1. Insertion cost 1, deletion cost 1, substitution cost 2.

drive
brief
s di
Edit distance is 4.

drive
divers
 d ii
Edit distance is 3.

2. Cross-entropy loss function: Cross-entropy, also known as logarithmic loss or log loss, is a popular loss function used in machine learning to measure the performance of a classification model. Namely, it measures the difference between the discovered probability distribution of a classification model and the predicted values.

Sigmoid function: The sigmoid function is called a squashing function as its domain is the set of all real numbers, and its range is (0, 1). The sigmoid function is used as an activation function in neural networks. When the activation function for a neuron is a sigmoid function it is a guarantee that the output of this unit will always be between 0 and 1. Also, as the sigmoid is a non-linear function, the output of this unit would be a non-linear function of the weighted sum of inputs.

Gradient-descent: Gradient Descent is an optimization algorithm for finding a local minimum of a differentiable function. Gradient descent in machine learning is simply used to find the values of a function's parameters (coefficients) that minimize a cost function as far as possible.

3. tf-idf vs Word2Vec: tf-idf

- A common baseline model
- **Sparse** vectors
- Words are represented by (a simple function of) the **counts** of nearby words

Word2vec

- **Dense** vectors
- Representation is created by training a classifier to **predict** whether a word is likely to appear nearby

4. We can represent word meaning using embedding vectors. To find the similarity between two vectors we can use cosine similarity which is based on the dot product of two vectors. This raw dot product, however, has a problem as a similarity metric: it favors long vectors. The dot product is higher if a vector is longer, with higher values in each dimension. More frequent words have longer vectors since they tend to co-occur with more words and have higher co-occurrence values with each of them. The raw dot product thus will be higher for frequent words. But this is a problem; we'd like a similarity metric that tells us how similar two words are regardless of their frequency. We modify the dot product to normalize for the vector length by dividing the dot product by the lengths of each of the two vectors. This normalized dot product turns out to be the same as the cosine of the angle between the two vectors.