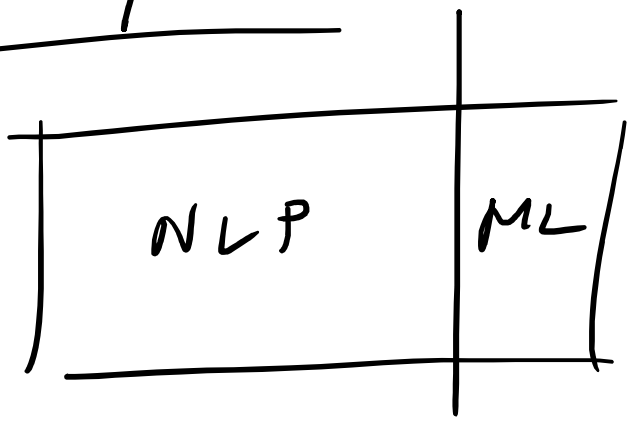
Mod 6 - CO6 - 2024-10-01 - Applications of NLP

01 October 2024 09:21

-> Text Classification:



-> Discrete Samples

[0, 1, 2, 3]

-> Preprocessing

||

Embedding

||

Model Building

-> Character Process (Regex)

↓

Effective Splitting (Sentence Conversion)

(The
the)

↓

Tokenization

↓

Casing (Lowercase)

↓

Stopwords

=> Embeddings :

↳ BoW : Bag of Words

" I like to go to the movies ".

" I do not like movies like this. "

↳ uses Stochastic Sparse Matrix.

	BoW1	BoW2	TF-IDF1	TF-IDF2
I	1	1		
like	1	2		2/7 log 7
to	2	0		
go	1	0		
the	1	0		

movies	1	1
do	0	1
not	0	1
this	0	1

-> **TF-IDF**: Term Frequency - Inverse Document Frequency.

$$TF(t, d) = \frac{\text{Occurrence of } t \text{ with } d}{\text{Total words in } d}$$

(t: Term, d: Document)

$$IDF(t, D) = \log \frac{N^{\sim}}{\text{documents with } t}$$

($N^{\sim}$: count(D); t: Term, D: all docs)

$$TF\text{-}IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$