09 May 2022 11:13

How to deal up with unstructured data

Text data --> online searching, comments, description, reviews, websites, pdf,..

1. Text - Pre-processing
2. Feature Extraction

Deep learning
Images, videos, audio, ……--> prediction

Text - Pre-processing

Collection of unique English meaning full words and removing the un-necessary English words from the whole document is called as "Text-preprocessing"

Tokenization --> word, sentence

Stemming

Lemmatization

Stop words

With two packages 1. NLTK, 2. Spacy

By Feature extraction we are going to fit a mode with Target Y --> Super vised learning
By Feature extraction we are going to topic modelling technique --> where this is works like cluster analysis

---

What is meant by tokenization?

I have separate every word from the sentence

What is stemming?

Feature Extraction

1. Word existence method

2. Word proportion method
3. TFIDF method --> Term Frequency and Inverse Document Frequency

1. This soundtrack was beautiful
2. The best soundtrack ever to anything
3. This soundtrack is my favourite music

---

1. Soundtrack beautiful
2. best soundtrack ever anything
3.soundtrack favourite music

B	C	D	E	F	G	H	I	J
	anythir	best	beautif	ever	favouri	music	soundtrack	
1	0	0	1	0	0	0	1	
2	1	1	0	1	0	0	1	
3	0	0	0	0	1	1	1	
							0	
	anythir	best	beautif	ever	favouri	music	soundtrack	
1	0	0	0.5	0	0	0	0.5	
2	0.25	0.25	0	0.25	0	0	0.25	
3	0	0	0	0	0.33	0.33	0.33	