec{x})$ – ответ классификатора на примере $\vec{x}$.

Логистическая регрессия является частным случаем линейного классификатора, но она обладает хорошим "умением" – прогнозировать вероятность p_+ отнесения примера $\overrightarrow{x_i}$ к классу "+":

$$p_+ = P\left(y_i = 1 \mid \overrightarrow{x_i}, \vec{w}\right)$$

Логистическая регрессия не просто дает ответ положительный отзыв или отрицательный, а именно выводит вероятность отнесения к определенному классу.

У нас есть линейный прогноз с помощью МНК: $b(\vec{x}) = \vec{w}^T \vec{x} \in \mathbb{R}.$
Для того чтобы преобразовать полученное значение в вероятность, пределы которой [0;1] нужна функция $f : \mathbb{R} \to [0, 1].$, в модели логистической регрессии для этого берется

$$\sigma(z) = \frac{1}{1 + \exp^{-z}}$$

Посмотрим, как логистическая регрессия будет делать прогноз $p_+ = P\left(y_i = 1 \mid \overrightarrow{x_i}, \vec{w}\right)$ (пока считаем, что веса $\vec{w}$ мы как-то получили (т.е. обучили модель), далее разберемся, как именно).

- **Шаг 1.** Вычислить значение $w_0 + w_1 x_1 + w_2 x_2 + \ldots = \vec{w}^T \vec{x}$. (уравнение $\vec{w}^T \vec{x} = 0$ задает гиперплоскость, разделяющую примеры на 2 класса);
- **Шаг 2.** Вычислить логарифм отношения шансов: $\log(OR_+) = \vec{w}^T \vec{x}$.
- **Шаг 3.** Имея прогноз шансов на отнесение к классу "+" – OR_+, вычислить p_+ с помощью простой зависимости:

$$p_+ = \frac{OR_+}{1 + OR_+} = \frac{\exp^{\vec{w}^T \vec{x}}}{1 + \exp^{\vec{w}^T \vec{x}}} = \frac{1}{1 + \exp^{-\vec{w}^T \vec{x}}} = \sigma(\vec{w}^T \vec{x})$$

Итак, логистическая регрессия прогнозирует вероятность отнесения примера к классу "+" (при условии, что мы знаем его признаки и веса модели) как сигмоид-преобразование линейной комбинации вектора весов модели и вектора признаков примера:

$$p_+(x_i) = P\left(y_i = 1 \mid \overrightarrow{x_i}, \vec{w}\right) = \sigma(\vec{w}^T \overrightarrow{x_i}).$$

Далее из принципа максимального правдоподобия получается оптимизационная задача, которую решает логистическая регрессия.

В библиотеке sklearn существует реализованный метод LogisticRegression, позволяющий использовать логистическую регрессию для своих задач.

Для анализа “важных” слов будем использовать математические бибилиотеки numpy и matplotlib для расшифровки важных признаков и построения графика, на котором будет видно важность слов для положительных отзывов и отрицательных.