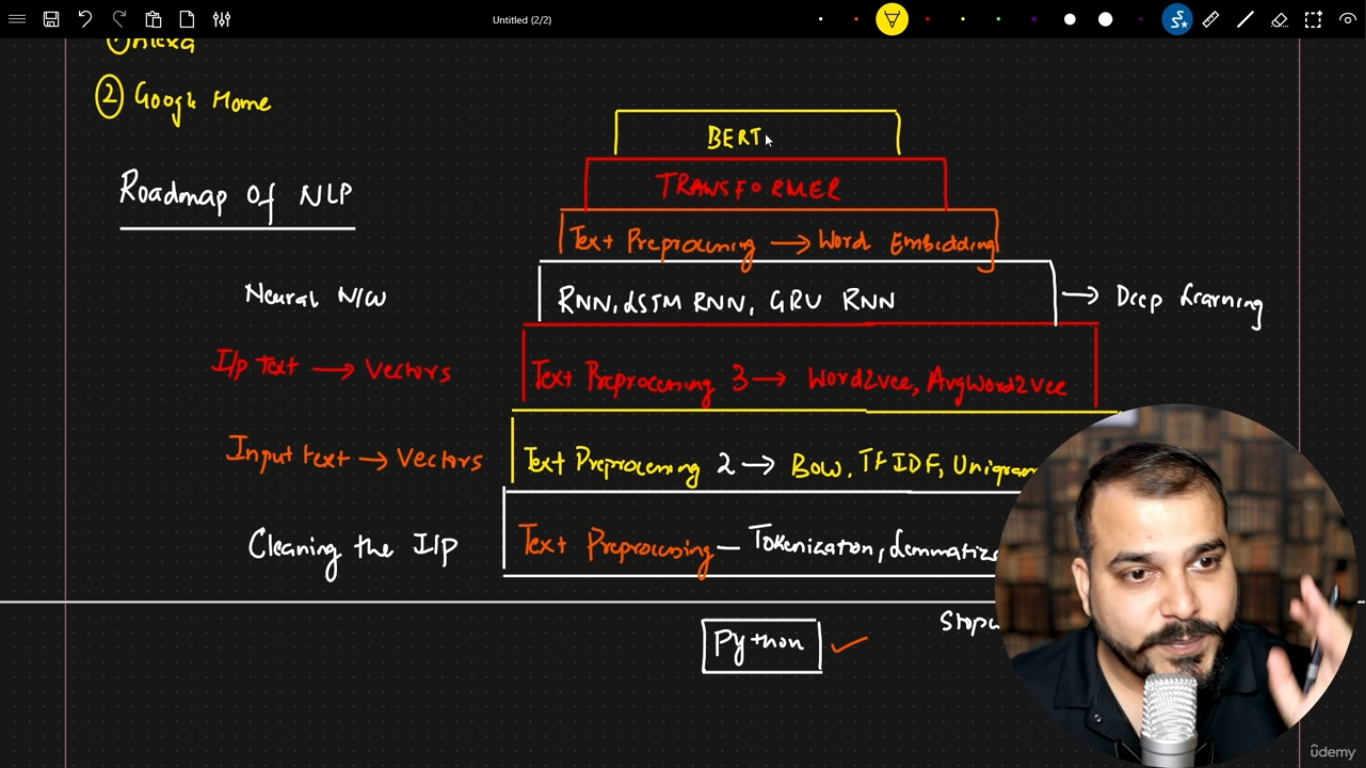
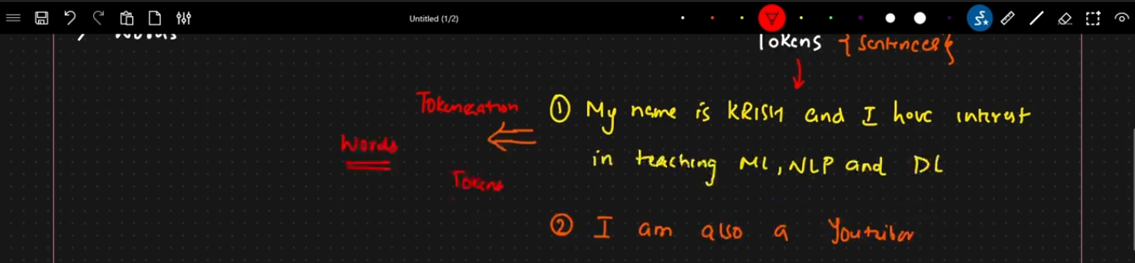
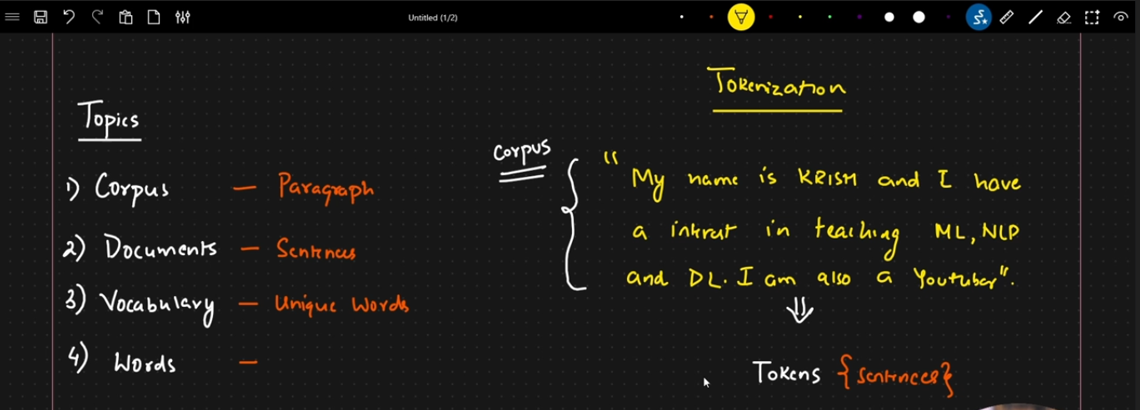
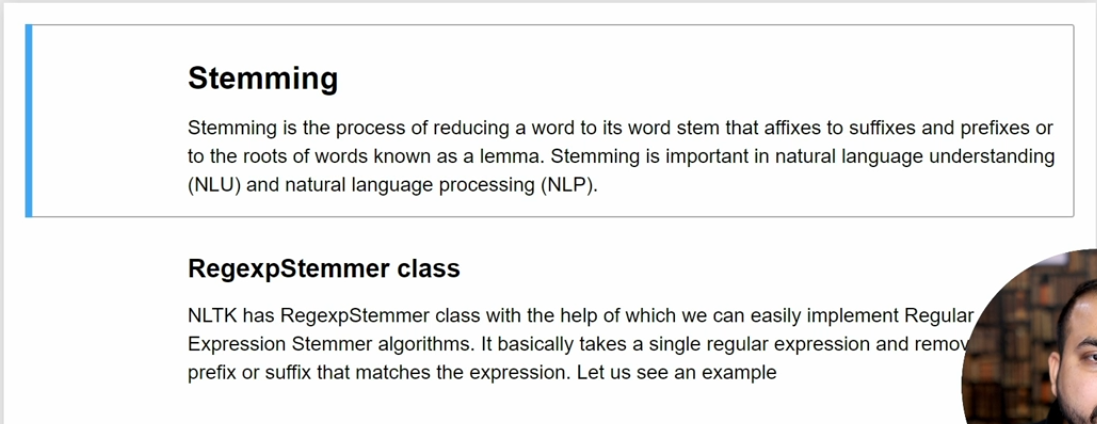
Roadmap.

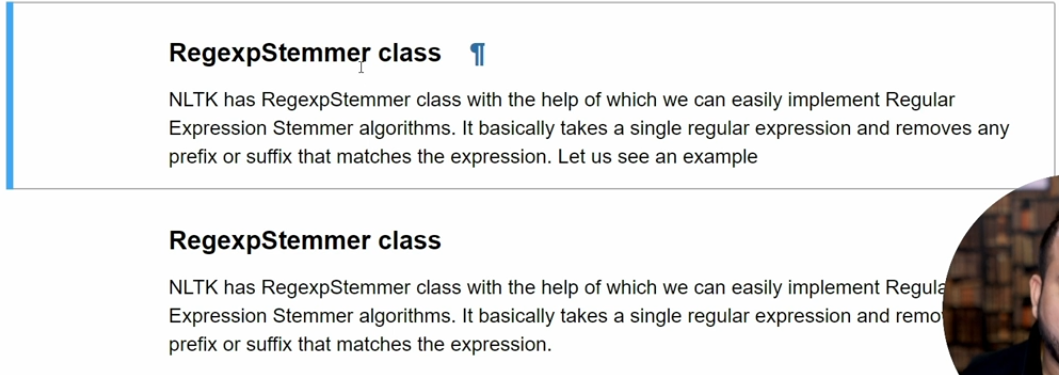


Tokenization.

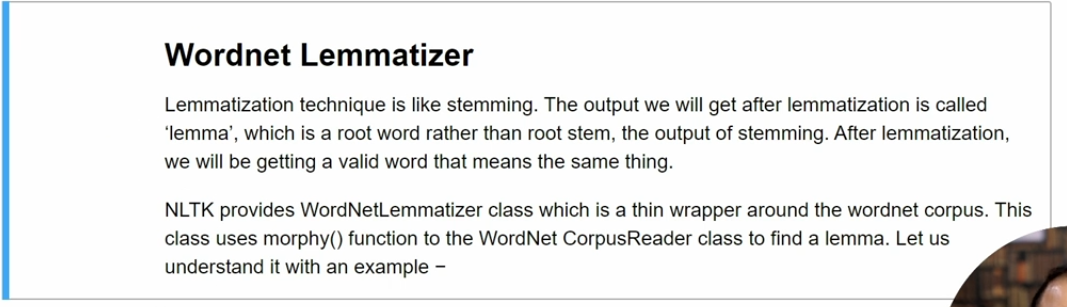


Stemming and its types.

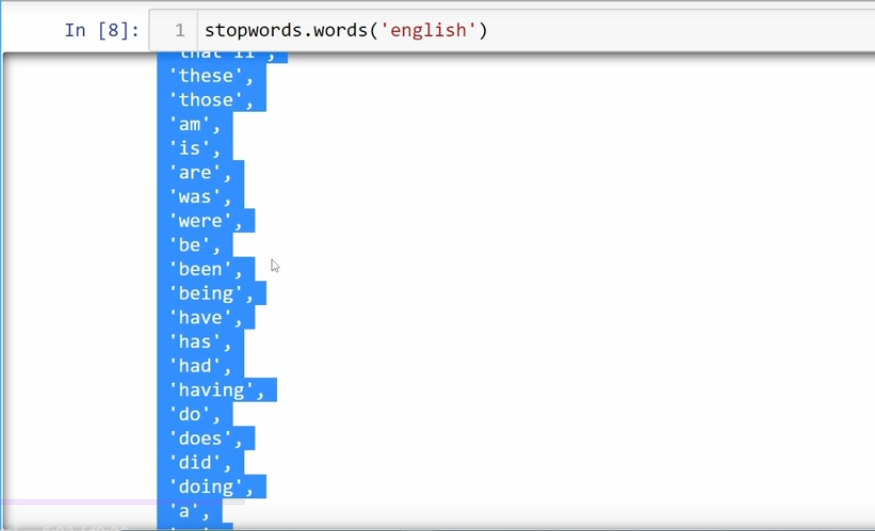




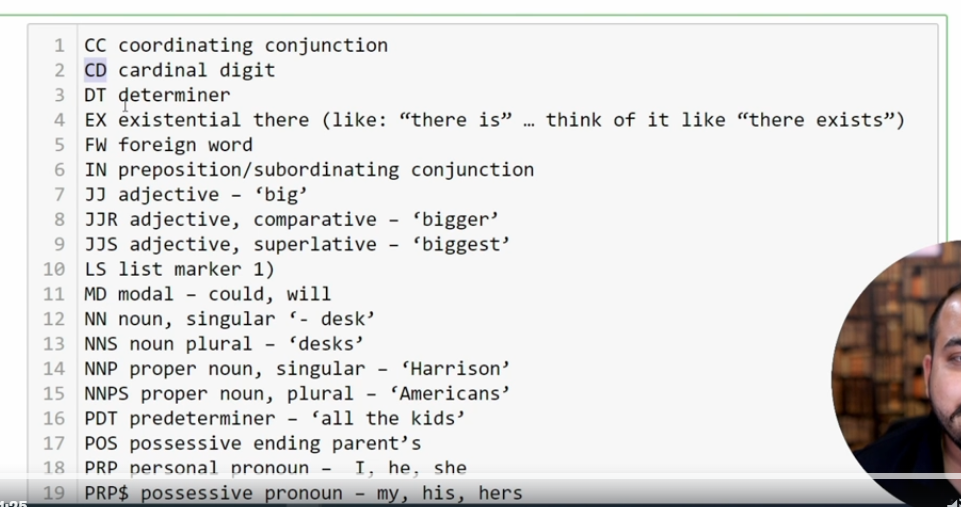
Lemmatization.



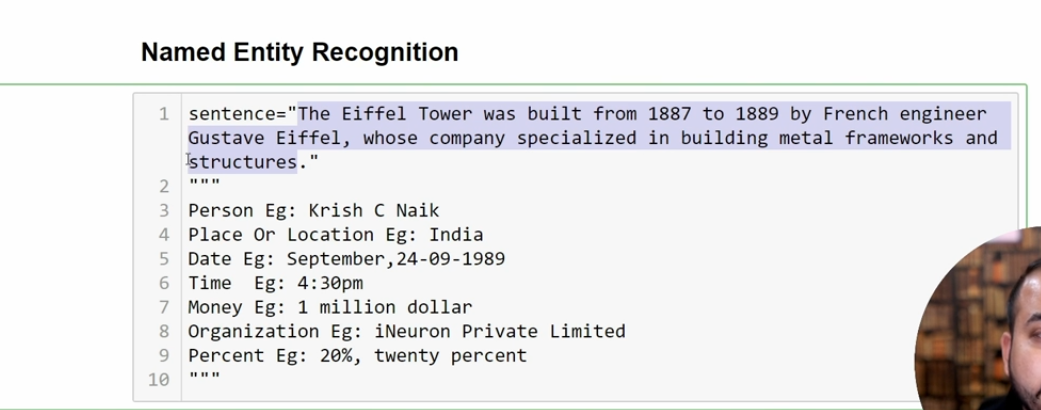
Stopwords.



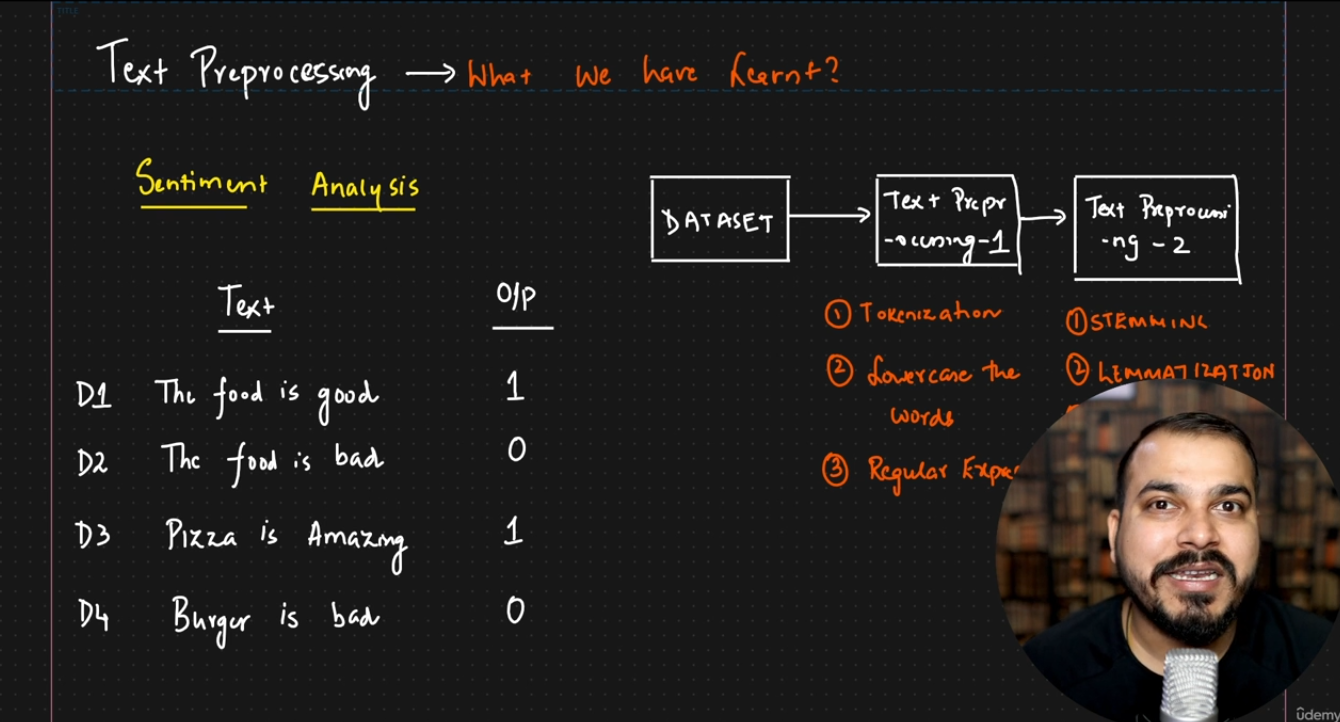
POS Tagging.



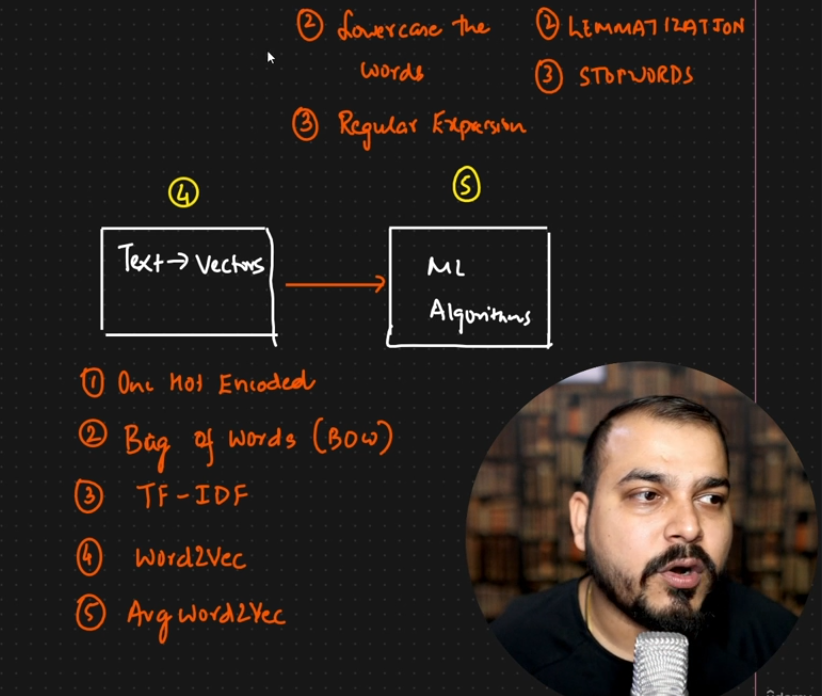
NER.



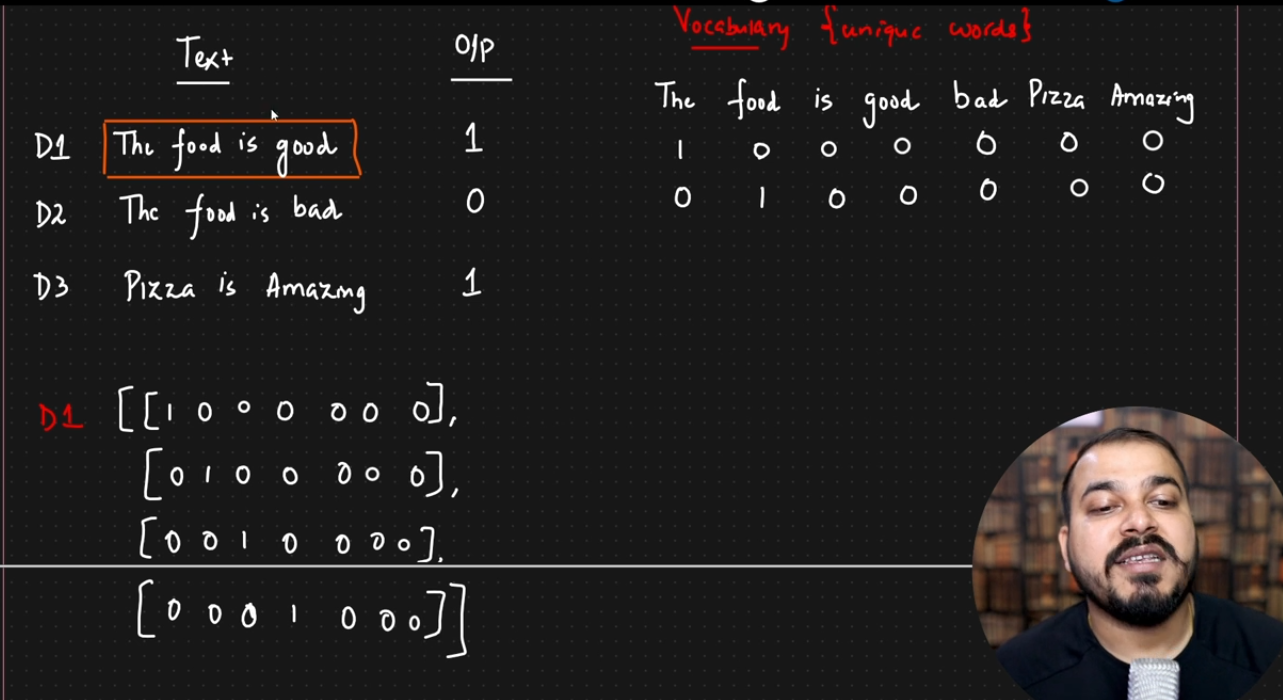
What we know till now.



What’s next.



One hot encoding



4x7 size of D1, D2. 3x7 size of D3.

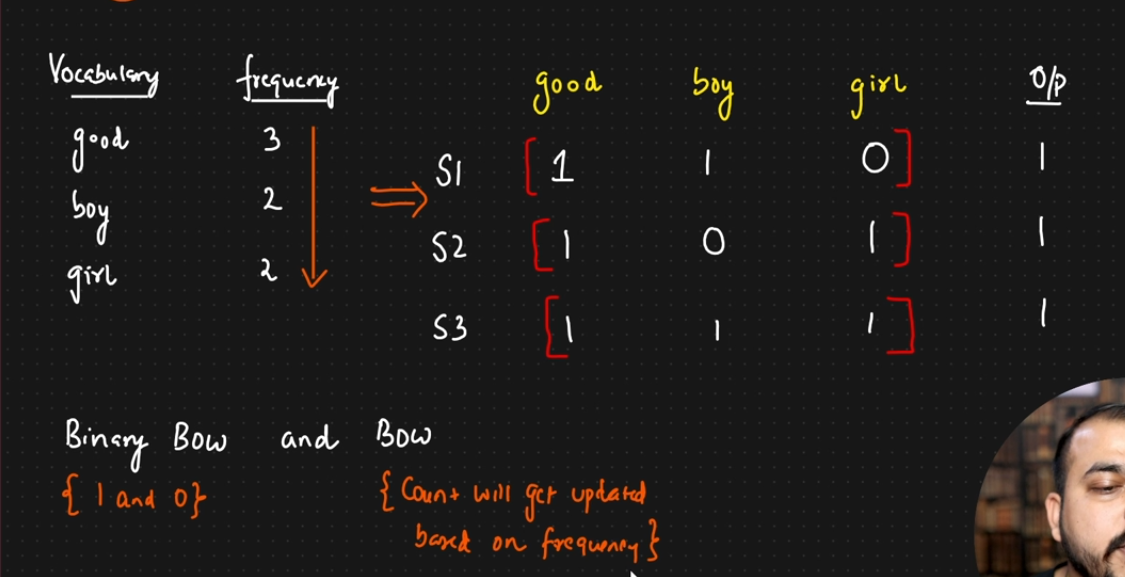
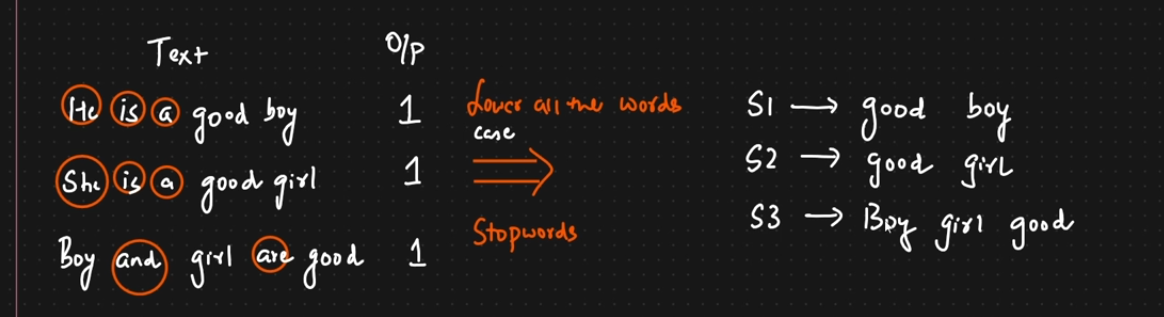
Advantages of One Hot Encoding:

1. Easy to implement (sklearn.preprocessing/pd.get\_dummies())

Disadvantages of One Hot Encoding:

1. Sparse Matrix (too many 0s and 1s).
2. No fixed size. (ML algos need that).
3. Semantic meaning is not captured.
4. Out of vocabulary.

Bag Of Words.



Advantages of BOW:

1. Simple and Intuitive.
2. Fixed size input.

Disadvantages of BOW:

1. Sparse Matrix.
2. Ordering of the word is getting changed.
3. OOV.
4. Semantic meaning is still not getting captured.

BOW N-Grams.

