LAPORAN DATA PRE-PROCESSING DENGAN PYTHON

ARIF FRIMA ARI SUWADJI - 221011700443 - 03SIFE003

DATA PRE-PROCESSING - Tokenizing

INPUT	OUTPUT
pwd	In [5]: pwd Out[5]: 'C:\\Users\\arifs.DESKTOP-EQJJN52\\Documents\\unpam\\Semester3\\Pengantar_Big_Data\\KelasBigData'
import nltk nltk.download('punkt') import pandas as pd	In [47]: import nltk nltk.download('punkt') import pandas as pd [nltk_data] Downloading package punkt to C:\Users\arifs.DESKTOP- [nltk_data] EQJJN52\AppData\Roaming\nltk_data... [nltk_data] Unzipping tokenizers\punkt.zip.
text = ['This is introduction to NLP', 'It is likely to be useful, to people ', 'Machine learning is the new electricity', 'There would be less hype around AI nd more action going forward', 'python is the best tool!'] df = pd.DataFrame({'data':text}) df	In [48]: text = ['This is introduction to NLP', 'It is likely to be useful, to people ', 'Machine learning is the new electricity', 'There would be less hype around AI nd more action going forward', 'python is the best tool!'] In [49]: df = pd.DataFrame({'data':text}) df Out[49]: data 0 This is introduction to NLP 1 It is likely to be useful, to people 2 Machine learning is the new electricity 3 There would be less hype around AI nd more act... 4 python is the best tool!

LAPORAN DATA PRE-PROCESSING DENGAN PYTHON

ARIF FRIMA ARI SUWADJI - 221011700443 - 03SIFE003

INPUT	OUTPUT

def proses(kalimat) :
  kalimat = kalimat.lower()
  kalimat = nltk.tokenize.word_tokenize(kalimat)
  return kalimat
 	
hasil = df['data']
hasil = hasil.apply(proses)
hasil
	In [50]:
	`def proses(kalimat) :`
	`  kalimat = kalimat.lower()`
	`  kalimat = nltk.tokenize.word_tokenize(kalimat)`
	`  return kalimat`
	
	In [51]:
	`hasil = df['data']`
	`hasil = hasil.apply(proses)`
	`hasil`
	
	Out[51]:
	0 [this, is, introduction, to, nlp]
	1 [it, is, likely, to, be, useful, ,, to, people]
	2 [machine, learning, is, the, new, electricity]
	3 [there, would, be, less, hype, around, ai, nd,...
	4 [python, is, the, best, tool, !]
	Name: data, dtype: object