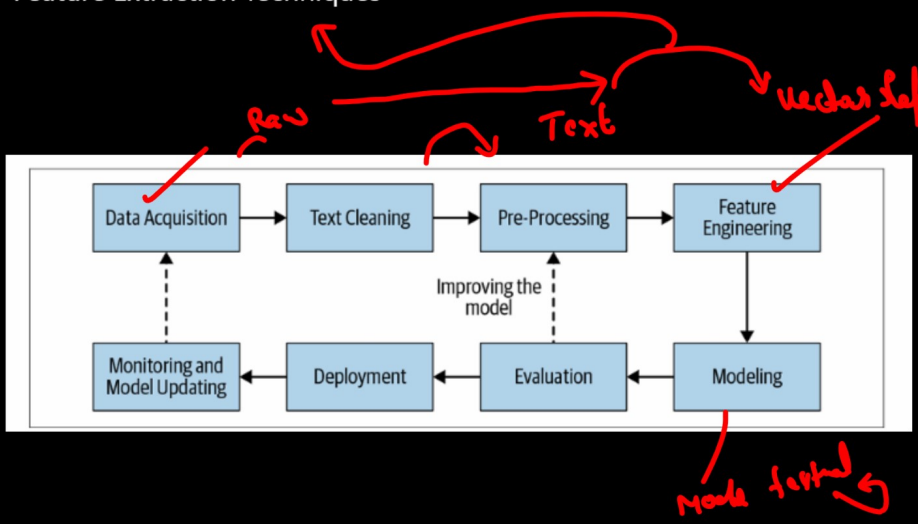
Feature Extraction Techniques



① One HOT encoding

[One hot encoding is a method used to represent text (words / tokens) as binary vectors in NLP]

Each unique word in the dataset is represented as a vector, where all elements are 0 except for single element that corresponds to that words which is set to 1

Example dataset:

1. "cat eats fish"
2. "dog eats fish"
3. "dog chases cat"

① create a vocabulary:

["cat", "eats", "fish", "dog", chases"]

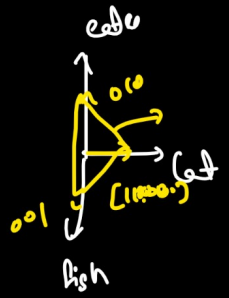
unique words (vocabulary)

② Assign index to each word

cat → 0
eats → 1
fish → 2
dog → 3
chases → 4

③ create one hot vectors

cat eats fish dog ch



"Cat cats fish"

[[1 0 0 0 0][0 1 0 0 0][0 0 1 0 0]]

"dog cats fish"

[[0 0 0 1 0][0 1 0 0 0][0 0 1 0 0]]

"dog chase cat"

[[0 0 0 1 0][0 0 0 0 1][1 0 0 0 0]]

[1 0 0 0 0]
[0 1 0 0 0]
[0 0 1 0 0]
[0 0 0 1 0]
[0 0 0 0 1]

Advantages

1. Easy to implement with python

Disadvantages

1. Sparse matrix → overfitting
2. issues with size
   RC Algorithms → fixed size input
3. out of vocabulary
4. No semantic meaning is getting captured