13 Jan, 2023

Semantic Analysis

Text Similarity, diff ppl hv a slightly different notion on what text similarity means

meaning	surface closeness
It is called as Semantic Similarity	— " — Lexical Similarity

eg: How similar are the below phrases?

The cat ate the mouse .

The mouse ate the cat food

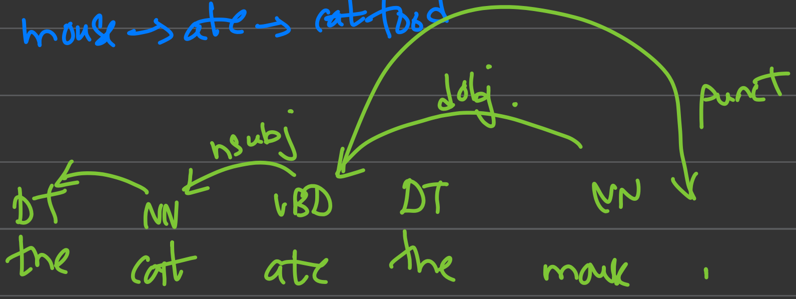
At surface level, these appear very similar.

∴ there are 5 similar words.

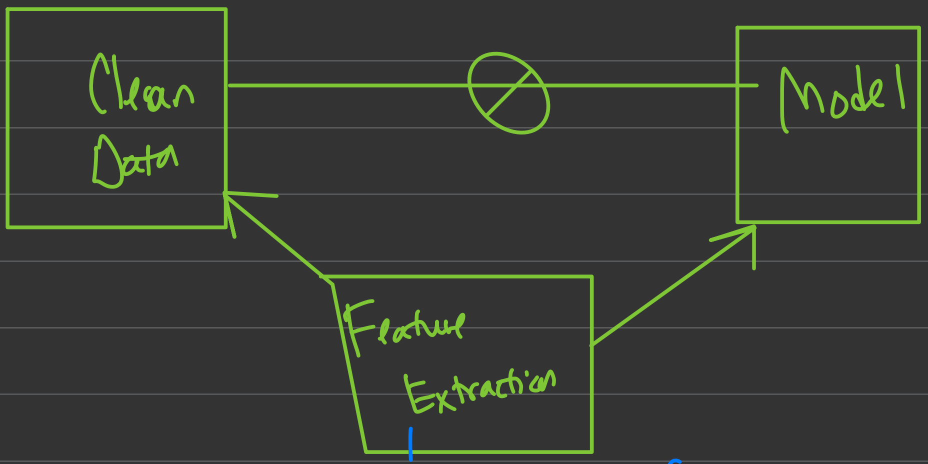
This is lexical Similarity.

Cat → ate → mouse

mouse → ate → cat food



Feature Extraction



Bag of Words (CountVectorizer)

N-gram

TF-IDF

Document Term Matrix

Cosine Similarity

Count Vectorizer (Bag of Words):

Ex:

1. I love playing.
2. We love playing cricket
3 Playing is fun

Preprocessing

lower()

1. i love playing.
2. we love playing cricket.
3. playing is fun.

Remove Punctuations

1. i love playing
2. we love playing cricket
3. playing is fun

Removal of Stopwords

1. love playing
2. love playing cricket
3. playing fun

Stemming

1. love play
2. love play cricket
3. play fun

Lets convert it into text using Feature Extraction techniques Here we r using Countvectorizer (Bag of Words BOW)

	love	play	cricket	fun
1. love play	1	1	0	0
2. love play cricket	1	1	1	0
3. play fun	0	1	0	1
Σ =	2	3	1	1

↓

Writing all the words in the corpus in

the descending order of count

	play	love	cricket	fun
1. love play	1	1	0	0
2. love play cricket	1	1	1	0
3 play fun	1	0	0	1

Original Corpus:

1. I love playing → [1 1 0 0]
2. We love playing cricket → [1 1 1 0]
3. playing is fun → [1 0 0 1]

____________ X ____________